BANK INDONESIA

International Statistical Institute
Regional Statistics Conference

PROCEEDINGS

of The

International Statistical Institute
Regional Statistics Conference 2017

"Enhancing Statistics, Prospering Human Life"

Bali, 20 - 24 March 2017

Traditional Balinese Kecak Dance

isi

BADAN PUSAT STATISTIK

IKATAN PERSTATISTIKAN INDONESIA

FMS Forum Masyarakat Statistik

International Statistical Institute
Regional Statistics Conference

Proceedings of The ISI Regional Statistics Conference 2017 (ISI RSC 2017)

ISI Regional Statistics Conference 2017 (ISI RSC 2017) brings together eminent statisticians and members of the statistical community from South East Asia and all around the world to present, discuss, promote and disseminate research and best practice in every field of Statistics and its applications toward enhancement statistics to prospering human life

Published by:

Statistics Department, Bank Indonesia
Jl. MH Thamrin No. 2 Jakarta Pusat, DKI Jakarta

INTERNATIONAL STATISTICAL INSTITUTE
REGIONAL STATISTICS CONFERENCE 2017 (ISI RSC 2017)

ISBN : 978-602-61659-0-9
Book Size : 16,5 cm x 24 cm
Total Pages : 1095 pages

Manuscript:
Statistics Department – Bank Indonesia

Cover Design:
Statistics Department – Bank Indonesia

Published by:
Statistics Department – Bank Indonesia

Printed by:
Statistics Department – Bank Indonesia

PREFACE

The ISI Regional Statistics Conference (RSC) 2017 was organized by the ISI and its South East Asia Regional Network (ISI-SEA Network) in collaboration with Bank Indonesia (BI) as the co-host, and supported by Badan Pusat Statistik (Statistics Indonesia), Ikatan Perstatistikan Indonesia (Association of Indonesia Statistician), and Forum Masyarakat Statistik (Indonesia Statistics Community Forum).

The second ISI Regional Statistics Conference (ISI RSC 2017) was a three day conference preceded by the ECB-BIS-BI Regional Seminar on Central Banking Statistics, the Irving Fisher Committee (IFC)-BI Satellite Seminar on Big Data, and short courses. The Main Conference program was held from 22 to 24 March 2017 in Bali International Conference Center, Nusa Dua, Bali. The conference theme "Enhancing Statistics, Prospering Human Life" was chosen to encourage researchers and practitioners who are actively involved in statistical science in academia, industry, national statistical offices, national and international agencies, central banks, as well as users to participate in discussions on how statistics may contribute to prosper human life. The conference theme was wide enough to accommodate participants with diverse interests.

There were three Plenary Sessions and parallel sessions discussing 193 papers in 29 Invited Paper Sessions (IPS) and 34 Contributed Paper Sessions (CPS).

This publication contains abstracts, papers and materials presented in the ISI RSC 2017. Release of this publication has been made possible by the assistance and contribution of all contributors. To all parties who have been involved in the completion of this publication, we would like to express our sincere gratitude and high appreciation. Hopefully this publication will be a useful resource for any purposes. Comments and suggestions to improve the publication are always welcome.

TABLE OF CONTENTS

Thursday, 23 March 2017

Welcoming Remarks

"Enhancing Statistics Prospering Human Life"

Dr. Sugeng
Deputy Governor, Bank Indonesia
Welcome Address
at the International Statistics Institute-Regional Statistics Conference
(ISI-RSC) 2017
hosted by the International Statistical Institute (ISI) and its South East Asia Regional Network in collaboration with Bank Indonesia (BI), Statistics Indonesia (Badan Pusat Statistik/ BPS) and Ikatan Statistisi Indonesia (ISI)
22 March 2017, Bali, Indonesia

Yang kami hormati Bapak Prof. Boediono, Vice President of Republic of Indonesia in 2009 to 2014;

President of the International Statistical Institute, Mr. Pedro Luis do Nascimento Silva;

Former Deputy Governor of Reserve Bank of Australia and ANU Professor, Prof. Stephen A. Grenville;

BIS Chief Representatives for Asia and the Pacific, Mr. Eli Remolona;

Distinguished resource persons, ladies and gentlemen,

1. It is my great honor and delight to welcome all of you to the International Statistics Institute – Regional Statistics Conference on "Enhancing Statistics Prospering Human Life". I would like to particularly extend a warm welcome to my honorable colleagues, distinguished speakers and guests from around the world who have travelled a long way to be with us today.
2. This conference is a joint collaboration between International Statistical Institute (ISI) and its South East Asia Regional Network with Bank Indonesia (BI), Statistics Indonesia (Badan Pusat Statistik/ BPS) and Ikatan Statistisi Indonesia (ISI).
3. I do appreciate the great enthusiasm from all participants, ranging from policy makers, economists, statisticians, scholars, practicioners and also students to attend this important conference that will discuss a broad set of statistical issues of interest not only to the central banks, but also to a broader usage in the economy. We have here today around 400 participants whereby 223 participants are presenters.
4. Prior to this conference, in the last two days, we had two seminars with interesting topics, namely Seminar for Asian Countries on Recent Developments in Central Bank Statistics and Seminar on Big Data, and also some workshops in recent statistical issues and methodology.

5. The theme of the conference today "Enhancing Statistics Prospering Human Life" echoes an often recurring question of how statistics affect human prosperity. How do we undertake statistics enhancement to attain prosperous human life?
6. In a finite world, under conventional wisdom, human prosperity in economic terms calls for continuing economic growth as the means to deliver it. Continuing growth means that it needs to be sustainable. Thus, statistics enhancement should touch upon all aspects pertaining to sustainable economic growth, as a means to attain human prosperity. This basic understanding is essential for policy makers, academia, economist, analysts as well as statistician to understand the statistics that reflects an amalgam of influences to the prospering human life.

Distinguished speakers and participants,

7. As we may recall, the year 2016 have brought about times of prolonged uncertainties. The global economy showed resilience but with sub-par economic growth at only 3.8%. In 2017, growth is set to pick up better compared to last year despite such optimism are still centered over a considerable amount of uncertainty. Indonesia recorded a 5.02% growth in 2016 and is expected to reach 5.2% in 2017 with all three major rating agencies now having a positive outlook.
8. In support of attaining sustainable economic growth, this conference shall discuss several aspects of statistics related to macro-financial statistics, which consist of three upmost important aspects that are statistics for central banks, sharia eonomic & finance and financial statistics. Other topics for discussion include statistical theories and official statistics.
9. The dynamic nature of macroeconomic and financial world undoubtedly demands reliable data and information of various economic events and indicators. Macro-Financial Statistics aids in the formulation of monetary policy and macro prudential measures. Moreover, a robust central banking statistics is critical to support policy analysis. This includes to understand the interaction between financial market conditions and economic activity, the identification and measurement of spill-over risks in financial sector activities as well as the interaction between monetary policy, financial stability and the payment system.
9. Another approach for sustainable economic growth is founded through Islamic Finance. Islamic Finance have been part of the global financial system that has developed rapidly in the last two decades. In short, the basic principles of Islamic finance have social and environmental goals that are drivers to sustainable economic development and enhancing resiliency of the financial sector. Therefore, to support the advancement of sharia finance and economy, statistically robust indicators must be available. This calls the need for speeding up the development of the statistics for sharia financial and economic as part of our dynamic economy.
10. To attain the betterment human life, official statistics should also be enhanced. The most important issue with regards to official statistics is its accuracy. Official statistics that form the basis for government policy formulation needs to reliable. Coordination should also be strengthened in the areas of statistical methodology development and among data sources.

Distinguished guests, ladies and gentlemen

12. To conclude, the rich perspectives and ideas taken from this conference will only be useful once it is put into concrete actions. We believe that a successful enhancement of these many areas of statistics will help us to effectively and efficiently address the critical focus in uplifting the economic progress and at the end, the quality of life.

13. In closing, let me offer my special thanks to all the distinguished speakers for taking part in this important endeavor and for sharing with us your expertise. All of us here very much look forward to your contributions.

    To all conference participants, I hope your active involvement to have productive discussions, as I have already noticed there are many experts present among us. Thank you for being here.

    Last but not least, I wish you a fruitful conference. And don't miss your chances here to also explore this beautiful island of Gods. Have a wonderful and enjoyable stay in Bali. Thank you.

Nusa Dua, Bali 22 March 2017

Wednesday, 22 March 2017

Plenary Seasson 01 (PS01): Banking and Finance

This page is intentionally left blank

Information and Statistics in Public Policy Making

Boediono
11th Vice President of the Republic Indonesia

BIConfBali22317.XX

INFORMATION AND STATISTICS IN PUBLIC POLICY MAKING Ψ

Boediono

Let me thank the organizers for inviting me to this important conference. To all participants, welcome to Bali. Before proceeding, though, I feel obliged to offer my reason why I, being a non-statistician, might usefully speak in the forum of professional statisticians such as this.

While I am not a statistician by profession, statistics have never been far from my works throughout most of my career. I took up my first job in government more than thirty years ago at our national planning agency - Bappenas – to lead the Bureau for Economics and Statistics. As the bureau's name suggests my main responsibility was to supply the institution's needs of data – especially economic data - for the planning process. It did not itself collect data but instead relied on other more formidable data collecting agencies. Conveniently as it turned out, I was also given the responsibility of overseeing the programs and the annual budget of our national statistical office - BPS.

In the subsequent years, as I increasingly took up the decision making responsibility, my role invariably shifted from facilitating the production of statistics to one of a principal user of statistics. So I thought it might be useful to share in this forum how we, the users of statistics and information in general, see their role in policy decision making in government.

Ψ Presented at the International Statistical Institute – Regional Statistics Conference 2017, Bali, March 22, 2017.

Let me start with the ideal information situation in which any policy maker would love to be whenever he/she has to make decision: all the relevant data with unquestionable accuracy are available in real time at his/her finger tip. Alas, that ideal situation is never to be. Even in the best of circumstances the hard reality is that, information wise, policy makers are always 'behind the curve'.

Why? The main reason why it is so is that a policy maker is always bound by a time table. At a particular juncture he or she has to come up with a decision on what actions to take on the basis of the 'best' information available to him or her at that critical time, which most probably are neither complete nor very accurate. Very often to get that 'best' information his or her team have to scramble to assemble data from different sources, in and outside the bureaucracy. The assembled information consists of data with differing completeness and quality, a kind of "information salad' or 'information soup'. The policy maker has to make the best use of it and make a decision.

To be fair to the statisticians and other data producers, I must add that in reality the problems of policy making do not come only from the 'supply side' or the availability and quality of information. Very serious problems could in fact occur on the 'demand side' or the way the available data are being used. The 'cook', if I could metaphorically call the supporting team who are tasked to process and analyze the assembled information and present actionable options, for weak technical expertise or lack of sound judgment, may not do a good job. The options are then flawed or misleading. Once such options find their way to the decision maker it is hard to expect a quality policy to be the outcome, unless we could assume that the decision maker happens to be a supremely wise and extremely knowledgeable person. A rarity indeed.

Important as they are, I will not dwell further on the 'demand side' problems. My comments that follow will be largely on the 'supply side' ones. Inevitably, my Indonesian experience will influence my story. And I will remain focus on public policy making.

Let me underscore that policy decision making is essentially a multistage input-output process. The quality of the resulting policy is the sum total of the qualities of all those inputs and outputs along the information chain. To improve the quality of the end product – the final policy outcome - therefore one must look into the possibility of improving the quality of the output of each related institution along that chain.

To begin, we should recognize the fact that in formulating policies, national governments rely mostly on information generated within and by its own institutions. The national statistical office usually stands out as its principal source for basic economic and social information. In this country three other institutions deserve special mention. The central bank is the sole source of monetary statistics, finance ministry for fiscal statistics and the financial services authority for data on banking and other financial instititutions. These four institutions are the first-tier information providers for policy making.

Certain other institutions also collect data related to their respective functions but with more limited coverage and generally of lesser quality. They are the second-tier information providers. To name a few: ministry of home affairs for regional finance and some social indicators, ministry of agriculture for agriculture-related statistics, ministry of industry for industrial production statistics, ministry of transports for air, sea and land transport capacities and traffics, ministry of public works for the state of road and irrigation systems.

The quality of the information vary greatly across institutions, notably among the second-tier ones. It shows the differing capacities in their information gathering and processing. But actually it reflects a deeper and more general problem - a lack of appreciation of the critical role of good information in making good decision. In today's world it is generally accepted that accumulated institutional knowledge and effective information system are the foundation of a "smart" institution (and hence smart policies). It seems though that such a view has not caught on in many government institutions. It is one of the fundamental challenges of a country's bureaucratic reformers.

The potential of improving information capability in the institutions I mentioned earlier is substantial. There are still enough rooms for raising the operational standards of even the first-tier institutions to the international best practices. And clearly there are plenty of rooms to level up the information capability of those second-tier agencies through redefining information gathering function in each of them, providing sufficient number of qualified personnel and securing adequate budget for them. To be sure partial efforts have been made along this line. But to make them stick the initiatives must be substantively incorporated in their respective long term reforms agenda. Better still if they are made to be an integral part of a broader plan for national bureacratic reform. Systematic efforts along this line in my view will give the greatest payoffs for policy making.

Recently I have been trying to follow the lively discussions among statisticians and data scientists on the potential benefits of using privately collected "big data" in improving the operations in both government and business. If we believe that the key to national progress is better public policies and better business conduct, then we must take the issue seriously. For a non specialist like myself, though, it is too complex an issue to jump in. So let me make only some general comments on it.

The first point I wish to make is that not only the private sector but the government could also be a producer of big data. There are many routine government processes at the national and subnational levels that could generate continuous streams of large scale and uptodate information. If digitized, they could become invaluable big data systems. Raising the standards of the digital technology usage and practices in government agencies would directly improve their 'traditional' activities in information gathering and processing while indirectly also raise the probability of success of any planned government's cooperation schemes with the private sector in utilizing other big data systems. Digitizing government's administration processes will give even larger payoffs as it helps raise the efficiency and integrity of the day-to-day operations of the bureaucracy.

This is a big, long term job with many challenges. Some of them may spring up at the very beginning. Thus a common problem is that the existing IT systems of government agencies are not compatible one another. Let me relate a story. I once was tasked to improve coordination of the agencies' IT development plans and I tell you how energy-sapping a job it was. It turned out that each agency had a legacy system not easily reshaped and reoriented. The reason, though, is not so much technological as institutional, namely bureaucratic inertia or resistance toward change. The important lesson from the case study was that getting a firm hold of their IT budgets was the minimum requirement. You need more than that. You must have some reserve energy for breaking many forms of institutional inertia and resistance. One form that we found particularly difficult to deal with has a root in the so-called 'vendor driven' planning practices. By the end of its term the task force at best registered only a partial success. Nevertheless I would reiterate that digitizing government processes and developing government-owned 'big data' is a truly worthwhile effort and should be redoubled in the future.

There is a big promise from the possibility of utilizing non-government big data which recently have grown exponentially as a result of the ever expanding digitization of ordinary social and economic processes. We are told that currently we are still at the beginning of a long process. If the government could tap these enormous sources of information, the quality of its administrative and policy decisions could be vastly improved with far less costs, and the society stands to gain.

These new sources of information are useful for strengthening and sharpening the 'traditional' policies. For instance, they potentially will make obsolete surveys such as those on consumers' confidence, investors' confidence and employment situation. Such surveys are essential for calibrating macroeconomic policy stance. Eventually they will be replaced by direct and real-time readings of the relevant big data. There are other instances, such as in health, education, poverty alleviation and transportation where the use of big data offers entirely new policy perspectives and possibilities.

The use of privately collected information by government involves a combination of the use of compulsion and voluntarism. Government can issue regulations compelling the private parties to share their information with the government. But in democracy and market economy there are political and economic limits to the application of the coercive power of the state. When the limits are reached we will have to rely on voluntary cooperation agreements between the government and the private parties. Such 'public-private partnership' in information sharing is essential but may not be easy to come by, especially in the newly digitized social and economic processes.

For the traditionally highly regulated sectors such as the financial sector, voluntary cooperations mean information sharing arrangements beyond what is mandated by prevailing prudential regulations which themselves are continually evolving. From the regulators' and policy makers' points of view, obviously, more, better and more timely data would be very helpful for their routine prudential surveillance job and, even more crucially, for policy makers in managing the fluid situation in times of crisis. But we know that beyond certain points compulsion becomes harmful for the efficient operations of financial institutions and markets, and most probably also for individual customers.

Judicious combination of regulations and cooperations is therefore key to the success of the endeavor. And since the use of big data for supporting policies most likely entails new institutional arrangements, new territories and new modus operandi, experts advice us to start with small scale experimentations then from there move on to scale them up, only after lessons have been learnt and extracted from the former.

To close my talk let me summarize its main points.

- The quality of policy making is determined by the quality of the available information and the way the available information is being used.
- In policy making governments still rely mainly on information generated by their own agencies. A key step to improve the quality of policy making is therefore by systematically raising the information producing capability of the relevant institutions.
- Digitizing routine government processes will improve the quality of policy making while indirectly also gives large benefits with the improvements in the efficiency and integrity of the government bureaucracy.
- The growth of privately collected big data opens up a new posibility of vastly improving public policies with far less costs. The key is how to evolve a judicious combination of regulations and voluntary cooperation schemes. The best way to move forward is to start with small experiments and as lessons gained, move on to scale them up.

Thank you.

@@@@@@@@

This page is intentionally left blank

Statistics for Banking and Finance

Stephen Grenville
Former Deputy Governor, Reserve Bank of Australia

Statistics for Banking and Finance

Stephen Grenville
Lowy institute for International Policy
Bali
23 March 2017

1

Introduction

- Narrow focus: how statistics can enhance the central bank's policies.
- I'll set out problems and hope you statisticians can offer help

2

Introduction

- Three primary roles:
  - monetary policy
  - Financial stability
  - Financial sector development

3

Introduction

- Hong Kong's former view that policy needed very few statistics.
- 2008 global financial crisis triggered demand for many more statistics, especially internationally comparable time-series.
- Of course more is better, but this is not costless. Prioritise between domestically oriented statistics and the demands of the global institutions.

4

Introduction

- IMF SDDS
- G20 Data Gaps Initiative
- BIS Banking Statistics

5

Introduction

- Why might (say) Indonesia's priorities differ from global priorities?
  - Very diverse economy, so broad aggregates don't capture complexity.
  - Growing fast, so changing quickly
  - Different vulnerabilities (e.g. capital flows)

6

Introduction

- Speed of collection may be important (e.g. during a crisis)
- Frequency of collection? For most policy issues, small advantage in frequent collection.
- Trade-off between quantity and quality

7

Monetary Policy

- Starting point might be mechanical policy rule (usually a Taylor Rule)

$$i = r^* + a\,(\underline{p} - \underline{p}^*) + b\,(Y - Y^*)$$

- Each one of these components presents measurement problems

8

Monetary policy

- Interest rate. What interest rate matters for policy?
- Natural (long-term) interest rate. Changing over time. Close to zero in the US now???

9

Monetary policy

- Inflation
  - Even historic rates are problematic (Boskin)
  - Inflation targeting frameworks emphasize INFLATION EXPECTATIONS, not actual
  - Too many different series (e.g. underlying versus headline)
  - Public scepticism about measurement accuracy

10

Monetary policy

- Output
  - Delays in measuring actual output
  - Problems of measuring potential output. Productivity; terms of trade; end-point problem with trend-based measures

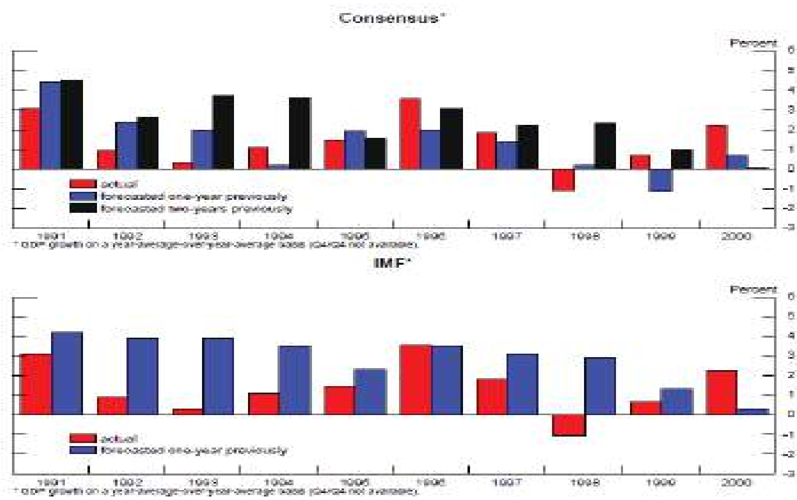
11

Monetary policy

- And all this has to be FORWARD-LOOKING
  - This involves forecasts and surveys of expectations, both problematic

12

We are not good at forecasting



13

Monetary policy

- As well, there are non-formula factors:
  - Headwinds from business/household confidence, bank credit conditions, political uncertainty
- So greater reliance on surveys, with all their problems (subjective, changeable, sensitive to environment)
- Greater emphasis on key role of asset prices

14

Broader macro

- Foreign capital flows. Not principally a statistical issue: rather analytical and structural, but there are important behavioural differences. Need for granularity and disaggregation. How much netting and what does it mean? Currency mismatches? Quality of the debt: who is hedged and with whom?
-

15

Broader macro

- Hence FLOW of FUNDS: “from whom and to whom”
- Plus assets/liabilities
- Plus risks and specific characteristics of assets

16

Financial stability

- Need for better financial stability policy was main 'take-away' from 2008 GFC. But main problems were operational (inadequate prudential supervision and forbearance) and policy, not shortage of statistics.

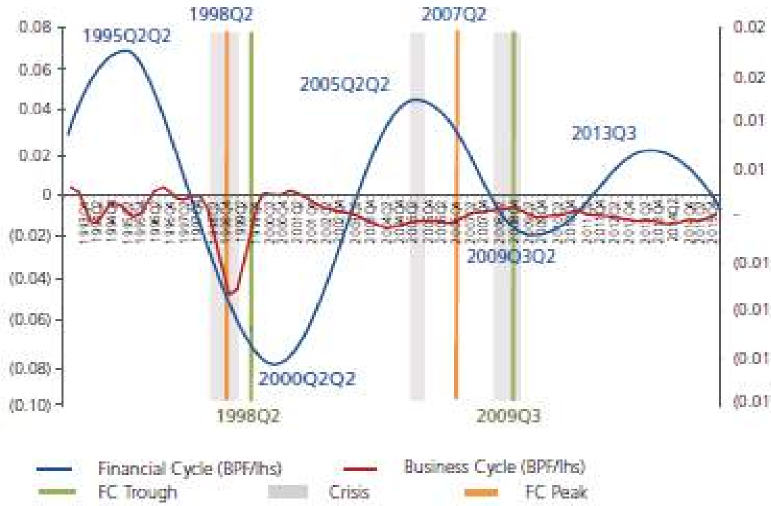
17

Financial stability

- From around 2000 onwards (thanks to Borio and Lowe etc) there was a recognition that financial cycle differs from business cycle, but even now we don't yet know what components should be in the measure of the financial cycle (clearly not just GDP, and it is more than credit) and how this cycle behaves.

18

The financial cycle



19

Financial stability

- Problems arise in sub-set of aggregates rather than total. Aggregates don't generally cause crisis, but which sub-sets?
- Gross versus net
- Problems of consolidation

20

Financial stability

- Usually about TAIL RISKS. Hard to measure and low probability, but usually not a surprise. But what to do even if tail risk is identified: enumeration of tail risks is unhelpful for policy-making. A probabilistic approach?

21

Financial development

- FLOW of FUNDS is key here too.

22

Conclusions

Priorities:

– Inflation expectations. The heart of inflation targeting: the anchor. How to maximise benefit from surveys?
– Better analysis of tail risks in financial stability.
– More generally, how to handle risk.
– Flow of funds-plus (i.e. with assets/libilities)

23

Conclusions

- Financial risks are best demonstrated in a narrative (heuristics, probabilities and additional anecdotal detail), with statistics forming just one element of the story.
- Start to develop ASEAN-wide statistics to encourage integrated thinking among policy-makers and analysts.

24

Conclusions

- For financial stability, use stress-tests and crisis simulations to explore the non-linearities and 'unknown unknowns' (testing tail-risk). Out of these exercises will emerge the gaps in data. For example, what data you need to determine 'systemic', and how quickly you need it.

25

Conclusions

- User-friendly access so that data is used for analysis and research. Public monitoring helps. And seek feedback.
- Statistics responding to political demands. Anticipate this.

26

Conclusions

- Is it now time to draw breath and prioritise domestic needs? Spend time analysing data, so that imperfections and gaps will become more apparent, from a bottom-up approach.
- DGI, SDDS and BIS banking statistics have allowed these institutions to tell important stories about global interactions and contagion risks. Comparing your country with another similar country using these data will be insightful.
- But now it is time to focus on the idiosyncratic nature of your own domestic economies.

27

This page is intentionally left blank

The New Cross-Border Finance in Asia

Eli Remolona
Chief Representative of the BIS Representative Office for Asia and the Pacific

BANK FOR INTERNATIONAL SETTLEMENTS

The new cross-border finance in Asia

Eli Remolona

ISI-RSC 2017 Bali Conference

Nusa Dua, 22 March 2017

The views expressed are our own and do not necessarily reflect those of the BIS

The new cross-border finance in Asia

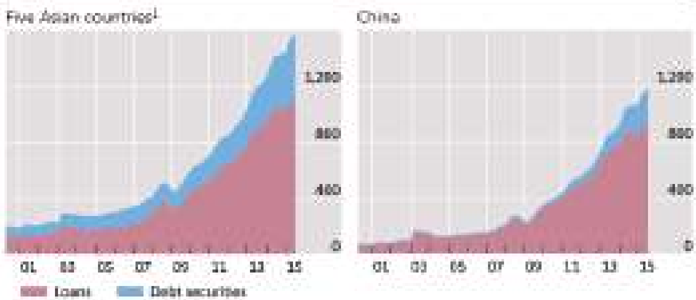
- When Asian non-banks now raise funds abroad, they tend to borrow in US dollars
- In borrowing from banks, there is good news and there is bad news
- In the shift to bond finance, what how do global investors differentiate between sovereign risks?

Restricted 2

Asian non-bank borrowers adore US dollar credit

US dollar credit to non-bank borrowers in five Asian countries

In billions of US dollars



¹ China, Indonesia, Korea, Malaysia and the Philippines.

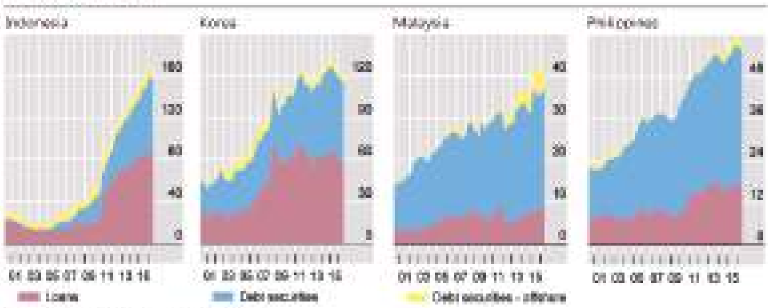
Source: McCauley, McGuire and Sushko (2015).

Restricted 3

Outside China, Asian firms go for bonds rather than bank loans

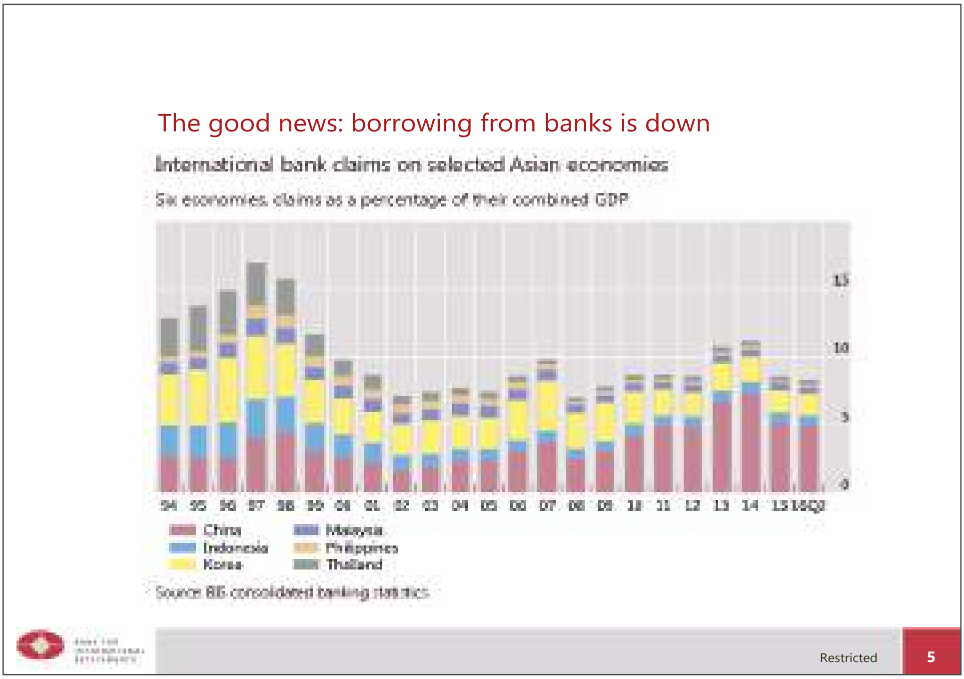
US dollar credit to non-bank borrowers in four Asian countries

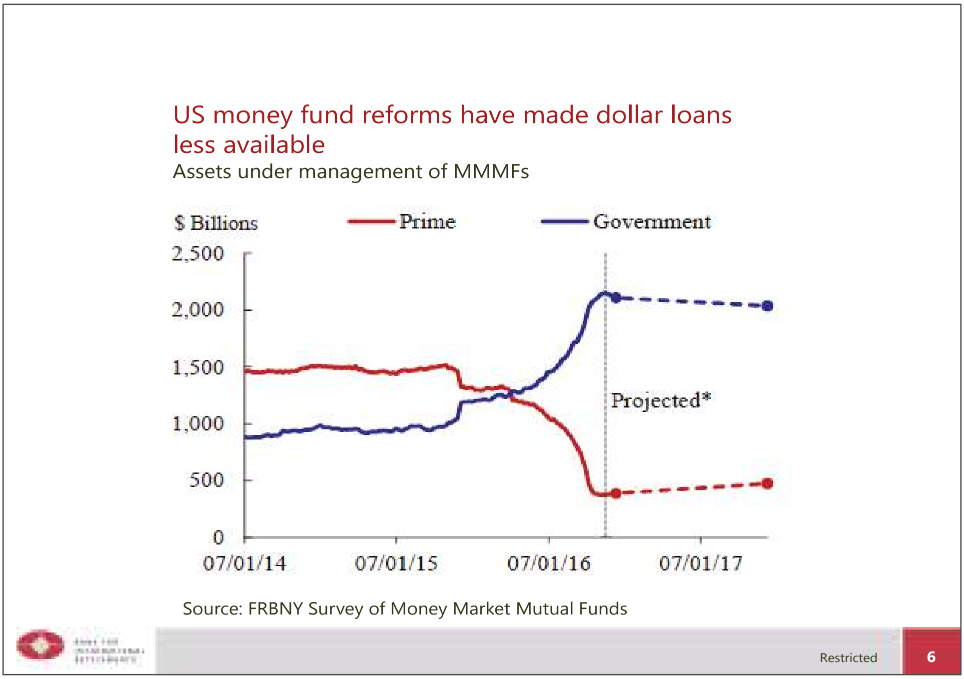
In billions of US dollars



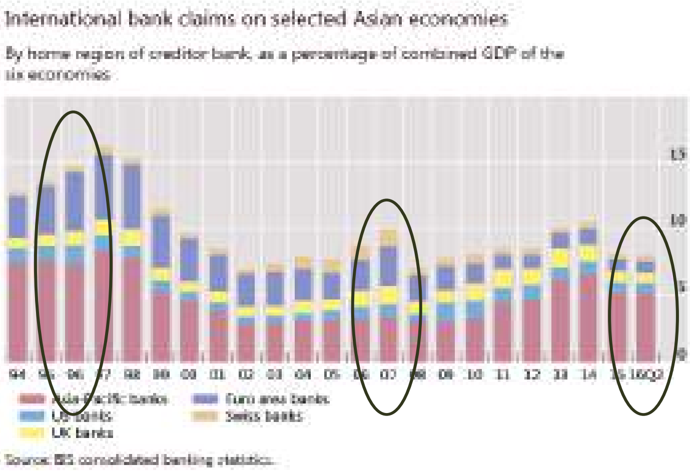
Source: McCauley, McGuire and Sushko (2015)

Restricted 4


The good news: borrowing from banks is down
International bank claims on selected Asian economies
Six economies claims as a percentage of their combined GDP
15
10
5
0
94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16Q2
China
Indonesia
Korea
Malaysia
Philippines
Thailand
Source: BIS consolidated banking statistics
Restricted
5

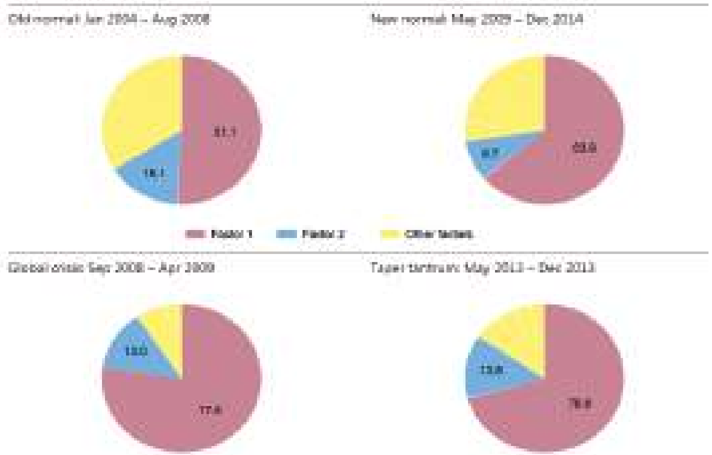

US money fund reforms have made dollar loans less available
Assets under management of MMMFs
$ Billions
Prime
Government
2,500
2,000
1,500
1,000
500
0
Projected*
07/01/14
07/01/15
07/01/16
07/01/17
Source: FRBNY Survey of Money Market Mutual Funds
Restricted
6

The bad news: concentration of funding is up



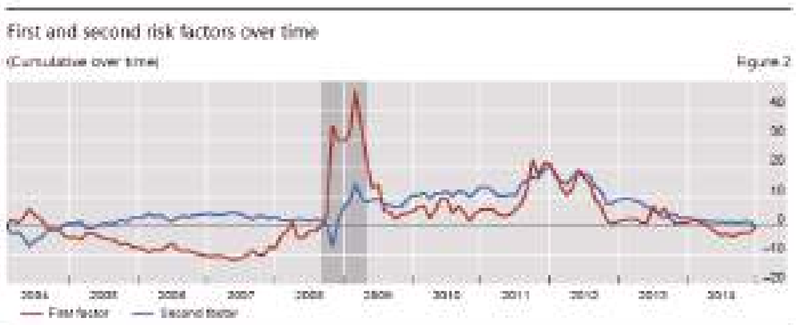
Restricted 7

How much can global risk factors explain?



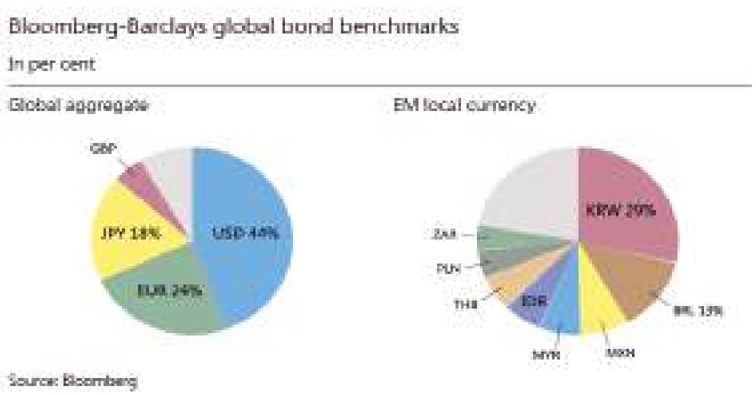
Restricted 8

Global risk factors drive what happens over time



Restricted 9

Bond benchmarks rule what happens across countries



Restricted 10

The new cross-border finance in Asia

- Asian non-bank borrowers have found US dollar credit hard to resist
- In cross-border bank loans, borrowing is down but concentration of funding is up
- In the shift to bond finance, global benchmarks are destiny

Restricted 11

References

- Amstad, Remolona and Shek (2016): How do global investors differentiate between sovereign risks? The new normal versus the old. *JIMF* 66.
- Bruno and Shin (2015): Global dollar credit and carry trades: A firm-level analysis, BIS Working Papers No 510 (August)
- Cohen and Remolona (2008): Information flows during the Asian crisis: Evidence from closed-end funds, *JIMF* 27 (June).
- McCauley, McGuire and Sushko (2015): Dollar credit to emerging market economies, *BIS Quarterly Review* (December).
- Remolona and Shim (2015): The rise of regional banking in Asia and the Pacific, *BIS Quarterly Review* (September), 119-134.
- Van Rijckeghem and Weder (2003): Spillovers through banking centers: A panel data analysis of bank flows, *JIMF* 22, 483-529.

Restricted 12

This page is intentionally left blank

IPS01:
Statistics for Sharia Economics and Finance

Could Big Data Take Islamic Finance to the Next Frontier?

Zamir Iqbal
The World Bank, Washington, D.C., USA – ziqbal@worldbank.org

Abstract

Islamic finance is experiencing significant growth globally but despite high growth the industry has yet to achieve its full potential. Islamic finance is a risk-sharing finance as opposed to conventional debt based risk-transfer finance. The key reason for low acceptance of risk-sharing and asset-based finance is typical information asymmetries and high monitoring costs leading to preference for debt-based financing. With the emergence of big data and increasing computing power enabling complex analysis and faster processing, information asymmetries could be overcome and thus reducing monitoring costs for risk-sharing financing such as equity and asset-based financing. With the development of sophisticated digital platforms with up to date information on the borrowers and with technologies like distributed ledgers, new financial products promoting risk-sharing could emerge. Such trends are promising for Islamic finance as it can promote financing mechanisms such as peer-to-peer lending, asset-based and equity financing, crowd funding, and venture capital which could become catalyst for Islamic finance especially for SME financing which could spur economic growth and development.

Keywords: Islamic finance; Big data; Asymmetrical information; Risk sharing finance

Measuring Islamic-Based Socioeconomic Progress: Issues and Challenges

Dr. Muhamed Zulkhibri Phd, CFP

Senior Economist,
Co-Editor, Journal Islamic Economic Studies (IES),
Islamic Research and Training Institute (IRTI),
Islamic Development Bank

Abstract

The recent global financial crisis and rising inequality around the world have countries, institutions and policy makers to rethinking new tools, statistics and measures for their development agenda. The importance of integrating other relevant indicators beyond the existing measurements is becoming even more relevant today. Furthermore, in the Muslim world, there is an increasing awareness among Muslim societies to measure socioeconomic progress based on Islamic principles.

The aim of the presentation is to address several emerging issues on how to measure socioeconomic progress based on Islamic paradigm; do we really need new measures and different indicators of socioeconomic progress for countries with Muslim majority, and what are the challenges in developing such indicators for OIC member countries.

Keywords: Islamic finance; Islamic paradigm; Socioeconomic

Prudential and Structural Islamic Financial Indicators (PSIFIs) – Soundness Indicators for Islamic Finance

Md. Salim Al Mamun
Member of Secretariat, Technical and Research

Abstract

The recent global financial crisis witnessed rapid innovation, product diversification, deregulation, and integration across sectors. The crisis also highlighted the need for both microprudential and macro-prudential data to support oversight of the financial industry, and market discipline over that industry. Moreover, there is a need to analyse the collected data in light with the changes in the structure and activities of the industry.

While the industry demonstrated resilience during and after the global financial crisis, the Islamic financial services industry (IFSI) in various jurisdictions was not immune to the overall macroeconomic environment and turbulence in the financial system. Therefore, stakeholders have reiterated the need for a well-developed global database on Islamic finance to track the progress of the industry and review its risks and vulnerabilities.

On the conventional side, the IMF's FSIs are macro- prudential indicators which measure the aggregate strength and vulnerabilities of financial systems. FSIs are also analysed in the context of Fund surveillance. The progress towards systemic and reliable collection, compilation and dissemination of Islamic finance statistics has been slow in the past. Most existing data platforms provide micro-level data/information on individual Islamic banks, while others publish macro-level data aggregated from individual Islamic financial institutions.

The Articles of Agreement approved by the IFSB Founding Members at the time of its establishment in 2002 had recognised the need for such a database, and as such one of the objectives of the IFSB aims at establishing a database of Islamic banks, financial institutions and industry experts [Article 4(h)]. After the two initial phases involving issuance of Compilation Guides and pilot data collection from five countries, the IFSB's PSIFIs project was launched in April 2015. So far, the IFSB disseminated the data on its own website five times. Now the data is available from 2013Q4 to 2016Q1 on the website. The project collects and disseminates reliable and regular data on the financial soundness and growth of the IFSI in participating IFSB member countries on quarterly basis, while addressing the specificities of Islamic banking. The PSIFIs are parallel to the IMF's FSIs, which measure the aggregate strength or vulnerabilities of the financial systems at the jurisdiction level.

Finally, against the industry need, within a short period of time, the PSIFIs database has been effectively able to fill the gap. The PSIFIs database now holds about 90% of global Islamic banking industry. The PSIFIs are useful as an input to a broader surveillance framework, together with other relevant indicators of economic and financial positions of a jurisdiction. Supervisory authorities can conduct macro-level stress tests using PSIFIs data on capital adequacy, asset quality, profitability, liquidity, and exposure to risks to understand vulnerability to shocks and capacity to absorb the resulting losses.

Keywords: Islamic finance, Database, Soundness
JEL: G2

Integration of Islamic Commercial and Social Finance for Socio-Economic Development and Financial System Stability

Ascarya
Islamic Economic and Finance Department, Bank Indonesia – ascarya@bi.go.id

Abstract

Different from conventional finance where commercial finance and social finance are separated, Islamic finance is essentially a blended combination of its commercial finance and social finance with similar guidance (maqashid Shariah or Shariah objectives) and principles to achieve similar goal of wellbeing in this world and in the hereafter. This study aims to formulate the integration between contemporary Islamic commercial and social finance in order to improve socio-economic development and strengthen financial system stability.

It can be concluded that the development of contemporary Islamic economic and finance is not only concern with commercial aspects, but also it concerns with social aspects (such as zakah, waqf and Islamic microfinance), where both aspects should not be separated. Therefore, the integration between the social and commercial sector could not only alleviate poverty, improve socio-economic wellbeing and improve holistic financial inclusion, but also it could improve financial system stability.

Baitul Maal wa Tamwil (BMT) is one well-known type of Islamic microfinance institution combining social finance (managing zakat, waqf, infaq and other Islamic charities) and commercial finance (providing microfinance services) to service its members (including the poor and MSEs) and the surrounding community, which is an ideal model of Islamic institution to carry out the above goals in micro-small scale. Islamic Financial Institution (IFI) such as Islamic Bank and Islamic Rural Bank in Indonesia, could also adopt BMT model combining Islamic commercial finance and Islamic social finance, by establishing its own Baitul Maal to manage Islamic social finance, to improve their stability and sustainability, as well as to contribute to financial system stability. The authority should provide incentives for IFI combining Islamic commercial finance and Islamic social finance.

Keywords: Islamic finance, Islamic commercial finance, Islamic social finance, Islamic microfinance, Baitul Maal Wa Tamwil

JEL Classifications: G21, G28, O17

Developing Real Sector Return Index

Jardine Husman
Islamic Economic and Finance Department, Bank Indonesia – jardine@bi.go.id

Abstract

The practices of Islamic banks in anchoring their pricing to conventional banks and market interest rate are ubiquitous. The absence of an indicator that allows measurement of the actual return of real economic activities is the major argument behind these practices. This study proposes a cash-based method in measuring company performance in generating return from their real economic activities, excluding any interest revenues and expenses. Aggregating the values from individuals firms in one economic sector, we constuct an index to reflect economic performance of the particular sector.

Keywords: Cash-based method; Real sector return index

IPS03: Advance Developments in Ordinal Data Analysis

Ordinal Model Effect Measures
Alan Agresti

Cluster Analysis for Ordered Categorical Data
Ivy Liu

Dispersion and Response Styles in Ordinal Regression
Gerhard Tutz

Ordinal Model Effect Measures

Alan Agresti
University of Florida, Gainesville, USA – aa@stat.ufl.edu

Abstract

We consider simple ordinal model-based probability effect measures for comparing distributions of two groups, adjusted for explanatory variables. An `ordinal superiority' measure summarizes the probability that an observation from one distribution falls above an independent observation from the other distribution, adjusted for explanatory variables in a model. The measure applies directly to normal linear models and to a normal latent variable model for ordinal response variables. We show the forms for cumulative link models with probit, logit, and log-log link functions. Another ordinal superiority measure generalizes the difference of proportions from binary to ordinal responses. We also present related measures directly for ordinal models for the observed response that need not assume corresponding latent response models. We present confidence intervals for the measures and illustrate with an example. We also discuss ordinal summary measures of predictive power that are analogues of R-squared and the multiple correlation. This is where the abstract is placed. It should include a statement about the problem being addressed in the presentation (and paper, if submitted). Continue with a discussion of why it is important to address this problem. This may be followed by some summary information about the models and methods developed and/or used to address the problem. Conclude with a description of the key results and contributions that will be covered in the presentation (and paper).

Keywords: Cumulative link model; Proportional odds; Latent variable model; Normal regression.

Cluster Analysis for Ordered Categorical Data

Ivy Liu
School of Mathematics and Statistics
Victoria University of Wellington
Wellington, New Zealand
ivy.liu@vuw.ac.nz

Abstract

For a data matrix with ordered categorical data, we establish likelihood-based multivariate methods to reduce the dimension of the matrix. It applies fuzzy clustering via finite mixtures to ordinal response models. Model-fitting is performed using the expectation-maximization algorithm and a fuzzy allocation of rows, columns, and rows and columns simultaneously to corresponding clusters is obtained. This talk gives a summary of recent results of our working group on this topic. It includes the development on various ordinal models, Bayesian approaches, alternative correlation structures, and data visualization methods. The application in real datasets is also shown.

Keywords: Cluster analysis; Ordered categorical data

Dispersion and Response Styles in Ordinal Regression

Gerhard Tutz

Ludwig-Maximilians-Universität, München, Germany– tutz@stat.uni-muenchen.de

Abstract

In ordinal regression the focus is typically on location effects, potential variation in the distribution of the probability mass over response categories referring to stronger or weaker concentration in the middle is mostly ignored. If dispersion effects are present but ignored goodness-of-fit suffers and, more severely, biased estimates of location effects are to be expected. A model specification is proposed that explicitly links varying dispersion to explanatory variables. It is able to explain why frequently some variables are found to have category-specific effects if the cumulative type model is used. For repeated measurements, which are used in survey-based research, dispersion effects may be seen as response styles that represent specific answering patterns of respondents. We consider an extension of the Partial Credit Model that explicitly accounts for response styles. A common problem in partial proportional odds models is the selection of the effect type, each covariate can be equipped with either a simple, global effect or a more flexible and complex effect which is specific to the response categories. A general penalty framework is proposed that allows for an automatic, data-driven selection between global and category-specific effects in all types of ordinal regression models.

Keywords: Ordinal regression; Dispersion; Response styles; Location-shift model.

IPS04:
Where To In Statistical Education Across School, All University Disciplines and Beyond?

The Good, The Bad and Lessons for The Way Forward for Teaching Statistics and Data

Professor Helen MacGillivray
President-elect, International Statistical Institute
Editor, Teaching Statistics
Principal Fellow, Higher Education Academy

Abstract

There are many reasons why the teaching and learning of statistics has produced, and continues to produce, so much debate, literature, research and challenges. The statistical sciences are the sciences of data, variation and uncertainty, and hence work within and across all contexts in which these occur, from the most theoretical to the most applied, and whether they arise in research, industry, business, government or everyday life. Statistics is therefore of importance in education across all levels, namely:

- School: primary, junior secondary, upper secondary
- Tertiary: other disciplines and the training of future professionals
- Workplace: professionals, researchers, managers, consultants
- Society: citizenship

Professional statisticians and information scientists have long been aware of the range of implications of rapidly increasing technological power and of massively increasing amounts of data, and the current more general and widespread awareness of this has given rise to terms such as 'big data', 'data science' and 'data literacy'. Recent descriptions of data literacy are essentially the same as the many and long-standing descriptions of statistcal literacy, and in some cases, even the same as descriptions of the statistical investigation process. Hence, to teach data literacy and data science, there are lessons from the long and many experiences of efforts to teach statistical literacy and statistics.

This presentation looks at some of what has gone right and what has gone wrong in statistics education, and what we can learn for the way forward. A first step is to recognise that there are no magic blueprints or solutions to cover all the complexities, but that progress in all countries requires real understanding, knowledge, skills, commitment and professional collaboration backed by management and government support across school, tertiary, and workplace learning.

Keywords: Teaching statistics; Statistics education

Teaching and Learning Statistics: Lessons from Malaysian Classrooms

Prof. Dr. Mohd Majid Konting
Universiti Putra Malaysia, Serdang MALAYSIA
mmk@upm.edu.my

Abstract

Statistics and statistical literacy are gaining greater importance in the rapidly technologically-based 21st century education and skills then before. The increasingly massive amounts of especially real time big data that are easily accessible and available to support education demands that teaching and learning of statistics in schools, tertiary educational institutions, workplaces and society at large have to be enhanced. This is especially pertinent in the case of Malaysia with the introduction of both school and higher education blueprints in 2013 and 2015 respectively, whereby higher order thinking skills have been promoted towards producing holistic, entrepreneurial and balanced graduates. This paper discusses the current practices in the teaching of statistics and statistical literacy in the Malaysian schools and higher learning institutions to support and improved education especially its investigation process of promoting higher order thinking skills. As statistics and statistical literacy are crucial in national education and development, this paper proposes strategies to enhance teaching for quality statistical learning in Malaysian schools and educational institutions.

Keywords: Statistics education; Teaching statistics; Statistical learning

Teaching Statistics in Indonesian Schools: Today and Future

Muhammad Arif Tiro
Head of Statistics Department,
FMIPA Universitas Negeri Makassar
arif_tiro@yahoo.com
ariftiro2016@gmail.com

Abstract

Statistics education that empowers students to gain life skills needs to emphasize the learning of statistical literacy in building statistical thinking. Statistical thinking is the essence of art owned by statisticians. The desire to inspire students with statistical thinking has led to a resurgence of attention in combining real investigation and education statistics. Thus, statistics education courses that prepare teachers to teach statistics in schools need to be initiated by the statisticians. Teachers who teach literacy of statistics in developing statistical thinking, encourages to active learning by using the data and assessment to improve the quality of students learning outcomes. This presentation will explain reality of teaching statistics at Indonesian schools and perspective of teaching statistics in the future, especially the implication to university programmed.

Keywords: Statistical literacy, Statistical thinking, Life skills, Learning outcomes

CPS01:
Environmental & Natural Resources Statistics

Growth Externalities on the Environmental Quality Index of East Java Indonesia, Spatial Econometrics Model

Rahma Fitriani *
Department of Mathematics University of Brawijaya, Malang, Indonesia – rahmafitriani@ub.ac.id

Wara Alfa Syukrilla
Department of Mathematics University of Brawijaya, Malang, Indonesia – warasyukrilla@ymail.com

Abstract

Environmental quality degradation will be unavoidable in the midst of rapid economic growth. In a regional development context, specifically for East Java Province, Indonesia, the growth in one region might affect the condition in its surrounded regions, including their environmental quality. The main source of the growth externalities is still debatable. It might be the spread of technological information and productivity, population growth, or investment. The tool to analyse the functional relationship between environmental quality and those factors is called an IPAT equation, which is extended to a STIRPAT (Stochastic Impacts by Regression on Population, Affluence and Technology) model. The latter model allows the estimation of the model's parameters followed by the hypothesis testing of their significance. In this study, spatial econometrics models are used to build the STIRPAT model of the East Java's environmental quality as the function of the influential factors, by accommodating the possible externalities. The knowledge regarding the source of externalities and their significance will be useful to promote the development without neglecting the environmental quality. Some models (SDM, SLX and SDEM) are proposed. Data from 38 cities/regencies in East Java at 2014 are used. Among the proposed model, SDEM best decribes the situation. The model indicate that the local as well as the neighbouring regions' population have significant impact on the environmental quality. The increase of local population decreases the local environmental quality, but the opposite holds for the increase of population in the neighbouring regions. The first case is in line with the result of the similar previous studies. The latter one indicates that in this study area, population is the main source of externalities. It also indicates the possibility of labour migration from one region to the neighbouring regions, leaving the origin region with less economic activity, less environmental degradation. However, within the context of SDEM, there is some evidence that the local environmental quality is also positively affected by other unmeasured externalities.

Keywords: Spatial econometrics; STIRPAT; Externalities; Environmental quality.

1. INTRODUCTION

Economic growth is a necessary process in a regional development. Growth of labour force, capital increase and better technology are the causes of the growth. However the process comes with certain consequences for the environmental quality. More industrial activities and residential energy consumption, increase the pollution level and lower the environmental quality. In order to preserve the environmental quality, a region must not necessarily restrict its economic growth. Performing a better technology which minimizes the negative impact on the environment per unit energy consumption would be preferrable. In a regional development context, Ertur and Koch (2007) indicates the phenomenon of interdependence growth among neighbouring regions. They suggest that the economic activities in one region is not only determined by the local socioeconomic characteristics, but also by the characteristics of the nearby regions. This phenomenon implies that the economic activities affect not only the local environmental quality but also the environmental quality of its neighbouring regions. It leads to the situation where the economic growth creates externalities on the environment.

East Java, is one of Indonesia's provinces. Until the third quarter of 2016, it has second largest GDP after Jakarta (Jawa Timur dalam Angka 2016) which contributes 25.28 percent to the national GDP. At the end of the first quarter of 2016, it performs 5.34% of economic growth, which is above the national growth rate of 4.75% (Jawa Timur dalam Angka 2016). In 2014 it is populated by 38,610,202 people (819 people per square km). The substantial economic growth comes at the cost of its environmental quality. In 2014 the environmental quality index of East Java is 56.48% (on the scale of 0% to 100%, from poor to perfect quality), which is below the national quality index of 63.4%. Those numbers are aggregated from 38 cities/regencies in East Java, which are the lower level administrative units. The interraction between cities/regencies depends on their spatial arrangment. Nearby cities/regencies have similar economic activities. Cities/regencies in the northern east of the province, near the capital city (Surabaya) mainly rely on industrial activity. Whereas cities/regencies in the western, eastern and southern part of the province still have active agricultural activity. Therefore, a spatial pattern of economic activities is emerged in this region as well as in other regions. Since the environment quality depends on the economic activity, the spatial pattern of the activities will be mirrored in the spatial pattern of the environmental quality.

The relationship between human activity and environment has been presented in IPAT equation (Ehrlich and Holdren 1971). It is an equation which states that environment performance is affected by Population size, Affluence and Technology ($I = P \times A \times T$). Affluence is commonly represented by consumption or production. It is an identity, in which $T = I/(P \times A)$, such that it cannot be used to measure the effect of each factor on the environment or test its significance. Since it is necessary to analyse the relative importance of each driving force on the environmental performance, the IPAT equation has been modified into a stochastic model, namely STIRPAT (Stochastic Impacts by Regression on Population, Affluence and Technology) (Dietz and Rosa 1997). The model is useful to predict the environmental impacts based on the influential factors, to estimate the magnitude of each of the effect on the impact, and to test the significance of each effect. When the data are available, both can be carried out within the context of econometrics.

Several studies analyse the effect of human economics activities (e.g. population size, population structure, GDP, consumption) on the level of CO2 emission based the STIRPAT model using time series data (Zhu and Peng 2012; Shahbaz, Loganathan et al. 2015) or panel data (Liddle 2013; Aguir Bargaoui, Liouane et al. 2014). Wang, Fang et al. (2016) also use panel data, but they starts to consider the spatiotemporal variations on the CO2 emission in China's 30 provinces over the period 1995 -2011. The latter indicates the presence of the spatial pattern on the environmental impact.

The accommodation the spatial pattern due to the growth externalities in the STIRPAT model, requires an explicit definition of the spatial arrangement between the locations of all cross section units. This type of data is defined as a set of spatial data. Further analysis of the STIRPAT model based on the spatial data is within the scope of spatial econometrics.

The objective of this study is to choose the best model among some spatial econometrics models of the STIRPAT, for the East Java's environmental quality. It is assumed that the environmental quality is a function of the influential factors, by accommodating the possible externalities. The model will be utilize further to discover the main source of externalities by conducting a hypothesis testing regarding its significance. This knowledge will be useful to promote the development without neglecting the environmental quality.

2. GROWTH EXTERNALITIES AND ENVIRONMENT QUALITY, IPAT – STIRPAT MODELS

Some recent regional studies (Fingleton and López-Bazo 2006; Ertur and Koch 2007; Tian, Wang et al. 2010) indicates that the well – being of a regional economy depends on the economic performance and socioeconomic characteristics of its surrounding regions. In other words, the situation in the neighbouring regions affect the regional development and economic growth. It confirms the influence of externalities on the regional growth. In this case, spatial externalities are the region specific economic effects which spread over several nearby regions.

The concept of externalities was originated from the increasing returns to scale assumption of productivity, which is defined in new geographer's theory (Krugman 1991; Fujita, Krugman et al. 1999). Combining it with Jacob's (1970) and MAR's (Marshall – Arrow – Romer) externalities (Glaeser, Kallal et al. 1992), the increasing returns promote the agglomeration of activity because agents receive external benefits when they interact, exchange information or technology. The agglomeration of activity leads to the growth of population. The information or technological exchange, however, cannot be constrained within the barriers of the location of the economic agent. It might spillover from one region to the other, such that the technological progress in one region improves not only the local but also the nearby regions. It implies that the local technology, which is the product of agglomeration of economic activity, is also a function of neighboring technology. Whereas the local technological progress is one of the determinants of the local productivity. Using this argument, the local productivity might also be affected by the productivity of its surrounding regions through their technological progress.

Every region needs growth in its economy. However, the growth comes with some consequences, especially on environment. The relation between the economic productivity and its impact on environment is defined by IPAT identity. It was formed by Ehrlich and Holdren (1971) as an attempt to analyse the effects of human activities on the environment. The identity defines the environment impact (*I*) as a multiplication of population (*P*), affluence (*A*) and technology (*T*). Affluence (*A*) is measured by per capita productivity and Technology (*I*) is in terms of impact per unit production. The following relation holds:

$$I = P \times A \times T = P \times \frac{GDP}{P} \times \frac{I}{GDP} = I \qquad (1)$$

In (1), it is assumed that each factor has proportional effect on the environment impact. However each of them do not affect the environment impact independently of one another. For example, when *A* change, it will also affect *P* and *T*, such that the effect of the change in *I* will not be solely caused by *A*. Thus, a doubling in *A* will not necessarily lead to a doubling of impact *I*. The effect on *I* can be more than double due to the change of other factors as well. Therefore IPAT is reformulated into a stochastic model, namely Stochastic Impacts by Regression on Population, Affluence and Technology (STIRPAT) (York, Rosa et al. 2003). This relation indicates that theoretically, the bigger the value of *P*, *A* and *T*, the greater the effect on the environmental impact. In other words, the increase those factors deteriorates the environment condition.

STIRPAT is not an identity equation. It is a stochastic model which is empirically useful for hypothesis testing. The STIRPAT model is defined as:

$$I_i = aP_i^b A_i^c T_i^d e_i \qquad (2)$$

in which, *a*, *b*, *c*, and *d* are the model's parameters that must be estimated by applying standard statistical techniques, and *e* is the error term. Each model's parameter measures how much the change of each factor affects the environment impact. The parameters estimation will be followed by the hypothesis testing regarding the significance of each factor in determining the environment impact. The definition of STIRPAT will be equivalent to IPAT identity when all the parameters and error term equal to one. The subscript *i* represents a particular observation unit. The observed/measured value of each factor/variable varies across the observation unit. Some studies use the subscript *i* to identify the cross sectional unit, or change the index into *t* when the model will be estimated based on time series of observation (Zhu and Peng 2012; Shahbaz, Loganathan et al. 2015). The cross sectional and time units might be used together, for panel data model's estimation (Liddle 2013; Aguir Bargaoui, Liouane et al. 2014).

When externalities are involved, different specification of the unit of observation is needed. The definition of externalities indicates that the variable value on a location will be affected by the variable value on its surrounding/neighbourhood. Thus, the spatial reference of each observation unit and a spatial arrangement of one unit relative to the others are needed. In that case a set of spatial data is obtained, and the analysis falls within the context of spatial econometric. This study uses STIRPAT in a broader sense to accommodate the growth externalities and other influential factors affecting the environmental quality, based on a set of spatial data.

In the presence of growth externalities, the following relations are assumed, in which $i=1,\ldots,N$, are the index for the location unit and $j \in$ neighbour of $i, j=1,\ldots,N$:

1. Local technology (*T*) is a function of neighboring technology (*T*), which depends on their productivity (*A*):

$$T_i = f(T_j) = f(A_j), \tag{3}$$

2. Local productivity (*A*) is a function of local capital (*K*), local labor (*L*), local technology (*T*) and neighbouring productivity (*A*):

$$A_i = f(K_i, P_i, T_i, A_j) \quad \text{implying that} \quad A_i = f(K_i, K_j, P_i, P_j, A_j) \tag{4}$$

3. Local environment impact (*I*) is a function of local labor/population (*P*), local productivity (*A*) (as defined in equation (4)) and local technology (*T*) (as defined in equation (3)):

Putting them together, local environment impact is a function of local labor/population, neighborhood

$$I_i = f(P_i, A_i, T_i) \tag{5}$$

labor/population, local capital, neighborhood capital, local productivity and neighborhood productivity:

$$I_i = f(P_i, P_j, K_i, K_j, A_{i,} A_j) \tag{6}$$

The inclusion of neighboring capital and neighboring population, is partly from the assumed relation in the local productivity. Apart from the relation in (6), it is also possible to assume that the neighboring environment quality affect the local environment quality, such that:

$$I_i = f(I_j, P_i, P_j, K_i, K_j, A_i, A_j) \tag{7}$$

The two relations in (6) and (7), are the basis to build an environment quality model which accommodates the possible externalities. The significance of at least one of the neighboring conditions (productiviy, population or capital) confirms the effect of externalities on the environment quality.

3. STIRPAT WITH EXTERNALITIES, SPATIAL ECONOMETRICS MODEL

In this study, STIRPAT model which accommodate the externalities will be estimated using the spatial data, within the context of spatial econometrics. A spatial version of STIRPAT will be useful to analyze how is the change of one of the driving forces, on a local level or on the neighboring locations, affecting the local environment impact. The latter effect refers to the externalities, which are the products of interaction between spatially distributed economic agents.

Spatial econometrics is one field in econometrics which treats the spatial interaction properly. The spatial interaction is commonly associated with spatial autocorrelation in regression model. Analysis in that field includes parameter's estimation, and the hypothesis testing of its significance.

In the spatial data, location index is attached to every unit of observation. This would differentiates between the spatial econometrics and other econometrics models. The location index will be used further to define the neighboring units of every location. Analogue to a time lag operator in time series analysis, the spatial model uses a spatial lag operator which is a weighted average of variables under study at 'neighboring' locations. The definition of neighbors is translated into an $n \times n$ spatial weight matrix W which describes the spatial arrangement. The matrix assigns a positive weight for its ij^{th}, $i = 1, \ldots, n$ and $j = 1, \ldots, n$ element, if location j is considered as a neighbor of location i and zero otherwise. It implies that the diagonal elements w_{ii} are zeros for all $i = 1, \ldots, n$. The term 'neighbors' of location i is used to define all locations which are assumed having significant influence on location i. The following concepts are the possible ways of choosing the value of the weight, that is: (i) positive if location i and j share borders on a spatially contiguous plane, 0 otherwise, (ii) decreasing with distance between location i and j, such that the further the distance, the less the weight will be, (iii) increasing with the length of the shared borders, (iv) positive, if location i is inside a certain bandwidth of the n^{th} nearest neighbor of location j, 0 otherwise, (v) decreasing with the rank of the distance between location i and j, (vi) positive if location i is inside a radius of d from location j, 0 otherwise (Elhorst 2014). This study applies the contiguity concept in (i), such that the following definition is used for each element of W:

$$w_{ij} = \frac{c_{ij}}{c_{i.}}, \; c_{ij} = \begin{cases} 1, & \text{if location i and i are shared border} \\ & 0, \; \text{otherwise} \end{cases}, \; c_{i.} = \sum_{j=1}^{n} c_{ij} \tag{8}$$

This matrix is then attached to every variable which is assumed as the source of externalities, it might be the response variable, one or all of the predictors, or the error terms. For a vector variable, says:

$$Y = [Y_1 \quad \dots \quad Y_n]^T$$

The i^{th} element of vector WY defines the spatial lag operator of Y, which is the average of Y_j, $j \in$ neighbourhood of i. When W is attached to the error terms, it is assumed that there are unmeasured predictors which could be the source of externalities.

The assumed relation in (6) and (7) can be tested based on two spatial econometrics models, Spatial Durbin Model (SDM):

$$Y = \rho WY + \beta_0 1_n + X\beta + WX\theta + u = (I_n - \rho W)^{-1}(\beta_0 1_n + X\beta + WX\theta + u) \qquad (9)$$

in which W is attached to the response as well as the predictors. Spatial of Lag X (SLX) is defined when W is attached only to the predictors:

$$Y = \beta_0 1_n + X\beta + WX\theta + u \qquad (10)$$

In (8) and (9), Y is the $n \times 1$ vector of response variable, 1_n is the vector of ones, X is the $n \times k$ matrix of predictors and u is the $n \times 1$ vector of the error terms, which are normally, identically and independently distributed, $u \sim N(0, \sigma^2 I_n)$.

In both models in (8) and (9), if there are some unmeasured variables as the possible sources of externalities, one might model the error terms as follows:

$$u = \lambda W u + \varepsilon, \qquad (11)$$

leading to the use of Spatial Durbin Error Model (SDEM). In those models: $\rho, \beta_0, \beta = [\beta_1 \quad \dots \quad \beta_k]^T, \theta = [\theta_1 \quad \dots \quad \theta_k]^T$ and λ are the model's parameters that should be estimated.

There have been three main estimation methods of the spatial models, maximum likelihood (ML) (Ord 1975), instrumental variables (IV) (Anselin 1988) or generalized method of moments (GMM) (Kelejian and Prucha 1998) and Bayesian Markov Chain Monte Carlo (MCMC) approach (LeSage 2000). This study will focus on ML as the estimation method. The MLEs of the model's parameters are asymptotically unbiased and approximately normal.

4. THE MODEL APPLICATION FOR THE CASE OF EAST JAVA

This study covers all 38 regencies/municipalities in East Java Indonesia. The regency/municipality is the unit of observation. The spatial econometrics models of East Java's environmental quality, based on the STIPAT approach, will be estimated based on each regency/municipality's Environmental Quality Index (I_i in percent), Per Capita GDP (A_i in Rupiah), Population Density (P_i in persons/km^2), Domestic Investment (K_i in Trillion Rupiah), for $i=1,\dots,38$. Data are available from Environment Agency and "Regions in Numbers" by *BPS* (Central Statistics Biro) for the condition in 2014. The assumed relations between variables are defined in equation (7) and (6), as SDM and SLX model respectively. Therefore, in terms of the defined variables, the SDM of East Java's environmental quality is:

$$\begin{aligned} I &= \rho WY + \beta_0 1_n + X\beta + WX\theta + u \\ &= (I_n - \rho W)^{-1}(\beta_0 1_n + X\beta + WX\theta + u) \end{aligned} \qquad (12)$$

and the SLX model is:

$$I = \beta_0 1_n + X\beta + WX\theta + u, \qquad (13)$$

where $I = [I_1 \dots I_{38}]^T$ is the vector of Environment Quality Index of all regencies/municipalities, $X = [A\ P\ K]$, with $A = [A_1 \dots A_{38}]^T$, $P = [P_1 \dots P_{38}]^T$, $K = [K_1 \dots K_{38}]^T$, is the matrix of predictors, β_0 is the model's intercept, $\beta = [\beta_1\ \beta_2\ \beta_3]$ is a vector of coefficients for the non spatially weigted variables, W is the 38 x 38 spatial weight matrix, $\theta = [\theta_1\ \theta_2\ \theta_3]$ is a vector of coefficients for the spatially weighted variables, and ρ

is a coeficient for spatially weighted response variable. When it is necessary to spatially model the error terms, the relation in (10) can be used.

The MLEs for all coefficients of SDM and SLX, and their significance are presented in Table 1. The negative estimated coefficients of the local GDP and Population are in accordance with the assumed theoretical model, in which local productivity and population decrease the environmental quality. Both variables are the indicators of the scale of economic activity in a regency/municipality. The opposite holds for the local investment (K) that has positive estimated coefficient. Even though it is not significant, it is possible that the investment will be useful to advance the technology such that the negative impact on the environment per unit energy consumption would be minimized.

In SDM and SLX, the use of spatial matrix W differentiates them from the non spatial model. In this case, W is attached to every variable which is assumed as the source of externalities. In SDM, it is attached to I (the environmental quality index) and all the predictors, whereas in SLX, only the predictors which are attached to W. In SDM it is assumed that the average environmental quality of the neighboring regencies/municipalities affects the local environmental quality. However the result in Table 1 indicates that neighborhood environmental quality does not create significant effect on the local environmental quality. Since the SLX is nested in SDM, the insignificance of the coefficient of WI (ρ) leads to the use of SLX. By comparing the AIC, SLX produces the least AIC, therefore among the two models, SLX better represents the East Java's environment quality.

So far, SLX has been considered as the better model to describe East Java's environmental condition. However, the residuals of SLX must be checked for possible unexplained source of spatial externalities. The Moran I test is conducted with p value = 0.0152, which indicates that the residuals still have spatial dependencies. The spatial dependence of the residuals signifies the presence of unmeasured factors as the source of externalities. Therefore, instead of only model the explainable terms spatially, the spatial model also applied to the error terms, leading to the use of SDEM. The estimated model's parameters of SDEM are presented in Table 2. The significance of estimated coefficient λ (spatial autocorrelation parameter of the error terms), confirms that it is necessary to model the error terms. In order to ensure that every possible sources of spatial dependence have been well accommodated by the model, the Moran I test is conducted for the residuals of the SDEM. The p value of the test = 0.7594 shows that the residuals are spatially independent. Thus, it is plausible to accept SDEM as the best model.

In all models, based on the estimated coefficient in Table 1 and Table 2, the average GDP and investment of neighboring regencies/municipalities decreases the environmental quality, whereas the average population density of neighboring regencies/municipalites increases the environmental quality. The GDP and investment (capital) create negative externalities, while the opposite holds for population density. The negative externalities from GDP and investment are expected from the theoretical model, since the local environmental quality is also affected by the neighbor economic productivity. The predicted direction of the effect of neighbourhood density indicates the missmatch between labour's residential and labour's workplace. Furthermore for the case of East Java there are unmeasured factors which lead to the positive externalities of the population density.

Table 1. MLE for all parameters of SLX and SDM, and their significance

Variables	Coefficient	SDM		SLX Model	
		Estimated Coefficient	P-value	Estimated Coefficient	P-value
Intercept	β_0	67.3	0.000000**	71.65	0.000000**
A (GDP per capita)	β_1	-0.000024	0.361267	-0.000030	0.324301
P (Population density)	β_2	-0.001934	0.004602**	-0.002408	0,001625**
K (Domestic Investment)	β_3	0.274	0.486575	0.325	0.476704
$\boldsymbol{WA}$ (Average of GDP of neighboring regencies/municipalities)	θ_1	-0.000065	0.029611**	-0.000058	0.096039*
$\boldsymbol{WP}$ (average of population density of neighboring regencies/municipalities)	θ_2	0.000348	0.719400	0.001338	0.133533
$\boldsymbol{WK}$ (Average of Domestic Investment of neighboring regencies/municipalities)	θ_3	-0.65	0.295051	-0.78	0.277365
$\boldsymbol{WI}$ (Average of Environmental Quality Index of neighboring regencies/municipalities)	ρ	0.079	0.123572	NA	NA
AIC		268,64		71.64901	

(**) the estimated coefficient is significant at any level of α,
(*) the estimated coefficient is significant at α = 10%

According to the best model (SDEM), local population density significantly affects the environmental quality at any level of significance. The average of population density of of neighboring regencies/ municipalities also create significant impact on the environmental quality, at α = 10%. The result indicates that the source of externalities is the population density or labour force.

Table 2. MLE for all parameters of SDEM, and their significance

Variables	Coefficient	Estimated Coefficient	*p-value*
Intercept	β_0	70.88	0.0000000**
A (GDP per capita)	β_1	-0.0000192	0.4904666
P (Population density)	β_2	-0.0023329	0.0002572**
K (Domestic Investment)	β_3	0.1615300	0.6363736
$\boldsymbol{WA}$ (Average of GDP of neighboring regencies/municipalities)	θ_1	-0.0000453	0.1033785
$\boldsymbol{WP}$ (average of population density of neighboring regencies/municipalities)	θ_2	0.0013348	0.0708089*
$\boldsymbol{WK}$ (Average of Domestic Investment of neighboring regencies/municipalities)	θ_3	-0.824010	0.1644801
$\boldsymbol{Wu}$	λ	0.4586200	0.0225860**

(**) the estimated coefficient is significant at any level of α,
(*) the estimated coefficient is significant at α = 10%

5. CONCLUDING REMARKS

This study modifies the crossection or time series based STIRPAT model for the East Java's environmental quality into the spatial version, using spatial data to estimate the model's coefficients. Three spatial econometrics models, SDM, SLX and SDEM are assumed to accommodate the possible source of growth externalities. The analysis indicates that SDEM is better explain the condition in East Java. In that model, local population/labour and neighborhood populatio/labour give significant effects on local environmental quality. Thus any policy aiming at controlling the negative effect on local environmental condition must focus on controlling those factors. For East Java's condition, neighborhood population density is the main source of externalities.

To reduce the negative effect of local population or neighborhood population on local environmental condition, the local authority should also take into account the local investment, since it potentially reduces the negative effect on environment. In this case investment must be directed to enhance the efficient production technology and waste management. The local authority should also cooperate with the authorities of its neighboring regencies/municipalities to deter the negative effect of externalities. And lastly, within the context of SDEM, there is some evidence that the local environmental quality is also positively affected by other unmeasured externalities.

REFERENCES

Aguir Bargaoui, S., N. Liouane, et al. (2014). "Environmental Impact Determinants: An Empirical Analysis based on the STIRPAT Model." Procedia - Social and Behavioral Sciences **109**: 449-458.

Anselin, L. (1988). Spatial econometrics : methods and models. Dordrecht ; Boston, Kluwer Academic Publishers.

Dietz, T. and E. A. Rosa (1997). "Environmental impacts of population and consumption." Environmentally significant consumption: Research directions: 92-99.

Ehrlich, P. R. and J. P. Holdren (1971). "Impact of population growth." Science, New Series **171**(3977): 1212 - 1217.

Elhorst, J. P. (2014). Spatial Econometrics: From Cross-Sectional Data to Spatial Panels. Dordrecht, Springer Briefs in Regional Science.

Ertur, C. and W. Koch (2007). "Growth, technological interdependence and spatial externalities: theory and evidence." Journal of Applied Econometrics **22**(6): 1033-1062.

Fingleton, B. and E. López-Bazo (2006). "Empirical growth models with spatial effects*." Papers in Regional Science **85**(2): 177-198.

Fujita, M., P. R. Krugman, et al. (1999). The spatial economy : cities, regions and international trade. Cambridge, Mass. ; London, MIT Press.

Glaeser, E. L., H. D. Kallal, et al. (1992). "Growth in Cities." Journal of Political Economy **100**(6).

Jacobs, J. (1970). "The economy of cities." The economy of cities.

Jawa Timur dalam Angka (2016). "http://jatim.bps.go.id/."

Statistical Analysis for NDVI Trend and Variation Using MODIS Data in the Cloud Forest of Khao Nan National Park, Thailand during 2000-2015

Anusa Suwanwong[1*]
[1]Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Songkhla, 90110, Thailand. – 5710220115@email.psu.ac.th

Noodchanath Kongchouy[2]
[2]Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Songkhla, 90110, Thailand. – nootchanath.k@psu.ac.th

Attachai Ueranantasun[3]
[3]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Pattani, 94000, Thailand. – attachai.u@psu.ac.th

Abstract

For satellite imaging data, Normalized Difference Vegetation Index (NDVI) is an important vegetation indicator that can be used for distinguishing plants in the area by considering a signal range, while Land Surface Temperature (LST) is a radiative skin temperature that looks through the atmosphere to the ground. Both data have been used several times in climatological and environmental studies, especially for natural preservation purpose. One of the valuable areas on earth that needs to be protected is a cloud forest. It is a rare resource, acting as a vital water source, and a concentration of biodiversity. In Thailand, a cloud forest can be found in Khao Nan National Park, south of the country. To provide information on vegetation in this area for helping preservation of its valuable natural resources, this study focused on using LST to examine the missing data of NDVI and then analyzed those data with weighted regression, cubic spline regression, and GEE to perceive patterns and variations of NDVI at Khao Nan during 2000-2015. The data were NDVI and LST recorded by MODIS, an imaging apparatus installed on NASA satellites, between 2000 and 2015. The results indicated 57 NDVI observations corresponding to missing LST day temperatures. Weighted regression showed that 37 observations were considered unreliable data. Cubic spline regression showed that the patterns of NDVI were constant around the central pixel. The result from GEE suggested a continuous decrease trend for NDVI at the region over 15 year's period, but it was not conclusive.

Keywords: NDVI; Cubic Spline Regression; Weighted Regression; GEE.
JEL Classification: Q51

1. INTRODUCTION

Normalized Difference Vegetation Index (NDVI) is a numerical indicator or an index of plant "greenness" or photosynthetic activity. It is used in several studies to investigate the evolution of the ecological status, the measurements of vegetation growth, the degree of cover, and the biomass (Zhang *et al.*, 2009). Lead & Yu (2014) suggested that Land Surface Temperature (LST) is one of the important indicators of the earth surface energy budget which is widely required in hydrology, meteorology, and climatology applications. The information of both indicators is provided by Moderate Resolution Imaging Spectroradiometer (MODIS). MODIS is installed in both Terra and Aqua satellites which are designed for monitoring the earth with spatial resolutions of 250 m^2 to 1 km^2. Furthermore, it can be used to substitute meteorological and environmental data for many purposes, especially when the area of interest is a complex terrain such as mountains (Neteler, 2010).

The cloud forests are immensely valuable areas that represent rare resources, water sources, and varieties of biodiversity. There are about 380,000 km^2 of the cloud forests or 0.26 percent of the world's forest areas. The potential cloud forest mostly occurs in Asia, Americas, and Africa (Philip Bubb *et al.*, 2014). In Asia, the spatial distribution of cloud forests is widely found in Indonesia, Malaysia, and Thailand by considering the forest areas with an altitude of 1,400 meters above sea level (Mulligan & Burke, 2005). Khao Nan National Park is one of the cloud forests in southern Thailand. Most of the area consists of productive rainforest which causes heavy continuous rains. In addition, the weather of the area has an effect on animal diversity. Khao Nan has an abundance of biodiversity and some of them are rare to find. As an example, there were two recorded species of Pteridophyte which have been found only once at one location of Khao Nan and in rather small numbers (Boonkerd *et al.*, 2008). Kittipanangkul & Ngamriabsakul (2011) found that the Zingiberaceae diversity of 29 species among this area was mostly distributed at an altitude of 90-300 m and the number of species decreased as the altitude increases. Therefore, a study of the vegetation change for cloud forest areas is beneficial to the preservation of the areas and their habitants.

In previous studies, time series analysis was frequently used to investigate trends and variations of NDVI. The Theil–Sen (TS) median slope method for time series showed the NDVI trend increased by 0.46 x 10^{-3} per year from 1982 to 2012 globally with decadal variations (Liu *et al.*, 2015). In China between 1982 and 1999, NDVI showed the largest increase of the total study area in spring and the smallest increase of the area in summer. The NDVI trends also showed a marked heterogeneity corresponding to regional and seasonal variations in climates by using time series (Piao *et al.*, 2003). Vaiphasa *et al.* (2011) suggested that the NDVI time series of the Idle Agriculture Land (IAL) had the values close to zero in Khonkean, Thailand, during 2003-2005 using data from LANDSAT satellite images. Another method, widely used to analyze vegetation change, is regression analysis (Eklundh & Olsson, 2003; Herrmann *et al.*, 2005). Linear regression analysis showed that the greatest increasing trends in NDVI were on re-vegetating abandoned gravel pads, old road material excavations, and stabilized river terraces in the area of the Northern Foothills of the Brooks Range, Alaska (Raynolds *et al.*, 2013). In central Asia, the overall trends of NDVI obviously differed before and after 1992. The least absolute deviation (LAD) in linear regression model showed trend of NDVI increase for most areas before 1991, but experienced a dramatic decrease in 1992-2011 as (Zhou *et al.*, 2015). Cubic spline regression has been various used in studies of demographics, and animal and plant sciences (White & Brotherstone, 1997; McNeil *et al.*, 2011). Weighted regression is one of popular methods to be mostly used in geographical and environment (Huang *et al.*, 2010), while GEE is commonly used in large biological studies for application of statistical analysis (Xie & Paik, 1997). Both of the techniques have been used to deal with spatial correlation and other correlations of the data.

The objective of this study was to identify the possible NDVI change of Khao Nan by combining the techniques of examining the doubtful data of NDVI by LST, and then applying weighted regression, cubic spline regression, and GEE to perceive patterns and variations of NDVI at Khao Nan. The study area covered parts of Khao Nan National Park and the study data ranged from 2000 to 2015.

2. MATERIALS AND METHODS

2.1. Study area

The study area was a part of Khao Nan National Park, of which the central area started from 8.491667'N to 8.50000'N and 99.642344'E to 99.650770'E, and had a total land area of 441 km^2. The area was equally divided into 81 pixels, covering the area of 250 x 250 m^2 each. Therefore, the total pixels in this study were 6561.

Figure 1. Map of Khao Nan National Park

2.2. Data and Definitions

The satellite data used in this study were Aqua/MODIS data from 2000 to 2015.Two MODIS land products, NDVI or 16-day 250-m land surface reflectance product (MOD13Q1), and LST or 8-day 1-km day and night land surface temperature product (MOD11A2), were collected from Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). Vermote & Vermeulen (1999) presented that MOD13Q1 provided land surface reflectance from seven spectral bands (red, NIR1, blue, green, NIR2, SWIR1, SWIR2), which were atmospherically corrected using MODIS atmospheric products. An area within a resolution of 250 m^2 for NDVI or 1 km^2 for LST, respectively, is called a pixel.

NDVI is probably the most common of these ratio indices for vegetation or measure of greenness. Tucker (1979) defined that it is calculated using the following equation:

$$NDVI = (\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red}) \quad (1)$$

Where ρ_{NIR} and ρ_{red} are the reflectance of near infrared (NIR) and red bands, respectively. NDVI values are represented as a ratio ranging value from -1 to 1 but the negative values represent water, values around zero represent bare soil and values over .6 represent dense green vegetation.

LST is the radiative skin temperature of ground. It depends on the albedo, the vegetation cover, and the soil moisture. Liang (2013) suggested that LST influenced the partition of energy between ground and vegetation, and determined the surface air temperature. Only land surface day temperatures were used in this study.

2.3. Preliminary data analysis of NDVI and LST

Considering for each pixel, for NDVI, the data values for a specified period was created by selecting the day from the 16-day period when NDVI presented maximum measurement quality. There was 23 different days from 1 to 353 for each year and, in turn, resulted in 345 observations over 15 years. Since there were fewer than 16 days remaining in the year from the day 353, the first few days in the following year were used both for the last period of one year and the first period in the following year. The data values were coded as NA when no measurement was recorded at any pixel. The coded NA indicated that no observation in the region in the 16-day period had sufficient measurement quality. For all pixels, the missing data of NDVI were found to be 0.87% during 2000-2015.

LST values were stored as Kelvin. There were 46 different days from day 1 to 361 every year. In one recording location, the total number of observations was 690 over 15 years. Also, these values were coded as NA, when no measurement was recorded at any location. For the whole study area, the missing data of LST were 1.16% during 15 years.

LST data were used to compare with NDVI data to check whether NDVI data were correctly recorded. Therefore, only LST data at the same days with NDVI data were selected in this study. As a result, the number of both NDVI and LST were finally 345 observations. The method to find incorrectly recorded

NDVI data was to mark any NDVI, recorded in the same day as LST with NA code, as a doubtful value. The reason for this was that these NDVI values could not be possible because they were measured with the same MODIS as LST values. At the same day, if an LST value was deemed NA, the corresponding NDVI value should be NA as well.

2.4. Statistical Analysis

Weighted regression is an efficient method that reflects the behavior of the random errors in the model; and it can be used with functions that are either linear or nonlinear in the parameters. It works by incorporating extra nonnegative constants, or weights, associated with each data point, into the fitting criterion (Cressie, 1985). For this study, the data were fitted with weighted regression to identify unreliable data points and they were subsequently removed.

Spline function is one of popular approximating functions in mathematics. Spline function is defined as piecewise polynomials of degree n. Knots and the first n-1 derivatives are condition of this function (Wold, 1974). In this study, the data were fitted by spline function by connecting the end of any year with the beginning of the next year. The result from fitted spline model was used as a seasonal pattern of the data. The formula for a cubic spline function is

$$s(t) = a + bt + \sum_{k=1}^{p} c_k (t - t_k)_+^3 \tag{2}$$

Where t denotes time, $t_1 < t_2 < \ldots < t_p$ are specified knots, and $(t - x)_+$ is $t - x$ for $t > x$ and 0 otherwise.

From (2), The boundary conditions require that s(t) for $t < t_1$ equals s(t) for $t > t_p$. One way of ensuring this is to put b = 0 and make $\sum_{k=1}^{p} c_k (t - t_k)_+^3$ vanish for $t > t_p$, so that $\sum_{k=1}^{p} c_k = 0$, $\sum_{k=1}^{p} c_k t_k = 0$, $\sum_{k=1}^{p} c_k t_k^2 = 0$, and $\sum_{k=1}^{p} c_k t_k^3 = 0$.

Generalized Estimating Equation (GEE) is a common statistical method to fit a marginal model for longitudinal and other correlated response data and it provides the population-averaged estimates of the parameters (Wang, 2014). For this work, GEE was aimed to handle spatial correlation between pixels after the doubtful and unreliable data were removed and the seasonal patterns were subtracted from the data. To give more details over the period of study, the NDVI data were separated into three periods of five years: 2000-2004, 2005-2009, and 2010-2015. GEE was then applied to each part to find a change in NDVI and its corresponding confidence interval (CI) for each period.

All methods were analyzed by R program version 3.22. GEE was also implemented in R statistical software, using 'geepack' package (Venables & Smith, 2016).

3. RESULTS AND DISCUSSION

For an example of all pixels, the central pixel (pixel 3281) is selected for displaying the results. Figure 2 shows the pattern for NDVI in a 250 x 250 m^2 pixel around latitude 8.4968 and longitude 99.645 for the 15-year period (mid-June 2000 to mid-June 2015) at the central pixel. It can be seen that some measurements of NDVI vary greatly in a short period of time. For example, the value of 0.17 on 289 day, in a red circle in Figure 2(a), is seen to jump to much higher values within just 16 days. Even though vegetation can be reduced in such a short time by burning or cleaning forest, it is impossible that it can be increased with this rate. Therefore, this value is doubtful and unreliable. To identify doubtful data, the NDVI data were then compared with LST data. The result in Figure 2(b) shows the blue points denoting 57 NDVI observations corresponding to missing LST day temperatures. However, some of error data still remain at low values.

A lot of observed NDVI values are still implausibly low as mentioned before. Thus, the method to use is weighted regression model that eliminates negative errors. By assuming that NDVI data values are more likely to be reduced during the rainy season when there are more water vapors and cloud in the sky. In the south-east Thailand, where Khao Nan is located, the rainy season is from early October at day 275 to mid-February at day 40. Our rule is that errors are given zero weight if they are negative and below

-0.02 during the rainy season or greater than 0.2 in magnitude at other times. With these threshold values, further 37 observations are identified as unreliable data and with a crossing sign in Figure 2(b). Both values.

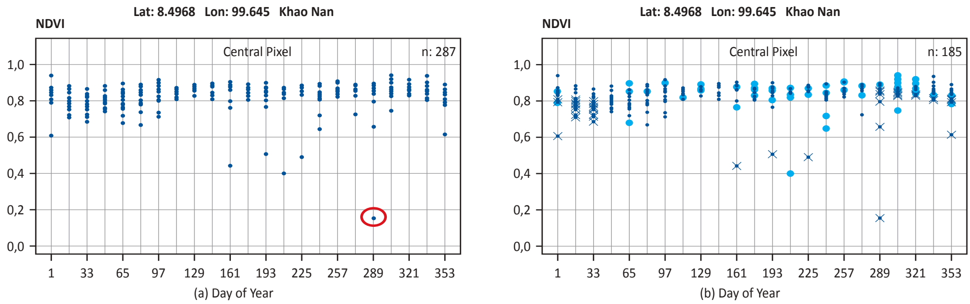


Figure 2 (a). The patterns NDVI for 15 years at central pixel, and (b) The NDVI data with marked doubtful and unreliable data

To find a seasonal pattern, cubic spline regression was applied by using six knots at days 10, 30, 120, 160, 200, and 350 indicated in blue crosses at the bottom of Figure 3(a). The model should be a smooth periodic function with boundary conditions.

The result from the model for the central pixel is shown in Figure 3(a) in a red line. The seasonal patterns from cubic spline functions were then subtracted from the data. Cubic spline regression fitted quite well on the data. For presentation purpose, only four vegetation pixels at the central location (25) were selected to show NDVI trends. These sample pixels were with IDs of 3034, 3040, 3522, and 3528. Figure 3(b) shows variation in NDVI trends for the central location (25). It is shown that the patterns are constant and local variation is slight around the central pixel.

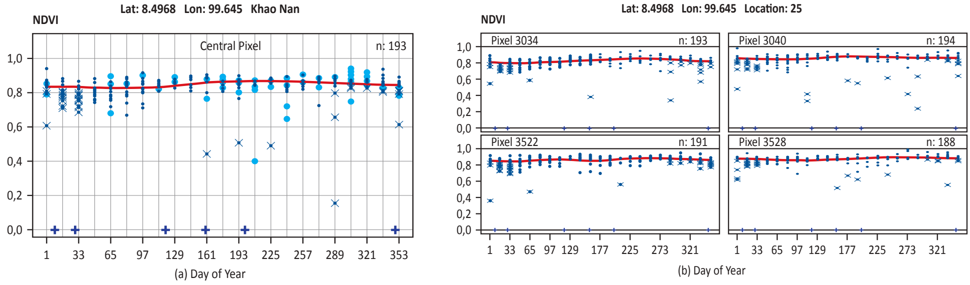


Figure 3 (a). The fitted spline regression with six knots for NDVI data, and (b) Local variation in NDVI Trends around the center pixel

The seasonally adjusted data were applied with GEE to handle spatial correlation and find a change for NDVI for three divided periods. As a result, the changes in NDVI for 2000-2004, 2005-2009, and 2010-2015, and their corresponding CIs are shown in Figure 4. It is shown that the change rates of NDVI in Khao Nan had been declined throughout the study period. However, the changes are inconclusive because the CIs of them are crossing the zero rate.

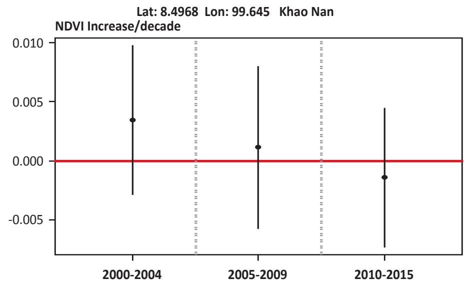


Figure 4. NDVI change patterns in 2000-2004, 2005-2009, and 2010-2015 with corresponding confidence interval

4. CONCLUSION

It is evident that the combination of techniques including comparison of data, weighted regression, cubic spline regression, and GEE can be used as a tool to examine the remote NDVI data from the satellite. These techniques have handled doubtful and unreliable data, seasonal pattern and spatial correlation, and the patterns and variation of NDVI for the study period can be determined. Even though, the change in NDVI from 2000 to 2015 cannot be conclusive and rather considered constant, the importance of preserving the cloud forests in Khao Nan National Park cannot be ignored. There is a hint of indication that the changing rate starts to be in a decreasing pattern. The study with more future data for the same area is useful to determine if the change could become significant.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Don McNeil for his comments and suggestions. This study was funded by the National Research Council of Thailand (Year 2016).

REFERENCES

Boonkerd, T., Chantanaorrapint, S., & Khwaiphan, W. (2008). Pteridophyte diversity in the tropical lowland rainforest of Khao Nan national park, Nakhon Si Thammarat province, Thailand. *Natural History Journal of Chulalongkorn University*, **8**, 83-97.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *J. Int. Assoc. Math. Geol.,* **17**, 563-586.

Eklundh, L., & Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *Geophys. Res. Lett*., **30**, DOI: 10.1029/2002GL016772.

Herrmann, S. M., Anyamba, A., & Tucker, C. J. (2005). Recent trends in vegetation dynamics in the African Sahel and their relationship to climate. *Global. Environ. Change*, **15,** 394-404.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int. J. Geogr. Inform. Sci*., **24**, 383-401.

Kittipanangkul, N., & Ngamriabsakul, C. (2011). Zingiberaceae diversity in Khao Nan and Khao Luang National Parks, Nakhon Si Thammarat, Thailand. *Walailak J.Sci. & Tech*., **5**, 17-27.

Lead, T., & Yu, Y. (2014). Land surface temperature (LST). https://www.star.nesdis.noaa.gov/jpss/lst.php. Accessed 5 December 2016.

Liang, S. Li, X. (2013). *Land Surface Observation, Modeling and Data Assimilation.* Singapore: World Scientific.

Liu, Y., Li, Y., Li, S., & Motesharrei, S. (2015). Spatial and temporal patterns of global NDVI trends: correlations with climate and human factors. *Rem. Sens*., **7**, 13233-13250.

McNeil, N., Odton, P., & Ueranantasun, A. (2011). Spline interpolation of demographic data revisited. *Sonklanakarin Journal of Science and Technology*, **33**, 117-120.

Mulligan, M., & Burke, S. M. (2005). Global cloud forests and environmental change in a hydrological context. http://www.ambiotek.com/cloudforests/cloudforest_finalrep.pdf. Accessed 15 December 2016.

Neteler, M. (2010). Estimating daily land surface temperatures in mountainous environments by reconstructed MODIS LST data. *Rem. Sens*., **2**, 333-351.

Philip Bubb, P. Ian May,L. Miles,L & Sayer J. (2014). Cloud forest agenda. http://www.ourplanet.com/wcmc/pdfs/cloudforests.pdf. Acessed 7 December 2016.

Piao, S., Fang, J., Zhou, L., Guo, Q., Henderson, M., Ji, W., & Tao, S. (2003). Interannual variations of monthly and seasonal normalized difference vegetation index (NDVI) in China from 1982 to 1999. *J. Geophys. Res. Atmos*., **108**, DOI: 10.1029/2002JD002848.

Raynolds, M. K., Walker, D. A., Verbyla, D., & Munger, C. A. (2013). Patterns of change within a tundra landscape: 22-year Landsat NDVI trends in an area of the northern foothills of the Brooks Range, Alaska. *Arctic. Antarct. Alpine. Res*., **45**, 249-260.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Rem. Sens. Environ.,* **8**, 127–150.

Vaiphasa, C., Piamduaytham, S., Vaiphasa, T., & Skidmore, A. K. (2011). A normalized difference vegetation index (NDVI) time-series of idle agriculture lands: a preliminary study. *Eng J*., **15**, 9-16.

Venables, W.N. & Smith, D.M. (2016). The R development core team an introduction to R: notes on R: a programming environment for data analysis and graphics Version 3.22. https://cran.r-project.org/doc/manuals/R- intro.pdf. Accessed 14 January 2017.

Vermote, E. F., & Vermeulen, A. (1999). Atmospheric correction algorithm: spectral reflectances (MOD09). http://dratmos.geog.umd.edu/files/pdf/atbd_mod09.pdf. Accessed 14 January 2017.

Wang, M. (2014). Generalized estimating equations in longitudinal data analysis: a review and

recent developments. *Advances in Statistics*, **2014**, 1-11.

White, I., & Brotherstone, S. (1997). Modelling lactation curves with cubic splines. *Interbull Bulletin*, **16**, 80-82.

Wold, S. (1974). Spline functions in data analysis. *Technometrics,* **16**, 1-11.

Xie, F., & Paik, M. C. (1997). Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics,* **53**, 1458-1466.

Zhang, X., Hu, Y., Zhuang, D., Qi, Y., & Ma, X. (2009). NDVI spatial pattern and its differentiation on the Mongolian Plateau. *J. Geogr. Sci*., **19**, 403-415.

Zhou, Y., Zhang, L., Fensholt, R., Wang, K., Vitkovskaya, I., & Tian, F. (2015). Climate contributions to vegetation variations in central Asian drylands: pre-and post-USSR collapse. *Rem. Sens*., **7**, 2449-2470.

Improved the Forecasting of ANN-ARIMA Model Performance: A Case Study of Water Quality at the Offshore Kuala Terengganu, Terengganu, Malaysia

Muhamad Safiih Lola[1]
School of Informatics and Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu, Malaysia- safiihmd@umt.edu.my
Mohd Noor Afiq Ramlee[1]
School of Informatics and Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu, Malaysia- afiqramlee91@gmail.com
Vigneswary a/p Ponniah[1]
School of Informatics and Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu, Malaysia- vignes2538@yahoo.com
Nurul Hila Zainuddin[1]
School of Informatics and Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu, Malaysia- Hila.zainuddin@gmail
Razak Zakariya[2]
School of Science Marin, University Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia-ajak@umt.edu.my
Md Suffian Idris[2]
School of Science Marin, University Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia-suffian@umt.edu.my
Idham Khalil[2]
School of Science Marin, University Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia- idham@umt.edu.my

Abstract

Researches obtain motivation to produce innovative models for managing water resources after precise predictions of time series have been made. Water quality time series data are complex in nature causing them to be difficult to predict. Hence, no single models of ARIMA and ANN are able to handle both the linear and nonlinear relationship well. In this study, a hybrid method that is able to utilize the advantages of time series approaches and also artificial neural networks. Findings revealed that that hybrid models water temperature, dissolved oxygen, pH and salinity provides much better prediction as compared to traditional model. Hence, hybrid approach can be an effective way in predicting water quality time series compared than using available models separately.

Keywords: Artificial Neural Network, Water Quality, ARIMA, Hybrid Models

[1]School of Informatics and Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu
[2]School of Science Marin, University Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu

1. INTRODUCTION

Numerous research have been done using time series model, Seasonal Autoregressive Integrated Moving Average (SARIMA), Artificial Neural Network (ANN) and Autoregressive Integrated Moving Average (ARIMA). However, one of the major weakness of ARIMA is the time series which will be generated are from linear component. It has difficulties in capturing the nonlinear component. On the other hand, Artificial Neural Network (ANN) are of nonlinear nature is influenced by the behavior of neurons in them. It can approximate the function to a satisfying level of accuracy. Hence, a hybrid model combining ARIMA and neural network back propagation model is proposed. The use of hybrid models in water quality time series data could be an added advantage in capturing patterns of data sets and could improve the prediction accuracy. The motivation behind this hybrid approach is mainly due to the reason that water quality are real data sets which are complex and any single model approaches would not be sufficient to determine the patterns well. In this study, we will able to predict the water quality time series from the developed hybrid model, NNARIMA and evaluate its performance. (Zhang, 2003).

2. MATERIALS AND METHODS

Study Area and Water Quality Data

This study was carried out around the coast of the South China Sea in the area of Kuala Terengganu, Kampung Marang, Kampung Setiu and Kuala Besut. The data that involved in this study are In-situ in 2015 (30th April to 3rd May). Figure 1 shows the research area in Terengganu in which the study was done. A 126 data was collected from 26 sampling stations at different depths.

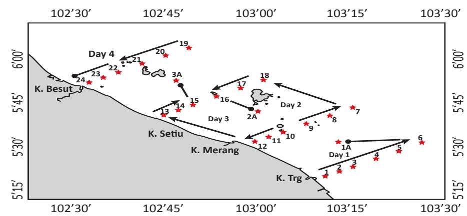


Figure 1: Research area at offshore Kuala Terengganu, Terengganu, Malaysia

ARIMA Modelling Approach

ARIMA model is formed only when the series is not stationary. The model is represented by a general term ARIMA (p, d, q) as follows:

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where p and q are the number of autoregressive terms and the number of lagged forecast errors in the prediction equation, respectively. The number of p, d and q are obtained by looking at the ACF and PACF plots. The ARIMA modelling approach consist of three steps: model identification, parameter estimation and diagnostic checking. The model identification compromises of two steps that is determining whether the series are stationary and also the examining the ACF and PACF functions. The model with the minimum Akaike's Crriterion is chosen as the best fit model.

Artificial Neural Network Modeling

One of the advantages of the neural network model compared to other nonlinear model is neural network is universal estimators that can emulate the class of functions with a high level of accuracy (Zhang et al., 1998). The strength of their estimation is a parallel information processing with data. The initial assumption is not necessary to establish this model when the model building process. Instead, the network model is largely determined by the characteristics of the data. A feed forward circuit (feed forward) are among the hidden layer model that is widely used to model time series and forecasting. This model is characterized by three mobile network layer of simple processing units linked by a series of a cyclic. The relationship between output ( ) and input ( ) has the following mathematical representation (Khashei and Bijari, 2011):

$$y_t = w_o + \sum_{j=1}^{q} w_j . g(w_o + \sum_{i=1}^{p} w_{ij} . x_{t,i}) + \varepsilon_t , \quad (2)$$

whereby (i=0,1,2,…,p) and (j=0,1,2,…,q) are parameter of the model which also called as connection weights, p, q, and and g are the number of input nodes, the number of hidden nodes, error term, weights of the arcs leaving from the bias terms and sigmoid equation, respectively.

Activation functions consist of a few forms and are represented by the condition of neurons in the network (Khashei and Bijari, 2011) as follows:

$$g(x) = \frac{1}{1+e^{(-x)}} \quad (3)$$

$$Tanh(x) = \frac{1 - e^{(-2x)}}{1 + e^{(-2x)}} \quad (4)$$

Normally, an artificial neural networks model in Eqs. (1) perform in terms of nonlinear functional mapping from the past observations i.e., to the future values of , which are

$$y_t = f\left(y_{t-1}, y_{t-2} \cdots y_{t-p}, \alpha\right) + \epsilon_t \tag{5}$$

where, α is a vector for all parameters, f(.)is a function determined by the network structure and connection weights and is error term.Then,the general structure of neural network as shown in Figure 2.

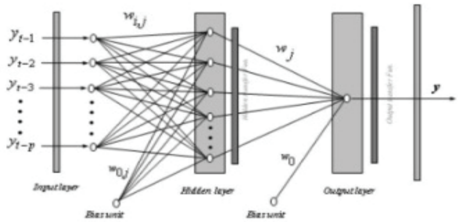


Figure 2: General structure of neural network

Hybrid Model of NNARIMA

In order to produce more general model, a linear hybrid model and more accurate nonlinear is produced naming as neural network autoregressive integrated moving average (NNARIMA) hybrid model. This model, the time series is also considered as a function of linear and nonlinear components.

$$y_t = l_t + n_t \tag{6}$$

where and are linear and nonlinear components respectively. In the first stage, the main objective is to obtain a linear model. Therefore, ARIMA model is used to model linear components. Error from the first stage which contains nonlinear relationship or could also be linear relationships which linear models could not capture or solve (Kashei and Bijari, 2011) here, we represent the error at time t as

$$\epsilon_t = y_t - \hat{y}_t \tag{7}$$

The values of the predictable and linear modelling error is the result of the first stage of which will be used in the next stage. In addition, the linear trend magnified by the ARIMA model to be used in the second stage. In the second stage, the main focus is nonlinear model. Thus, the multi-layer perceptron is used to model the nonlinear relationship and the possibility of simultaneous linear model which remains in error linear models and linear and nonlinear relationships in the original data. Therefore, errors can be modelled using neural network to identify the nonlinear relationship. With n input nodes, neural network model for the error is as follows:

$$\epsilon_t = f\left(\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-n}\right) + \omega_t \tag{8}$$

where f is nonlinear function which is dependent on the neural network and is random variable. Then the combined forecasting model would be

$$yF_t = lF_t + nF_t \tag{9}$$

3. COMPARISON OF ARIMA, ANN AND HYBRID MODEL OF NNARIMA

Both linear and non-linear model was used to set data, although more or less linearity have been found in this series. Only one step ahead predictions considered. Two key performance indicators including MAE (mean absolute error) and RMSE (mean square error), which is calculated from the following equation, will be used to measure the performance of the predicted models. RMSE checks the overall performance of the model while MAE evaluates the model.

$$\text{MAE} = \frac{\left|\sum_{i=1}^{n} y_i - yF_t\right|}{n} \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_t - yF_t)^2}{n}} \qquad (11)$$

4. RESULTS AND DISCUSSION

ARIMA Modelling

In this study, several steps are made to choose the ideal ARIMA model parameters. The model parameters that satisfy the residuals diagnostic checking. In the identification stage, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to study the stationary of the data and to determine the possible best fit models. The best fit model then has been determined by using the Akaike's Criterion (AIC) for all the parameters that is water temperature, pH, salinity and dissolved oxygen. The models were then checked for adequacy by analyzing the independence of the residuals.

Table 1: Best Fit Model for all Parameters

Parameters	Type of Model	MSE	AIC
Water Temperature	ARIMA (1,1,1)	1.364	0.3422
pH	ARIMA (2,1,2)	0.03571	-3.2688
Salinity	ARIMA (0,1,2)	189.2	5.2749
Dissolved Oxygen(DO)	ARIMA (1,1,1)	19.20	2.9871

(a)

(b)

(c)

(d)

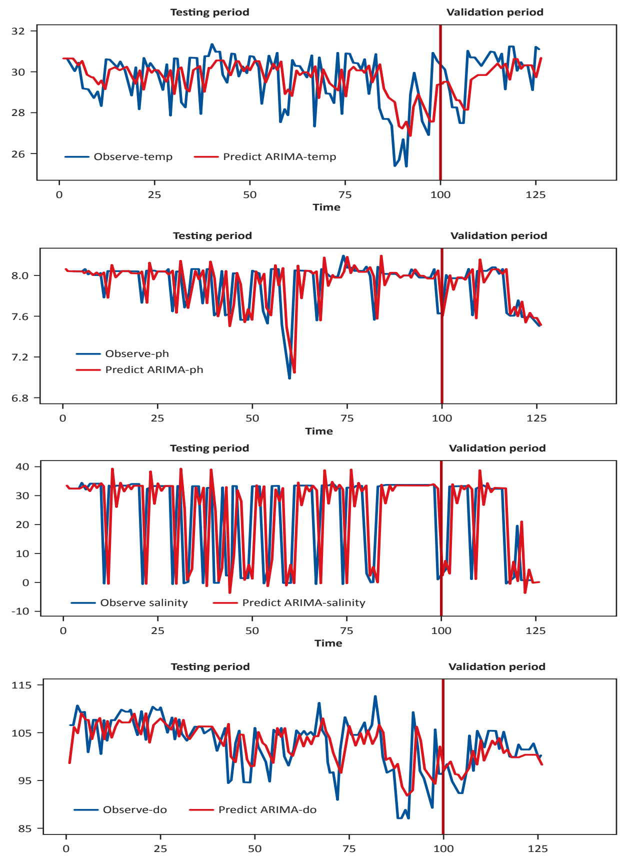


Figure 3: Observed versus ARIMA model predicted data for each water quality parameters; (a) temperature, (b) pH, (c) salinity, (d) DO

A suitable model to predict water quality time series were built using ARIMA. As shown Figure 3 although ARIMA models vary with the range, the model predictions are not adequate. This is due to the limitation of the linear modelling algorithm in ARIMA model which resulting in unsatisfactory in identifying and predicting nonlinear time series of water quality data.

ANN modelling approach

A neural network was developed to predict optimal model to predict water quality time series whereby water temperature, pH, and salinity and dissolved oxygen were used as the input data. The target would be changed simultaneously accordingly. For an example, if we need to predict water temperature, water temperature is the target and the other parameters would be set as the input data. There are 3 partition in neural network model which comprises of training testing and validation. During training the inputs data will be selected in the network and it will customize itself based on the error contain in the model. Next step is testing which is independent. Lastly, validation is used to measure the network ability to generalize and is the stopping criteria for training sample. The total data used in this study is 70% for training, 20% for testing and 10% for validation purposes.

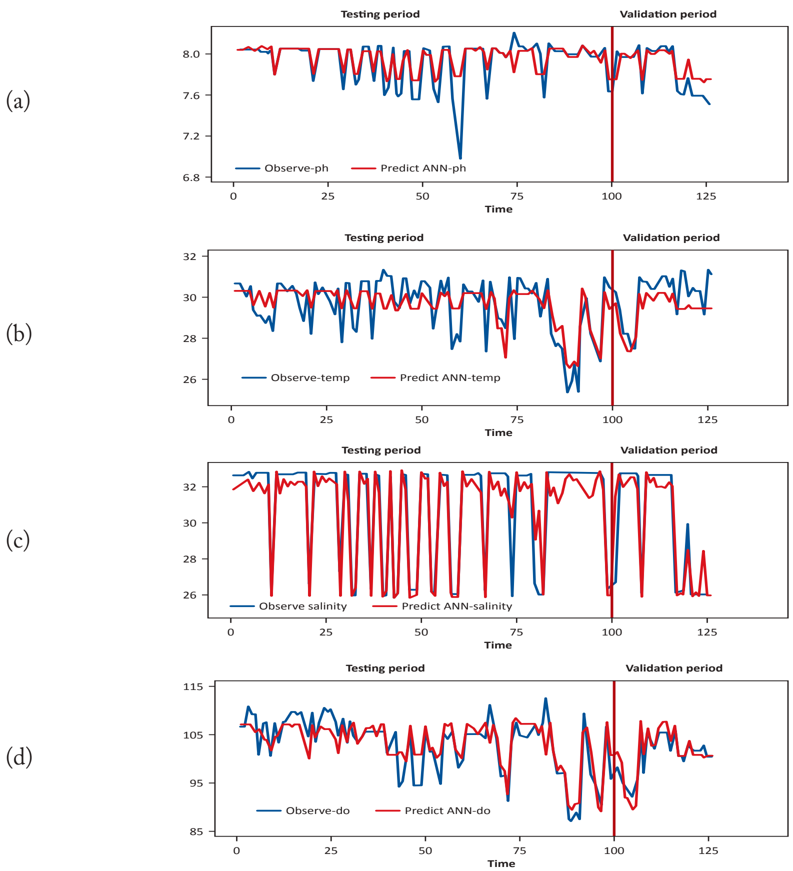


Figure 4: Observed versus ANN model predicted data for each water quality parameters; (a) temperature, (b) pH, (c) salinity, (d) DO

These results indicate that the neural network that was developed are able to detect the pattern in water quality parameters to provide prediction of the daily variations data due to the predicted graph is almost similar to the observed graph.

The Hybrid Modelling Approach

The testing and validation period for all parameters based on hybrid models are shown in Figure 4 (a),(b), (c) and (d). The figures shows that predict data are closed with the observed data for all water quality parameters. The predicted data was able to identify the pattern of the input data to provide desired and valid predictions better than the ARIMA and neural network models. Hence, hybrid models are proven to give the most reliable prediction when compared to single models.

(a)

(b)

(c)

(d)

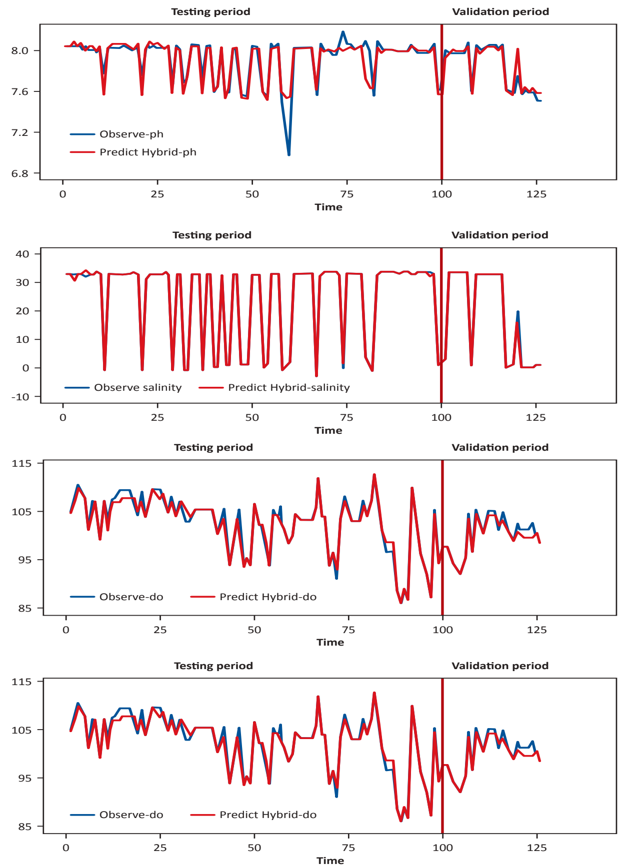


Figure 5: Observed versus hybrid model predicted data for each water quality parameters; (a) temperature, (b) pH, (c) salinity, (d) DO

Comparison of the Models Performance

To evaluate the performance of the models developed, this study used two statistical performance evaluation criteria, mean absolute error (MAE) and root mean square error (RMSE) as in Eqs (9) and (10). The comparative performance of ARIMA, ANN and NNARIMA for all parameters are tabulated in Table 2.

Table 2: Comparative of models performances using MAE and RMSE

Parameter Models	Temperature °C		pH		Salinity (ppt)		DO (ppm)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ARIMA	0.9100	1.1585	0.1303	0.2134	9.5102	18.9201	3.1811	4.2905
ANN	0.7031	3.6207	0.07206	0.1323	3.4192	6.0801	2.2211	2.8948
NNARIMA	0.3275	0.2935	0.0431	0.0889	3.0437	4.8789	1.7780	0.5204

Table 3: MAE and RMSE Reduced Error Percentage for all parameters (%)

Parameters	MAE Reduced Error (%)		RMSE Reduced Error (%)	
	ARIMA-ANN	ARIMA-HYBRID	ARIMA-ANN	ARIMA-HYBRID
Temperature (°C)	22.74	64.01	10.52	74.67
pH	44.68	66.92	38.02	58.34
Salinity (ppt)	64.05	68.00	61.81	74.21
DO (ppm)	30.18	44.11	32.53	87.87

From the Table 2, we can see that the hybrid model of NNARIMA has the lowest MAE and RMSE for all the parameters which indicates that it has the highest accuracy compared to the single models of ANN and ARIMA. The RMSE's of water temperature, DO, pH and salinity are 1.185°C, 0.2134pH, 18.9102ppt and 4.2905ppm for ARIMA modelling approach. Table 3, we can see that the MAE reduced error percentage decrease by 22.73%, 44.67%, 64.06% and 30.18% for water temperature, pH, salinity and DO when ANN models were used. Applying the hybrid models, the MAE reduced error percentage the values decrease by 53.42%, 40.18%, 10.89% and 19.94% in the MAE values when hybrid model is used for water temperature, pH, salinity and DO respectively. Comparatively, the RMSE reduced error percentage decrease by 10.52%, 38.03%, 61.81%, and 32.55% for water temperature, pH, salinity and DO when ANN models were used. As the hybrid models were used the RMSE reduced error percentage decreases by 71.68%, 32.73%, 19.76%, 82.02% for the parameters.

5. CONCLUSION

This study used ARIMA, neural network and hybrid NNARIMA models to predict the water quality time series. The hybrid model developed would be able to utilize the benefits of both the traditional methods and ANN. The result obtained shows that ANN model is more reliable and suitable to be hybrid with ARIMA model in predicting water quality time series. The hybrid model developed in this study can be much useful in water quality management efforts to ensure that water resource are sustainable for the coming years. In this study, two accuracy measures, RMSE and MAE were formulated in order to demonstrate the performance of the developed models in predicting water quality time series. The hybrid models performance were compared relatively with single models respectively ANN and ARIMA. The least values of MAE and RMSE gives an improved performance in the predicting of water quality time series.

REFERENCES

Khashei, M. & Bijari, M. (2011). An artificial neural network (p,d,q) model for time series forecasting. Expert Systems with Applications. 37, 479-489

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14, 35–62

Linear Mixed Models for Analyzing Total Weights of Fish in Na Thap River, Southern Thailand

Yenni Angraini*
Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia - y_angraini@apps.ipb.ac.id
Khairil Anwar Notodiputro
Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia - khairilnotodiputro@gmail.com
Kusman Sadik
Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia - kusmansadik@gmail.com
Sarawuth Chesoh
Department of Science, Prince of Songkla University, Thailand - chesoh99@gmail.com

Abstract

In this article we have developed mixed models to understand the relationship between the total weight of fish which was assumed to be normally distributed, and some explanatory variables based on data collected along the Na Thap river in Southern Thailand. The response and explanatory variables were observed on a regular basis and these formed longitudinal data. Moreover, since the explanatory variables consisted of fixed and random effects then we used the Linear Mixed Model (LMM) approach. The explanatory variables in the model consisted of salinity, BOD, DO, pH, and site. The site effect was considered as a random effect, whereas the effect of other explanatory variables were fixed.

The results showed that the best model consisted of explanatory variables including zone, time, salinity, BOD, DO and pH which have significantly affected the average amount of fish weight in the Na Thap River. In this case zone, salinity, BOD and PH produced negative effects on the average amount of fish weights. Moreover, in this model we found that the variation of the average amount of fish weights among site was not significant. However, the parameter σ^2 was significant indicating that the variation of fish weights was existed within sites.

Keywords: Longitudinal data, Linear mixed model (LMM), Na thap river, Total weight of fish.

1. INTRODUCTION

Na Thap River in south Thailand is the house of aquatic fauna and flora and produces fishery products that are exploited by the communities around the river (Chesoh 2010 and 2011). Since the Electricity Generating Authority of Thailand (EGAT) dumps its waste to Na Thap River, the quality of the water, the growth of the organism, the total weight and existence of fish species and other organism are affected. According to Chesoh et al (2004) the local community fears the power plant cooling system will damage the ecosystem and water life of Na Thap River which can directly affect their livelihood. Other study by Chesoh and Lim (2015) said that the power plant waste do not affect significantly to the quality of river water but they recommend to conduct further study to monitor the river ecology. According to Russev 1972 in Saheem et al, 2014 the diversity of total weight of fish in an ecosystem are affected by the structure of fish population, interaction between species and environmental conditions (ecology factor). Ecology factor includes water depth, temperature, distance to the shore, nutrition contents and water quality.

This study will develop a model to recognize whether or not the diversity of total weight of fish in each zone of Na Thap river is affected by EGAT construction. Response variable is total weight of fish that is assumed normal distribution. Factors affected response variable not only fixed effects but also random effects. For example location or site of the survey and it is classified as independent variabel but random effects. Data is observed from time to time for every site that it has longitudinal data structure. Between sites assumed to be independent but within site correlation exists between observation time. For this reason, the appropriate model to develop is Linier Mixed Model (LMM) for longitudinal data.

2. MIXED MODEL FOR LONGITUDINAL DATA

In general, the model for every i subject consists of fixed effect (β) , random effect u and can be written in the following matrix equation (Welch 2009):

$$Y_i = X_i\beta + Z_i u_i + \varepsilon_i$$

with $u_i \sim N(0,G)$ and $\varepsilon_i \sim N(0, R_i)$. Matriks G is variance-covariance for the q random effect u_i for the i-th subject. If there were only one random effect per subject (for example a random intercept) then G 1 x 1 matrix. Furthermore, if there were two random effect per subject (for example intercept and a random slope), then G would be 2 x 2. Two common structures for G is variance component matrix (VC) and unstructured matrix (UN) such as

$$G = Var(u_i) = \begin{bmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{bmatrix}, G = Var(u_i) = \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u1,u2} \\ \sigma_{u1,u2} & \sigma_{u2}^2 \end{bmatrix}$$

R is variance-covariance matrix in which the dimension depends on number of observation for subject i. First-order Autoregressive AR(1), Toeplitz (Toep), Toeplitz(2) (Toep(2)), Heterogeneous Compound Symetry (CSH), Heterogeneous first-order Autoregressive (ARH(1)) and Heterogeneous Toeplitz (Toeph) can be accommodated in if there are autocorrelation among observations.

A common linear model or marginal model does not use random effects and could be written as

$$Y_i = X_i\beta + \varepsilon_i^*, \ \varepsilon_i^* \sim N(0, V_i)$$

V_i is variance-covariance matrix of Y_i. Implied marginal distribution of Y_i based on linear mixed model is $Y_i \sim N(X_i\beta, Z_i GZ_i' + R_i)$, $E(Y_i) = X_i\beta$ dan $Var(Y_i) = V_i = Z_i GZ_i' + R_i$. V_i is required to be definite positif but G and R_i do not have to be positive-definite.

Variance-covariance G can be estimated using maximum likelihood (ML) and restricted ML (REML). $\hat{u}_{ML}$ is estimated by maximizing $l_p(u) = -\frac{1}{2}\{\ln|V| + (y - X\tilde{\beta})' V^{-1}(y - X\tilde{\beta})\}$ where $\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$. $\hat{u}_{REML}$ is estimated by maximizing $l_R(u) = l_p(u) - \frac{1}{2}\ln|X'V^{-1}X|$. Fixed and random effects are estimated by $\hat{\beta} = (X\hat{V}^{-1}X)^{-1}X\hat{V}^{-1}Y$, $\hat{u} = \hat{G}Z'\hat{V}^{-1}(Y - X\hat{\beta})$, where $\hat{V} = V(\hat{u}_{ML})$ or $\hat{V} = V(\hat{u}_{REML})$.

The best model can be selected using AIC (Akaike Information Criteria) and BIC (Bayes Information Criteria).

3. DATA AND METHODOLOGY

The data used is survey data from Electricity Generating Authority of Thailand (EGAT) between January and December 2013. In the survey that conducted by EGAT, Na Thap river was divided into 3 zones, freshwater, brackish and saline. EGAT is located near a site along the river.

The total weight of fish was measured in gram/1000m^2 and assumed to follow a normal distribution. Total weight of fish is a response variable measured repeatedly per site every month. Site is as random effect. Fixed effects in this study is water zone (freshwater, brackish dan saline) and time. Other fixed effects is ecology associated variable, i.e. salinity level in ppt, dissolved oxygen in mg/liter, Biochemical Oxygen Demand in mg/liter and pH.

Three models have been developed in this study,

Model 1 $y_{jklmno} = \gamma_{00} + u_{0j} + \beta_1 W_k + \beta_2 L_l + \beta_3 B_m + \beta_4 D_n + \beta_5 P_o + \varepsilon_{jklmno}$

where $\varepsilon_{jklmno} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau_{00}^2)$,

Model 2 $y_{ijklmno} = \gamma_{00} + u_{0ij} + \gamma_{01} Z_i + \beta_1 W_k + \beta_2 L_l + \beta_3 B_m + \beta_4 D_n + \beta_5 P_o + \varepsilon_{ijklmno}$

where $\varepsilon_{ijklmno} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau_{00}^2)$,

Model 3 $y_{ijklmno} = \gamma_{00} + u_{0ij} + \gamma_{01} Z_i + \beta_1 W_k + \beta_2 L_l + \beta_3 B_m + \beta_4 D_n + \beta_5 P_o + \varepsilon_{ijklmno}$

where $\varepsilon_{ijklmno} \sim N(0, R)$, R is variance-covariance matrix of random effects $\varepsilon_{ijklmno}$ with structures AR(1) and $u_{0j} \sim N(0, \tau_{00}^2)$.

4. RESULTS

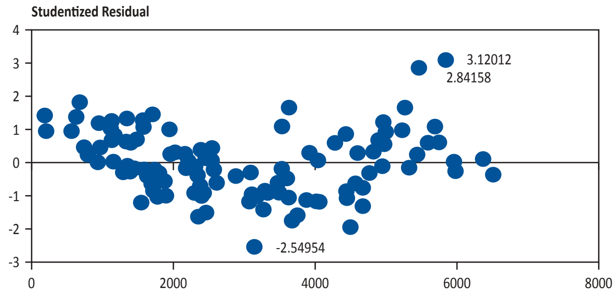


Figure 1. Residual plot of Model 2

The results showed that model 2 has the lowest AIC and BIC values, 1705.1 and 1705.8. Plots between residual and predicted values in Figure 1 show that the residual distribution around middle value zero but several observations have residual greater than $|\pm 2|$, and a non-random pattern. To resolve it, Winsor technique was applied (Huber and Ronchetti, 2009).

Table 1. Fixed effects of Model 2

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	10546.00	1702.97	7	6.19	0.0004
fresh water	-2997.27	380.92	7	-7.87	0.0001
brackish	-1800.36	391.12	7	-4.6	0.0025
Jan	1006.42	276.39	95	3.64	0.0004
Feb	1848.51	352.52	95	5.24	<.0001
Mar	2005.99	406.51	95	4.93	<.0001
Apr	2291.78	471.33	95	4.86	<.0001
May	1643.22	423.48	95	3.88	0.0002
Jun	1536.25	417.89	95	3.68	0.0004
Jul	1059.06	376.8	95	2.81	0.0060
Aug	1471.77	432.06	95	3.41	0.0010
Sept	2884.49	496.27	95	5.81	<.0001
Oct	752.2	351.37	95	2.14	0.0349
Nov	583.04	297.3	95	1.96	0.0528
SAL	-231.74	65.63	95	-3.53	0.0006
DO	23.8008	8.9571	95	2.66	0.0092
BOD	-138.32	67.2905	95	-2.06	0.0426

AIC and BIC of Model 2 with response resulted from Winsor approach is smaller than previous model, which are 1641.3 and 1641.9. Table 1 shows the estimates of the model as well as the significance of parameters. Variance parameter estimate for random effect yielded σ^2_{u1} = 192425 (p value =0.1022) and σ^2 = 273607 (p value < .0001). Variance parameter estimate of random effect in this model is not significant so it suggest that sites do not differ in their average of total fish weight. However, variance parameter estimate of residual is significant so it can be said that more variation among times within sites. The residuals also seem to meet the usual assumption.

The average amount of fish weight is 10546 gram/1000m^2. The fixed effect of zone to the average amount of fish weight is significant at 10% level. The average amount of fish weight decreased by 2997.27 gram/1000m^2 in freshwater zone and 1800.36 gram/1000m^2 in brackish zone compared to the average amont of fish weight in saline zone. The fixed effect of time and other fixed effects to the average amount of fish weight per site is also significant at 10% level. The average amount of fish weight increased in January by 1006.42 gram/1000m^2 when compared to the average amount of fish weight in December. The same interpretation can be done for other months.

5. CONCLUSIONS

The explanatory variables zone, time, salinity, BOD, DO and pH have significantly affected the average amount of fish weight. In this case zone, salinity, BOD and PH produced negative effects on the average amount of fish weights. The other fixed effects, time and DO, yielded positive effects on the average amount of fish weights.

REFERENCES

Chesoh, S., Luangtuwapraneet, C., Ruangchuay, R. et a. (2004). Environmental impacts assessment of aquatic ecosystem and fisheries resources utilization of Nathab canal before Songkhla thermal power plant construction.In proceeding of annual fisheries seminar at Department of Fisheries of Thailand during 25-27 July 2006. 77–94.

Chesoh, S. (2010). Community perception, satisfaction and participation toward power plant development in southernmost of Thailand. Journal of Sustainable Development, 3(2), 84-88. http://dx.doi.org/10.5539/jsd.v3n2p84

Chesoh, S. (2011). Environmental impact assessment of power development project: Lessons from Thailand experiences. Asian Social Science, 7(9), 119-123. http://dx.doi.org/10.5539/ass.v7n9p119

Chesoh, S dan A. Lim. (2015). Monitorng of Thermal Power Palnt Operation on Aquatic Ecosystem in Tropical Estuarine River of Thailand, 2008-2013. E-proceedings of the 36th IAHR World Congress 28 June – 3 July, 2015, The Hague, the Netherlands

Huber, P.J. and Ronchetti, E.M. (2009) Robust Statistics. Wiley Series in Probability and Statistics. 2nd Edition, John Wiley & Sons, Inc., Hoboken.

Saheem, N., A. Lim dan S. Chesoh. (2014). Predicting standing crop using lagged fingerling density of freshwater fish in the Na Thap River of Southern Thailand. Songklanakarin J. Sci.Technol. 36: 13–19

Welch, K. (2009). Random Effects vs. Marginal Models: Different Approaches to Analyzing Repeated Measures / Longitudinal Data. Presented at Midwest SAS Users Group. CSCAR, The University of Michigan.

CPS02:
Health & Social Statistics (1)

Modeling Seasonality in Epidemic Surveillance Data Using Count Time Series Models

Manik S. Awale*
Savitribai Phule Pune University, Pune, India - manik.stats@gmail.com

Abstract

Most of the epidemic outbreaks are seasonal in nature and hence reoccur after a specified time lag. Epidemic outbreaks brings burden on the economy of a country, as the government has to spend a lot of money and resources on the epidemic outbreak control program. To avoid this, one should have proper models to identify the progression of the disease, in advance. It seems that the susceptibility, infectivity and immunity of individuals are also changes with seasons and hence the probability of catching the infection and escaping from infection changes with seasons and with time. The researchers have been trying to model the phenomenon of infection as a function of time. The epidemic surveillance data is always in the integer form and therefore it needs to be tackled by appropriate integer-valued models. Usual Integer Valued Autoregressive (INAR) models cannot capture the seasonality in the data and hence one has to use the seasonal INAR models. In this paper an attempt is made to model the epidemic outbreak using a seasonal INAR time series model. In many of the existing models the thinning parameter is considered as constant while, in the suggested model the thinning parameter is allowed to vary over time deterministically using some trigonometric functions. Probabilistic and inferential properties of the suggested model are studied. The estimates of the parameters are obtained using Conditional Maximum Likelihood (CML) method. The model can be applied to any seasonal count time series data. The predictive performance of the model is studied using simulation. The model is applied to the epidemic data which has seasonality inherent in it. The suggested model is easy to use for practical purpose. It works better in terms of the predictive performance as compared to the existing models.

Keywords: Conditional maximum likelihood; Epidemic; Count time series; Integer-valued autoregressive model.

1. INTRODUCTION

INAR models have been widely used in various disciplines like economics, epidemiology, insurance etc., where the process under consideration take non-negative integer values. INAR(1) model with Poisson marginal distribution was first proposed by Al-Osh and Alzaid (1987). The INAR models with other marginal distributions are considered in McKenzie (1985, 1986). An INAR(1) model with geometric marginal distribution based on negative binomial thinning was proposed by Ristic et al. (2009). The forecasting aspect in INAR(1) models was first studied by Freeland and McCabe (2004) where they proposed the idea of data coherent fore-casts. Jung and Tremyne (2006) extended the idea of coherent forecasting to the INAR(2) model. Coherent forecasting in Binomial counting series of order p was considered by Kim and Park (2010). Maiti et al. (2015) explored the usefulness of the usual AR(p) models for the coherent forecasting under INAR setup. Maiti and Biswas (2015) proposed the coherent forecasting in INAR(1) model with geometric marginal distribution derived by McKenzie (1986).

Monteiro et al. (2010) proposed periodically stationary INAR(1) model with period T, INAR(1)T. This model takes care of the seasonality in the data with period T. But the model su ers some serious drawbacks as the parameters are time dependent, the number of parameters to be estimated is very high and hence less useful for practitioners. Bourguignon et al. (2015) proposed a stationary seasonal INAR(1)s model with constant thinning parameter having a seasonal period s, the seasonality brought in through auto-regression on the number of observations in the (t-s)th time point. The forecasting was carried out using traditional conditional mean. Various properties similar to INAR(1) are studied by the authors.

Seasonality in the epidemic data is evidenced by vast literature available on this topic. A good exposition on seasonality of infectious diseases is available in Grassly and Fraser (2006) and the references therein. In this paper we propose a INAR(1) model with seasonal thinning parameter called SPINAR(1) model. The model takes care of the inherent seasonality in the surveillance data. This model can be used to non-medical data having seasonal pattern. We propose the coherent forecasting of future cases using the said model. Various properties of the model are studied.

Section 2 deals with the model and its various properties. Estimation of parameters using CML and CLS methods is given in Section 3. Real data analysis is considered in Section 4. Section 5 concludes the paper.

2. SEASONAL INAR(1) MODEL (SPINAR(1))

In this section we introduce a seasonal INAR(1) model. Let us consider a process $\{X_t\}$ given by

$$X_t = \phi_t \, o X_{t-1} + \varepsilon_t, \qquad \phi_t \in (0,1) \tag{1}$$

Where,

$$\phi_t = \frac{\exp(\delta_\phi + \beta_\phi \sin(wt) + \gamma_\phi \cos(wt))}{1 + \exp(\delta_\phi + \beta_\phi \sin(wt) + \gamma_\phi \cos(wt))} \tag{2}$$

the operator 'o' is de ned as $\phi_t o\, X = \sum_{i=1}^{X} Y_i$ and $Y_i \sim$ Bernoulli (ϕ_t). Here $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables which is independent of $\{Y_i\}$ and X_t; w = 2π/52 for weekly data and 2π/12 for monthly data. The unconditional distribution of X_0 and ε_t are Poisson $(\frac{\lambda}{1-\phi_0})$ and Poisson (λ) respectively. We call such an INAR(1) model as seasonal INAR(1) or SPINAR(1).

We may write, $$X_{t+k} = \prod_{i=1}^{k} \phi_{t+i} \; o \; X_t + \sum_{j=2}^{k+1} \left(\prod_{l=j}^{k} \phi_{t+l}\right) o \; \varepsilon_{t+j-1}. \tag{3}$$

The k-step ahead conditional probability generating function (pgf) can be derived as,

$$G_{X_{t+k}|X_t} = \left(1 - \prod_{i=1}^{k} \phi_{t+i} + s \prod_{i=1}^{k} \phi_{t+i}\right)^{X_t} e^{-\lambda\left(1+\sum_{j=2}^{k} \prod_{l=j}^{k} \phi_{t+l}\right)(1-s)} \tag{4}$$

Using this, the k-step ahead conditional mean and variance can be obtained as

$$E(X_{t+k} \,|X_t) = \prod_{i=1}^{k} \phi_{t+i} \; X_t + \lambda\left(1 + \sum_{j=2}^{k} \prod_{l=j}^{k} \phi_{t+l}\right) \tag{5}$$

and

$$V(X_{t+k} \,|X_t) = \prod_{i=1}^{k} \phi_{t+i} \; \left(1 - \prod_{i=1}^{k} \phi_{t+i}\right) X_t + \lambda\left(1 + \sum_{j=2}^{k} \prod_{l=j}^{k} \phi_{t+l}\right). \tag{6}$$

From (5), (6) and (7), it can be observed that, as $k \longrightarrow \infty$, the conditional pgf approaches to the pgf of Poisson (λ) and conditional mean and variance converges to unconditional mean and variance of Poisson (λ).

Theorem 1.
The k-step ahead conditional probability mass function of the SPINAR(1) process is given as,

$$P_k(y|x) =$$
$$\sum_{r=0}^{\min(y,x)} \binom{x}{r} \left(\prod_{i=1}^{k} \phi_{t+i}\right)^r \left(1 - \prod_{i=1}^{k} \phi_{t+i}\right)^{x-r} e^{-\lambda\left(1+\sum_{j=2}^{k} \prod_{l=j}^{k} \phi_{t+l}\right)} \frac{\left(\lambda\left(1+\sum_{j=2}^{k} \prod_{l=j}^{k} \phi_{t+l}\right)\right)^{y-r}}{(y-r)!}$$

Proof. See Appendix
Remark. The auto covariance function $\upsilon(k)$ can be obtained as

$$\upsilon(X_{t+k}, X_t) = \prod_{l=1}^{k} \phi_{t+l} \left(\frac{\lambda}{1+\phi_t}\left(1 + \frac{\lambda}{1+\phi_t}\right)\right) - \frac{\lambda^2}{(1-\phi_t)^2} \prod_{l=1}^{k} \phi_{t+l} \, .$$

Hence, it can be proved that the autocorrelation decays as lag increases i.e.,

$$\rho(X_{t+k}, X_t) \rightarrow 0, \ \ as \ \ k \rightarrow \infty. \tag{7}$$

The sample path, sample ACF and sample PACF of the process are shown in Figure 1.

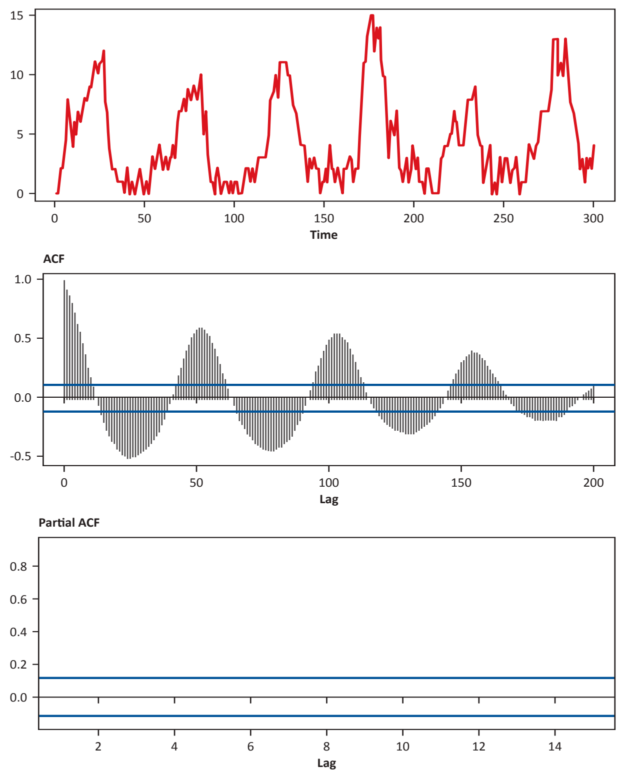


Figure 1: Sample Path of SPINAR(1) Process with its Sample ACF and Sample PACF (λ=1, δ_φ=0.8, β_φ=1.5 and γ_φ=-1).

3. ESTIMATION

Conditional maximum likelihood estimators can be obtained by maximizing the log-likelihood function

$$\log L\left(x_1, \dots, x_n; \lambda, \delta_\phi, \beta_\phi, \gamma_\phi\right) = \sum_{t=2}^{n} \log P(X_t | X_{t-1}).$$

Numerical estimates can be obtained using the constrained optimization in R software. Similarly, the conditional least squares estimates of the parameters can be obtained by minimizing the function.

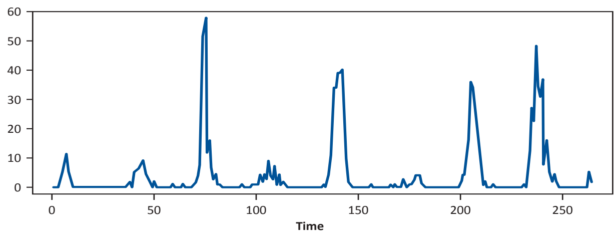
$$S_n\left(\lambda, \delta_\phi, \beta_\phi, \gamma_\phi\right) = \sum_{t=2}^{n} (X_t - E(X_t | X_{t-1}))^2$$

One thousand simulations were carried out with different parameter combinations and sample sizes. We have considered the period as s = 52 for the simulation. The estimates and their mean square errors are reported in Table 1. It may be noted from Table 1 that all the parameters are estimated quite accurately using both the estimation procedures.

Sample Size	$\hat{\lambda}_{cls}$	$\hat{\delta}_{\phi cls}$	$\hat{\beta}_{\phi cls}$	$\hat{\gamma}_{\phi cls}$	$\hat{\lambda}_{cml}$	$\hat{\delta}_{\phi cml}$	$\hat{\beta}_{\phi cml}$	$\hat{\gamma}_{\phi cml}$
			$\lambda = 0.5,\ \delta_\phi = 1,\ \beta_\phi = 1,\ \gamma_\phi = 2$					
200	0.516	-1.599	1.307	2.433	0.502	-1.425	1.243	2.394
	0.005	2.553	1.133	2.354	0.004	1.845	1.032	1.547
500	0.507	-1.151	1.053	2.111	0.500	-1.105	1.064	2.092
	0.001	0.237	0.134	0.248	0.001	0.163	0.107	0.197
1000	0.504	-1.088	1.028	2.061	0.500	-1.053	1.030	2.051
	0.001	0.094	0.055	0.102	0.001	0.066	0.049	0.089
			$\lambda = 1,\ \delta_\phi = 1,\ \beta_\phi = 2,\ \gamma_\phi = 3$					
200	1.026	0.890	2.138	3.069	0.999	0.977	2.066	3.079
	0.013	0.122	0.162	0.268	0.008	0.060	0.104	0.129
500	1.007	0.973	2.045	3.033	1.000	0.988	2.024	3.017
	0.005	0.044	0.050	0.092	0.003	0.022	0.038	0.041
1000	1.006	0.978	2.027	3.022	1.001	0.992	2.015	3.015
	0.002	0.019	0.022	0.041	0.002	0.011	0.018	0.020
			$\lambda = 2,\ \delta_\phi = 2,\ \beta_\phi = 3,\ \gamma_\phi = 5$					
200	2.014	1.975	3.087	5.064	1.993	2.009	3.035	5.049
	0.017	0.104	0.133	0.293	0.016	0.073	0.103	0.199
500	2.011	1.982	3.031	5.011	1.999	2.000	3.017	5.016
	0.007	0.040	0.048	0.112	0.005	0.027	0.039	0.068
1000	2.004	1.989	3.026	5.0139	2.002	2.007	3.012	5.018
	0.003	0.019	0.025	0.056	0.003	0.013	0.020	0.038

4. APPLICATION TO REAL DATA SET

We consider the weekly number of influenza cases from city of Stuttgart, Germany for the period 2001-2008, observed for 264 weeks, available from the Robert Koch Institute (website https://survstat.rki.de). From the time series and correlation plots in Figure 2, it can be seen that the data has got seasonal behaviour and AR(1) structure and therefore data can be modelled using (1). In Table 2 we have assessed three models viz. INAR(1) with Poisson marginal called as INAR(1), INAR(1) with Poisson marginal and seasonal period s called as INAR(1)s and Seasonal INAR(1) model with Poisson marginal called as SPINAR(1). It is found that SPINAR(1) model fits well to the given data. We have compared the predictive performance of the INAR(1) model with SPINAR(1) model. As the INAR(1)s is not giving good fit, we have excluded it from this comparison. We have computed Predictive root mean square error (PRMSE), Predictive median absolute error (PMAE) and 100(1-γ)% Prediction Intervals (PI). The results are tabulated in Table 3. From the forecast error measures, it can be observed that the suggested model outperforms the usual INAR(1) model.



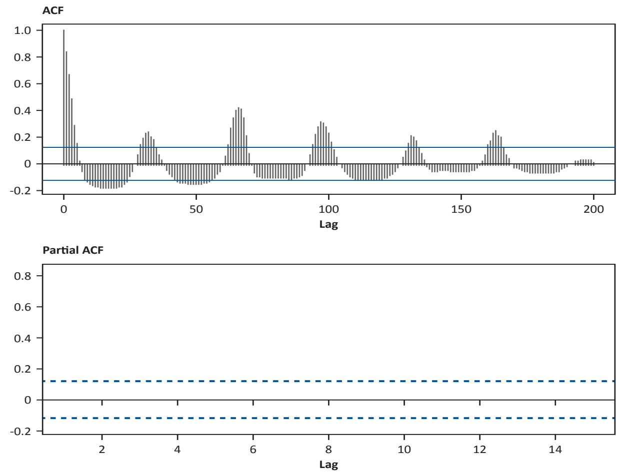


Figure 2: Time series plot, sample ACF and sample PACF of In uenza data

Table 2: Model selection using AIC and BIC

Model	Estimates	AIC	BIC
PoINAR(1)	ϕ=0.6792, λ= 1:2944	1854.602	1861.702
PoINAR(1)$_s$	ϕ= 0:0000, λ = 4:5611	3267.966	3275.118
SPINAR(1)	λ=1.3 λ=1.3004, δ =0.5597, β = -0.5600, γ =-0.1177	1838.991	1853.295

Table 3: Forecasting Values and 100(1-γ)% Prediction Interval for Influenza Data

		Mean Prediction		Median Prediction		Prediction Interval (γ =0.1)	
k	X_{295+k}	INAR(1)	SPINAR(1)	INAR(1)	SPINAR(1)	INAR(1)	SPINAR(1)
1	0	1.294	1.300	1	1	[0, 4)	[0, 4)
2	0	2.174	2.071	2	2	[0, 5)	[0, 5)
3	0	2.771	2.496	3	2	[0, 6)	[0, 6)
4	5	3.176	2.704	3	3	[0, 7)	[0, 6)
5	2	3.452	2.784	3	3	[0, 8)	[0, 6)
PRMSE		0.883	0.851				
PMAE				1.8	1.6		

5. CONCLUSIONS

The usual INAR(1) model is not suitable for a seasonal integer-valued time series data as revealed from the data analysis. Using information criterion such as AIC and BIC, we found that the INAR(1)s model is not a good alternative for the given data set. Hence, there is a need to have a model which can be fitted reasonably well to this type of data. The new model (SPINAR(1)) proposed in this paper would be a better choice for modelling the seasonal data. The model SPINAR(1) is easy to use in terms of estimation of parameters and forecasting the future values. Various scenarios like epidemic and endemic components in the model are subjects of future research.

REFERENCES

Al-Osh, M. A. & Alzaid, A. A. (1987). First order integer-valued autoregressive (INAR(1)) processes. Jour-nal of Time Series Analysis, 8, 261-275.

Bourguignon, M., Vasconcellos, K. L. P., Reisen, V. A., & Ispany, M. (2015). A Poisson INAR(1) process with seasonal structure. Journal of Statistical Computing and Simulation. 86(2), 373-387.

Freeland, R. K. & McCabe, B. P. M. (2004). Forecasting discrete valued low count time series. International Journal of Forecasting. 20, 427-434.

Grassly, N. C. & Fraser, C. (2006). Seasonal infectious disease epidemiology. Proceedings of Royal Society B. 273, 2541-2550.

Jung, R. C. & Tremyne, A. R. (2006). Coherent forecasting in integer time series models. International Journal of Forecasting. 22, 223-238.

Kim, H-Y. & Park, Y. (2010). Coherent Forecasting in Binomial AR(p) Model. Communications of the Korean Statistical Society. 17(1), 27-37.

Maiti, R. & Biswas, A. (2015). Coherent forecasting for over-dispersed time series of count data. Brazilian Journal of Probability and Statistics. 29(4), 747-766.

Maiti, R., Biswas, A. & Das, S. (2015). Coherent forecasting for count time series using Box-Jenkins's AR(p) model. Statistica Neerlandica. 70(2), 123-145.

McKenzie, E. (1985). Some simple models for discrete variate time series. Water Resource Bulletin. 21, 645-560.

McKenzie, E. (1986). Autoregressive moving-average processes with negative binomial and geometric distri-butions. Advances in Applied Probability. 18, 679-705.

Monteiro, M., Scotto, M. G. & Pereira, I. (2010). Integer-valued autoregressive processes with periodic struc-ture. Journal of Statistical Planning and Inference. 140, 1529-1541.

Ristic, M. M., Bakouch, H. S. & Nastic, A. S. (2009). A new geometric rst order integer-valued autoregres-sive (NGINAR(1)) process. Journal of Statistical Planning and Inference. 139, 2218-2226.

APPENDIX:

Proof of Theorem 1:

Using the property of the binomial thinning operator we can write,

$$\prod_{i=1}^{k}\phi_{t+i}\ o\ X_t\,|X_t \sim Binomial\left(X_t,\ \prod_{i=1}^{k}\phi_{t+i}\right)$$

and

$$\sum_{j=2}^{k+1}\left(\prod_{l=j}^{k}\phi_{t+l}\right) o\ \varepsilon_{t+j-1} \sim Poisson\ \left(\lambda\left(1+\sum_{j=2}^{k}\prod_{l=j}^{k}\phi_{t+l}\right)\right).$$

We know that the distribution of the convolution of the two random variables is

$$P_k\,(y|x) = P(X_{t+k} = y|X_t = x) = \sum_{r=0}^{\min(y,x)} P\left(\prod_{i=1}^{k}\phi_{t+i}\ o\ x = r\right) P\left(\sum_{j=2}^{k+1}\left(\prod_{l=j}^{k}\phi_{t+l}\right) o\ \varepsilon_{t+j-1} = y-r\right).$$

Using this argument the proof follows.

Transformation Cure Models for Enrichment Design in Targeted Clinical Trials

Chih-Yuan Hsu*
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C cyhsu@stat.sinica.edu.tw
Chen-Hsin Chen
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C. Graduate Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, 10055, Taiwan, R.O.C. chchen@stat.sinica.edu.tw

Abstract

Targeted therapies have been successfully utilized in treatments of some cancer patients with specific molecular targets. The enrichment design provides an innovative research approach to targeted clinical trials. An enrichment design selects patients by detection of the molecular targets using a diagnostic device before allocation of the treatment regimens. However, no diagnosis is a perfect test without any error, i.e., achieving the positive predictive value of 100%. This may lead to underestimation of the treatment effect for the patients truly having the molecular targets when a new targeted therapy is developed in a clinical trial. In consideration of the positive predictive value, Chen, Lin and Liu (2013) recently proposed a statistical method, only under the assumption of exponential survival distributions, for inference on censored data under enrichment design. Moreover, with a great advance in biomedical technologies, a proportion of patients may be cured after receiving the targeted therapy. This motivates us to consider developing the transformation cure model under enrichment design. To cope with practical concerns, we, in this study, extend the exponential model to the more flexible transformation models, which include commonly-used proportional hazards models and proportional odds models. Based on the likelihood approach and simulation studies, our proposed method performs well in estimation of the model parameters under the enrichment design.

Keywords: Diagnostic error; Positive predictive value; Proportional hazards models; Proportional odds models.

1. INTRODUCTION

In recent years, the issue of personalized medicine becomes more popular due to the pursuit of individual optimal therapies. For example, in cancer studies, target therapies are commonly used to treat the patients with some specific molecular targets because they will exert the greatest efficacy for these patients. To attain the ideal medical treatment, an accurate diagnosis testing for genetic information will become particularly important. However, no diagnosis is exactly accurate and perfect without any error, so the recruited patients may be contaminated. It consists of patients with and without the molecular targets. This maybe leads us to underestimate the actual treatment effect in the final data analysis when these target therapies are used. Since the diagnosis error is unavoidable, some researchers proposed suitable statistical methods to overcome the problem of underestimation. Chen et al. (2013) used the EM algorithm to deal with the unobservable molecular status and inferred the true hazard ratios of one-parameter exponential survival distributions under the enrichment design. Later, Tsai et al. (2016) extended the one-parameter exponential survival distributions to the Cox proportional hazards model.

With the great advance in medical technologies, the situation of no recurrence in some patients is frequently observed at the end of a long-term study. It makes us believe that some patients may be cured eventually after receiving a treatment, and one reason for patients' cure may be associated with their biomarkers. Berkson and Gage (1952) first used a mixture of exponential distributions and a permanent cure proportion to analyze the survival data. Under the structure of the mixture models, various parametric and nonparametric methods, including the approaches of estimating the parameters, have been proposed. Farewell (1986) used a logistic regression to fit the cure proportion, and assumed the Weibull survival distributions for the non-cure patients. Peng et al. (1998) extended the Weibull distributions to the generalized F distributions. On the other hand, in these nonparametric methods, Kuk and Chen (1992), Peng and Dear (2000) and Sy and Taylor (2000) considered the Cox proportional hazards cure model. The first one proposed a Monte Carlo approximation of a marginal likelihood together with the EM algorithm on the estimation of parameters, while the last two used the approach of maximizing the joint parametric- nonparametric likelihood function. Lu and Ying (2004) considered a more flexible transformation cure model, which contains the proportional hazards cure model and the proportional odds cure model, and took advantage of the martingale estimating equations for the estimation of parameters.

In this paper, we propose an approach to handle the survival data with cure fractions together with the diagnosis error on patients' molecular targets by assuming a transformation cure model and using the EM algorithm.

2. THE PROPOSED MODEL

2.1. Molecular Status

Let ξ be the indicator of molecular target in the recruited patients with the positive diagnosis. It is unobservable and is assumed to follow a Bernoulli distribution with the mean p,

$$f(\xi) = p^{\xi}(1-p)1-^{\xi},\ \xi = 0, 1 \tag{2.1}$$

Here, p is the positive predictive value (PPV).

2.2. Transformation Cure Model

Denote D as the indicator that a patient will not be cured eventually, and T as his/her survival time. We consider the transformation cure model,

$$S_{mix}(t|x_\xi, z_\xi) = \pi(x_\xi)S_u(t|z_\xi) + 1 - \pi(x_\xi), \tag{2.2}$$

with the non-cure proportion

$$\pi(x_\xi) = P[D = 1|x_\xi] = \frac{exp\{\Upsilon^T x_\xi\}}{1 + exp\{\Upsilon^T x_\xi\}}\ , \tag{2.3}$$

and the survival function of the non-cure patients

$$S_u(t|z_\xi) = exp\left\{-G\left(\int_0^t Y^*(s)e^{\beta^T z_\xi(s)}dQ(s)\right)\right\}, \tag{2.4}$$

where $x_\xi = (1, x^T, \xi, \xi x^T)^T$ and $z_\xi(s) = (z^T(s), \xi, \xi z^T(s))^T$ are vectors of covariates consisting of the known x, z(s) and the unobservable molecular status ξ; $\Upsilon = (\Upsilon_0, \Upsilon_1^T, \Upsilon_2, \Upsilon_3^T)^T$ and $\beta = (\beta_1^T, \beta_2, \beta_3^T)^T$ are the corresponding vectors of parameters. Also, the increasing function Q(s) is a nuisance parameter. Particularly, z(s) can be allowed to be time-dependent. $Y^*(s) = I(T \geq s)$ where $I(\cdot)$ is the indicator function, and G is a specified strictly-increasing and continuously- differentiable function. If G(t) = t, T satisfies a proportional hazards (PH) model; if G(t) = log(1 + t), T satisfies a proportional odds (PO) model.

From (2.4), the corresponding cumulative hazard function/intensity is

$$\Lambda_u(t|z_\xi) = G\left(\int_0^t Y^*(s)e^{\beta^T z_\xi(s)}dQ(s)\right), \tag{2.5}$$

and the hazard function

$$d\Lambda_u(t|z_\xi) = g\left(\int_0^t Y^*(s)e^{\beta^T z_\xi(s)}dQ(s)\right)Y^*(t)e^{\beta^T z_\xi(s)}dQ(t)\ , \tag{2.6}$$

where g(t) = dG(t)/dt. Finally, we denote W as the censoring time and assume that it is independent of T conditional on $z_\xi(.)$. Also, we define Y = min(T, W) and δ = I(T < W). If δ = 1, then D = 1 and δD = 1. If δ = 0, then δD = 0. This shows δ = δD.

3. PARAMETER ESTIMATION

Suppose n patients with the positive diagnosis are enrolled into the study, and their data {Oi = (xi, zi(t), Ni(t), Yi(t)) : i = 1, . . . , n, 0 ≤ t ≤ τ } are collected, where Ni(t) = I[δi = 1, Yi ≤ t] denotes the counting process, Yi(t) = I[Yi ≥ t] denotes the at-risk process, and τ is the maximum follow-up time. We define $x_{i,\xi_i} = (x_i^T, \xi_i, \xi_i x_i^T)^T$ and $z_{i,\xi_i}(.) = \left(z_i^T(.), \xi_i, \xi_i z_i^T(.)\right)^T$ i = 1, . . . , n, and assume that Oi's are mutually independent for all i. It follows from (2.1), (2.3) and (2.4) that the complete-data log-likelihood function is

$$
\begin{aligned}
l(\theta) &= l(\theta; O, \xi, D) \\
&\alpha \log \prod_{i=1}^{n} p^{\xi_i}(1-p)^{1-\xi_i} + log \prod_{i=1}^{n} \pi\left(x_{i,\xi_i}\right)^{D_i}\left\{1-\pi\left(x_{i,\xi_i}\right)\right\}^{1-D_i} \\
&+log \prod_{i=1}^{n}\left\{d\Lambda_u\left(y_i|z_{i,\xi_i}\right)\right\}^{\delta_i}\left\{S_u\left(y_i|z_{i,\xi_i}\right)\right\}^{D_i} \\
&\triangleq logL_m(p; O, \xi) + logL_c(\Upsilon; dQ, O, \xi, D) + logL_s(\beta; dQ, O, \xi, D),
\end{aligned}
\tag{3.1}
$$

where θ = (p, γ, β, dQ).

For simplicity of notations, we define

$$
\eta\left(t, z_{i,\xi_i}, \delta_i; \beta, Q\right) = \int_0^t Y_i(u) e^{\beta^T z_{i,\xi_i}(u)} dQ(u),
\tag{3.2}
$$

so $g_i(t;\beta,Q) = g\left(\eta\left(t, z_{i,\xi_i}, \delta_i; \beta, Q\right)\right)$ and $G_i(\beta,Q) = G\left(\eta\left(\tau, z_{i,\xi_i}, \delta_i; \beta, Q\right)\right)$ From (2.4) and $G_i(\beta,Q) = -logS_u\left(y_i|z_{i,\xi_i}\right) = \int_0^{y_i} d\Lambda_u\left(t|z_{i,\xi_i}\right)$.(2.6), According to Chen (2009), gi(t−; β, Q) can ensure better finite sample properties than gi(t; β, Q) even though they are asymptotically equivalent. Hence, we consider another representation for Gi(β, Q) by

$$
G_i(\beta, Q) = \int_0^T Y_i(t) e^{\beta^T z_{i,\xi_i}(t)} g_i(t-;\beta,Q) dQ(t),
\tag{3.3}
$$

and then rewrite

$$
\begin{aligned}
&\log L_s(\beta, dQ; \boldsymbol{O}, \boldsymbol{\xi}, \boldsymbol{D}) \left[= \sum_{i=1}^{n}\left\{\delta_i logd\Lambda_u\left(y_i|z_{i,\xi_i}\right) + D_i logS_u\left(y_i|z_{i,\xi_i}\right)\right\}\right] \\
&= \sum_{i=1}^{n} D_i \left[\left[\int_0^T (\beta^T z_{i,\xi i}(t) + logg_i(t-;\beta,Q)\right.\right. \\
&\left. + logdQ(t))dN_i(t) - \int_0^T Y_i(t) e^{\beta^T z_{i\xi_i}(t)} g_i(t-;\beta,Q) dQ(t)\right].
\end{aligned}
\tag{3.4}
$$

Due to the completely/partial-completely unobservable ξ′s and D′ s, we take advantage of the EM algorithm to estimate θ. E-step: we define

$$
\begin{aligned}
C(\boldsymbol{\theta}|\boldsymbol{\theta}') &= E_{\theta'}[l(\boldsymbol{\theta})] = l(\theta; \boldsymbol{O}, E_{\theta'}[\xi|\boldsymbol{O}], E_{\theta'}[\boldsymbol{D}|\boldsymbol{O}]) \\
&= \sum_{i=1}^{n} \tilde{\xi}_{i,\theta'} \log p + \left(1-\tilde{\xi}_{i,\theta'}\right) log(1|-p) \\
&+ \sum_{i=1}^{n}\sum_{j=0}^{1} \tilde{\xi}_{i,\theta'}^{j}\left(1-\tilde{\xi}_{i,\theta'}\right)^{1-j}\left[\tilde{D}_{ij,\theta'} log\, \pi(x_{ij}) + \left(1-\tilde{D}_{ij,\theta'}\right) log\left(1-\pi(x_{ij})\right)\right] \\
&+ \sum_{i=1}^{n}\sum_{j=0}^{1} \tilde{\xi}_{i,\theta'}^{j}\left(1-\tilde{\xi}_{i,\theta'}\right)^{1-j} \tilde{D}_{ij,\theta'}\left[\left[\int_0^T \left(\beta^T z_{ij}(t) + logg_{ij}(t-;\beta,Q)\right.\right.\right. \\
&\left. + logdQ(t)\right) dN_i(t) - \int_0^T Y_i(t) e^{\beta^T z_{ij}(t)} g_{ij}(t-;\beta,Q) dQ(t)\bigg],
\end{aligned}
$$

where $\theta' = (p', \gamma', \beta', dQ')$, and x_{ij} , z_{ij} $(\cdot)$ and $(\cdot)_{ij}$ $(t; \beta, Q)$ are the shorthand for x_{i,ξ_i} , z_{i,ξ_i} $(\cdot)$ and $(\cdot)_i(t; \beta, Q)$, respectively, with $\xi_i = j$, $j = 0$ or 1. Also, $\tilde{\xi}_{i,\theta'} = E_{\theta'} [\xi_i|O_i]$ and $\tilde{D}_{ij,\theta'} = E_{\theta'} [D_i|\xi_i = j, O_i]$ are the conditional expectation of ξ_i and D_i, respectively, with $\theta' = \theta(k)$, k being a nonnegative integer.

M-step: Given $\theta' = \theta(k)$, we find the maximizer of $C(\theta|\theta(k))$ by solving the score equation of $\partial C(\theta|\theta(k))/\partial\theta = 0$. Then, we update θ' with the new maximizer $\theta(k+1)$ and repeat the iterated procedure until $\{\theta(k), k = 1, \ldots\}$ converges.

4. NUMERICAL STUDIES

To better understand the advantages of the proposed method, we compare ours with one which ignores the diagnostic error (i.e., misleading $p = 1$). We simulate a two-arm enrichment clinical trial: Suppose n patients with a positive diagnosis for a specific molecular target are recruited, and they are randomly assigned to the two treatment groups with 1-1 allocation ratio. We assume that the non-cure proportion of the patients with the molecular target in the new treatment groups is 0.20 and their survival time follows an exponential distribution with the median of 1; for the other patients, their non-cure proportion is 0.35 and their survival times follows an exponential distribution with median of 0.5. In other words, only the patients having the molecular target and receiving the new treatment have a better performance on their survival time. Hence, the mathematical notations of the scenario can be expressed as follows: x and $z(t) = z$ denote the indicator of the treatment groups, and $\pi = \pi_0 = 0.35$ or $\pi = \pi_1 = 0.20$ according to $\xi x = 0$ or $\xi x = 1$; $S_u(t) = \exp\{-2 \log(2)t\}$ or $= \exp\{- \log(2)t\}$ according to $\xi z = 0$ or $\xi z = 1$. The corresponding parameters of interest are $\gamma = (\gamma_0, 0, 0, \gamma_1)^T$ with $\gamma_0 = -0.619$ and $\gamma_1 = -0.767$, and $\beta = \beta_1 = -0.693$; the nuisance parameter $Q(t) = 2 \log(2)t$. In addition, we assume that the censoring time follows a uniform distribution, and that it causes a 10% censoring proportion in the non-cure patients. We consider the two cases of the PPV = 0.8 and 0.9, and carry out one thousand replications for each case. The simulation results are presented in Table 1. The performance of the two methods on γ_0 are comparable, but on γ_1 and β_1 the one ignoring the diagnostic error has a quite significant bias compared to the proposed one even as $n = 1200$.

5. CONCLUSIONS

The enrichment design helps the development of new targeted therapies in treatment of some patients with specific molecular targets in clinical trials. However, the misclassification of patients' molecular statuses by a diagnostic device may lead to the underestimation of the actual treatment effect and the wrong conclusion on the benefit of the targeted therapies. To overcome the underestimation caused by the diagnostic error, we use the EM approach to deal with the uncertain patients' molecular statuses. Also, we consider a flexible transformation cure models, which contains the commonly-used proportional hazards models, to handle the survival data with a cure fraction. We demonstrate by simulation that the proposed method can not only estimate the treatment effect with a much smaller bias compared to one ignoring the diagnostic error but also infer the actual PPV.

REFERENCES

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. Journal of the American Statistical Association 47 501-515.

Chen, Y.-H. (2009). Weighted Breslow-type and maximum likelihood estimation in semiparametric transformation models. Biometrika 96, 591-600.

Chen, C.-F., Lin, J.-R. & Liu, J.-P. (2013). Statistical inference on censored data for targeted clinical trials under enrichment design. Pharmaceutical Statistics 12, 165-173.

Farewell V. T. (1986) Mixture models in survival analysis: are they worth the risk? Canadian Journal of Statistics 14, 257-262.

Kuk, A. Y. C. & Chen, C.-H. (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79, 531-541.

Lu, W. & Ying, Z. (2004). On semiparametric transformation cure model. Biometrika 91, 331-343.

Murphy, S. A., Rossini, A. J. & Van Der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. Journal of the American Statistical Association 92, 968-976.

Peng, Y., Dear, K. B. G. & Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. Statistics in Medicine 17, 813-830.

Peng, Y. & Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation.

Biometrics 56, 237-243.

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. Biomet- rics 56, 227-236.

Tsai, C.-A., Lee, K.-T. & Liu, J.-P. (2016). Addressing loss of efficiency due to misclassification error in enriched clinical trials for the evaluation of targeted therapies based on the Cox proportional hazards model. PLOS one DOI:10.1371/journal.pone.0153525.

Table 1: Results of parameters estimation in the PH model with $Q(t) = 2\log(2)t$ and a 10% conditional censoring proportion. The median survival times of $Su(t|\cdot)$ in the subgroups of $\xi z = 0$ and $\xi z = 1$ are 0.5 and 1.0, respectively, and the non-cure proportions are 0.35 and 0.20. That is, $\gamma 0 = -0.619$, $\gamma 1 = -0.767$, $\beta 1 = -0.693$, $\pi 0 = 0.35$, $\pi 1 = 0.20$ and $HR = 0.5$.

p = 0.8			p^	γ^0	γ^1	β^1	π^0	π^1	H^R
n = 600	New	Bias	0.044	0.003	0.016	0.089	0.001	0.005	0.07
		SD	0.075	0.135	0.262	0.296	0.03	0.038	0.173
	Ignore	Bias	0.200*	0.005	0.147	0.232	0.002	0.027	0.148
		SD	0.000*	0.135	0.221	0.233	0.031	0.033	0.153
n = 1200	New	Bias	0.029	0.003	0.019	0.07	0.001	0.005	0.046
		SD	0.054	0.098	0.183	0.183	0.022	0.026	0.104
	Ignore	Bias	0.200*	0.005	0.161	0.229	0.001	0.029	0.137
		SD	0.000*	0.098	0.152	0.165	0.022	0.022	0.106
p = 0.9			p^	γ^0	γ^1	β^1	π^0	π^1	H^R
n = 600	New	Bias	-0.005	0	-0.024	0.027	0.001	-0.001	0.031
		SD	0.06	0.133	0.262	0.255	0.03	0.037	0.152
	Ignore	Bias	0.100*	0.002	0.073	0.131	0.001	0.014	0.088
		SD	0.000*	0.133	0.232	0.248	0.03	0.034	0.146
n = 1200	New	Bias	-0.016	0.003	-0.031	0.005	0.001	-0.003	0.01
		SD	0.043	0.098	0.18	0.176	0.022	0.025	0.094
	Ignore	Bias	0.100*	0.004	0.075	0.13	0.001	0.014	0.078
		SD	0.000*	0.099	0.156	0.173	0.022	0.022	0.1

* The method ignoring the diagnostic error misleads p = 1.

A Powerful Method to Meta-Analysis for Testing No Treatment Effects

Kuang Fu Cheng*
School of Business, Asia University, Taichung, Taiwan, and Biostatistics Center,
Taipei Medical University, Taipei, Taiwan - kfcheng@tmu.edu.tw

Abstract

In a meta-analysis of multiple trials, a fundamental problem is to test whether a new treatment of interest is better than the placebo or an active treatment. Because the data of a meta-analysis may be heterogeneous, fixed-effects approach and random-effects approach are often used in the literature. The fixed-effects approach assumes that the all trial effects are the same, while the random-effects approach usually assumes the effects are random and follow a normal distribution. Under the random-effects model, Han, B. & Eskin, E. (2011) suggested a test for testing the averaged effect and variance of the effects are both zero, that is, all treatment effects (log odds ratios) are zeros. In this paper, we propose a new approach for solving the same problem. The novel method has greater advantage in that no distributional assumption of normality is required. Furthermore, a simulation study indicated that it was more powerful than Han-Eskin test under wide range of simulation conditions.

Keywords: fixed-effects; Power; Random-effects; treatment.

Journal of Economic Literature (JEL) classification:C12, C18, C33

1. INTRODUCTION

Meta-analysis has been considered as a powerful methodology for analyzing results from multiple independent trials that are relevant to specific research questions. Typically, in a meta-analysis, the result of each trial is summarized by means of an effect size and its variance estimate, and fixed-effects model or random-effects model (see DerSimonian, R. & Laird, N. (1986)) are used to obtain a global index about the effect magnitude of the studied relation, together with a confidence interval and its statistical significance. A fixed effect model assumes a single effect parameter value is common to all trials, though the assumption is seldom reliable in biomedical and social sciences. On the other hand, random-effects model often assumes that the effect parameters of the underlying trials follow a normal distribution.

An important first step in meta-analysis is to determine which trials to include in the analysis, since the effect size of the two treatments comparison often varies between trials. This phenomenon is referred to as between-study heterogeneity. Thus, in a meta-analysis, it is important to choose trials so that the combined sample size and the degree of heterogeneity can be balanced and to report how consistent of the results are among different trials. A test of heterogeneity of the trial effects often uses Cochran's Q statistic (Cochran, W.G. (1937)). This test is susceptible to the number of trials included in the meta-analysis. To overcome the shortcomings of the Q test, an I^2 index was proposed by Higgins et al. (2003). I^2 index is similar to an intra-class correlation in cluster sampling, which can be interpreted as the percentage of total variation of the effect sizes, not due to chance but because of the true heterogeneity.

Tests of heterogeneity are also commonly used to decide methods for combining trials. If the test is not significant, the fixed-effects model is used, otherwise, the random-effects model is more appropriate. To test whether the effects are significant, a further testing procedure based on confidence interval for the common effect can be applied, if fixed-effects model is suggested. Unfortunately, confidence interval approach is asymptotic-based, and its coverage probability could be very different from the designated value when the number of studies is not sufficiently large. On the other hand, under a random-effects model, Han, B. & Eskin, E. (2011) proposed a statistic for jointly testing zero mean effect size and zero effect variation. Their statistic can be written as a sum of two components. The first component is a likelihood ratio statistic for testing zero mean effect under a fixed-effects model, and the second component is a statistic for testing heterogeneity under a random-effects model.

The statistical significance of the Han-Eskin test can be accurately assessed by permuting the data within each study, for example. However, this test may be underpowered, if the assumption of normality distribution for the random effects fails to be satisfied. It is therefore desirable to consider alternative tests without the need of this assumption. In this paper, we discuss the application of the test, proposed by Cheng et al.(2014) in meta-analysis. The test is originally designed for detecting association signals in genetic studies. We focus on the trials with dichotomous outcomes in treatment and control arms, and log-odds-ratios being the effects of interest. Since the treatment and control sample sizes are known in each study, thus the information about sample log-odds-ratio and its error estimate is equivalent to the information given by a 2x2 table. That is, the numbers of positive outcomes in treatment and control arms are also known.

The test by Cheng, et al (2014). is a procedure based on testing zero proportion of the trials in the meta-analysis showing positive effect, against the alternative that the proportion is nonzero. Conceptually, this approach is very different from those tests based on testing specific parameter values based on a distributional assumption for the random effects. The two approaches use different parameters to describe the same hypotheses of interest. The first approach is designed to test proportionality parameter equal zero, without normality assumption. On the other hand, the random-effects model approach tests mean and variance parameters equal to zero, with normality assumption.

In the Methods section, we briefly introduce the test of Han-Eskin and that by Cheng et al. (2014). Our reports of the simulation results are given in the section of Simulations and Results. Some discussions are given in the Conclusion section.

2. METHODS

We focus on K placebo controlled trials with binary outcomes in a meta-analysis. . Usually, in meta-analysis, the effect (log-odds-ratio) estimate X_i and its variance estimate v_i, along with the sample size are given. Thus, in essence, we have available the data of 2x2 table for each trial. Suppose that for the k th trial, we had N_{1k} subjects received the new treatment and among them, there were n_{1k} subjects with positive outcome. The corresponding numbers in the placebo treated sample are denoted by N_{0k} and n_{0k}. Define $\theta_k = P_{1k}/(P_{1k}+P_{0k})$ where P_{1k} (P_{0k}) is the probability of having positive outcome for the subjects receiving treatment (placebo). Therefore, if the treatment is not effective, then θ_k =0.5, otherwise, θ_k >0.5 is likely. The relationship between the effect parameter and usual odds ratio, OR_k, is given by $\theta_k = OR_k \ / \{OR_k(1+p_{0k})+(1\text{-}p_{0k})\}$. For a particular treatment of interest, the proportion of $\theta_k > 0.5$ in trials of meta-analysis is the parameter of concern, denoted by α. Cheng et al. (2014) proposed to test $\alpha = 0$, which is equivalent to no treatment effect in meta-analysis. The basic concept of their approach is simple. In the case of no treatment effect, they considered θ_k as a random variable following a degenerate distribution at 0.5. In the case that the treatment is effective, they considered θ_k as a random variable following a uniform distribution between 0.5 and B_k, where the upper bond $B_k = \frac{OR^*}{OR^*+1+p_{0k}(OR^*-1)} < 1$, and OR^* is a chosen upper bound of all odds ratios of all trials included in meta-analysis. Since OR^* is known and p_{0k} can be estimated from the data of the control sample, B_k can be considered as a known quantity. The marginal probability of n_{1k} is proportional to

$$p(n_{1k} \mid \alpha) = \alpha(B_k - 0.5)^{-1} \times \int_{0.5}^{B_k} \{p_{0k}\theta_k/(1-\theta_k)\}^{n_{1k}} \left[1-\{p_{0k}\theta_k/(1-\theta_k)\}\right]^{N_{1k}-n_{1k}} d\theta_k + (1-\alpha)p_{0k}{}^{n_{1k}} \ (1-p_{0k})^{N_{1k}-n_{1k}}.$$

The test of Cheng, et. al. is based on the likelihood ratio statistic $T_{REM} = 2\log\{L(\hat{\alpha})/L(0)\}$ where $L(\alpha) = \prod_{k=1}^{K} p(n_{1k} \mid \alpha)$ and $\hat{\alpha}$ is the parameter value maximizing the likelihood function over [0,1]. If the likelihood ratio statistic is too large, then the null hypothesis is rejected. The p-value of the statistic can be computed based on permuting the data within each trial or using bootstrap method. It also can be approximated by using asymptotic distribution.

The Han-Eskin's method is based on the random-effects model. Their approach assumed the true effect size of each trail is sampled from a normal distribution with true mean effect size μ and variance σ^2. Under these conditions, the likelihood is given by $\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi(v_k+\sigma^2)}} \exp(-\frac{(X_k-\mu)^2}{2(v_k+\sigma^2)})$, and under the null hypothesis $\mu = \sigma^2 = 0$, the likelihood is $\prod_{k=1}^{K} \frac{1}{\sqrt{2\pi(v_k)}} \exp(-\frac{X_k^{\ 2}}{2v_k})$. The maximum likelihood estimates $\hat{\mu}$ and $\hat{}$ can be obtained by iterative method suggested by Hardy, R. J. & Thompson, S. G.(1998). The Han-Eskin test is based on the likelihood ratio statistic

$\sum_{k=1}^{K} \log(\frac{v_k}{v_k+\hat{\sigma}^2}) + \sum_{k=1}^{K} \frac{X_k^2}{v_k} + \sum_{k=1}^{K} \frac{(X_k-\hat{\mu})^2}{v_k+\sigma^2}$. The p-value of the likelihood ratio test also can be approximated by using the permutation method.

3. SIMULATIONS AND RESULTS

We compare the performance of two tests T_{REM} and Han-Eskin by simulations. The number of trials$K$ in the meta-analysis was either 5 or 15. All trials were balanced in the sense that they had same number $N_{1k} = N_{0k} = N_k$ of subjects receiving placebo or the new treatment. The sample sizes were 30, 50, 90 or 150, ranging from small trail size to moderate trial size. In each trial, we assumed the corresponding effect size (log-odds-ratio) θ_k, was generated from a normal distribution with mean μ and variance σ^2. In the simulations, μ =0, 0.25 or 0.5 and σ = 0, 0.1, 0.3, 0.1 μ or 0.3 μ were assumed. Also P_{0k}'s, the probability of having positive outcome due to taking placebo treatment in trial k, were assumed to be the same and equal to 0.1, 0.3 or 0.5. The same probability for the new treatment was given by $p_{1k} = \theta_k O_k /(1+\theta_k O_k)$,where $O_k = p_{0k} / (1-p_{0k})$. The data in the ith trial were generated by $B(N_k, p_{0k})$ and $B(N_k, p_{1k})$.

In each simulation, the number of replications for computing type I errors and power was 2000 and significance level was 5%. Further, in each replication, the number of permutations was 2000 for computing p-values.

The simulation results are given in Table 1. Scenario 1 considered the case of small number of trials and small sample size. In this case, the type I errors of two tests were well-controlled within acceptable ranges. The tests had greater power when P_{0k} was closer to 0.5 in general. Also, the power was greater when the mean and/or variance of the normal distribution became larger. However, T_{REM} seemed to have better powerful performance under all conditions studied in the simulations. In addition, using T_{REM} test seemed to achieve greater percentage of power improvement over Han-Eskin test under the fixed-effects model. Scenario 2 and 3 considered the similar cases as Scenario 1, but in the former simulations, the sample sizes were increased, and in the latter situations, the number of trials was increased. However, the total sample size of all trials in Scenario 2 was the greatest. Under both situations, the power of two tests was simultaneously enlarged. But comparing the power performances under scenarios 1 and 2, and that under scenario 1 and 3, the relative improvement due to increasing sample size but not trial number seemed greater. In any case, T_{REM} had the best power performance.

4. CONCLUSION

Meta-analysis provides an objective way of combining information from separate studies looking at the same clinical question and has been applied most often to treatment effects in randomized clinical trials. The usual approach of meta-analysis is to provide a global estimate of effect size, along with a confidence interval. Han, B. & Eskin, E. (2011) considered the usual normal-based random-effects model and proposed a joint test of zero mean and zero variance of the random effect. If the test shows the evidence of no significance, one may conclude that comparing to the placebo, the new treatment has no effect. In this paper, we consider a different method by Cheng, et al. (2014), for testing no treatment effect. We show by simulations that the type I error of this test can be properly controlled by using permutation method. We also find the method to be more advantageous in that it is not only more powerful but also independent of model assumption. The latter property is important, since the validity of the model assumption for the random effects is often questionable.

Table 1. Empirical Type I errors and Power of the Competing Tests

			Scenario 1: 5 trials; N_k=50, k=1,2, N_k=30, k= 3,4,5		Scenario 2: 5 trials; N_k=150, k= 1,2, N_k=90, k= 3,4,5		Scenario 3: 15 trials; N_k=50, k= 1,2,...,5 N_k=30, k= 6,7,..,15	
p_{0k}	μ	σ	Han-Eskin	TREM	Han-Eskin	TREM	Han-Eskin	TREM
0.1	0	0	0.0480	0.0575	0.0510	0.0485	0.0450	0.0475
	0	0.1	0.0595	0.0850	0.0555	0.1005	0.0505	0.0730
		0.3	0.1175	0.1840	0.2690	0.4255	0.1505	0.2720
	0.25	0	0.1285	0.1915	0.2680	0.3570	0.2520	0.3495
		0.1μ	0.1385	0.2060	0.3240	0.4100	0.2675	0.3665
		0.3μ	0.1635	0.2420	0.3705	0.4660	0.3120	0.4160
	0.5	0	0.3760	0.5115	0.8150	0.8595	0.7575	0.8255
		0.1μ	0.4315	0.5550	0.8730	0.9055	0.8045	0.8555
		0.3μ	0.5455	0.6545	0.9480	0.9635	0.8820	0.9240
0.3	0	0	0.0480	0.0425	0.0520	0.0560	0.0510	0.0495
	0	0.1	0.0670	0.0965	0.0940	0.1515	0.0755	0.1050
		0.3	0.2020	0.3195	0.5540	0.7180	0.3415	0.4455
	0.25	0	0.2185	0.3155	0.5200	0.5565	0.4505	0.5330
		0.1μ	0.2365	0.3370	0.5795	0.6140	0.5115	0.6020
		0.3μ	0.2895	0.4050	0.6890	0.7265	0.5625	0.6315
	0.5	0	0.6470	0.7340	0.9890	0.9880	0.9665	0.9705
		0.1μ	0.7175	0.7890	0.9965	0.9965	0.9800	0.9825
		0.3μ	0.8225	0.8850	0.9995	1.0000	0.9925	0.9955
0.5	0	0	0.0565	0.0555	0.0435	0.0400	0.0510	0.0505
	0	0.1	0.0685	0.0955	0.1020	0.1590	0.0700	0.0905
		0.3	0.1805	0.2955	0.5690	0.7275	0.3489	0.4285
	0.25	0	0.2350	0.3330	0.5785	0.6115	0.5085	0.5735
		0.1μ	0.2490	0.3515	0.6310	0.6675	0.5480	0.6135
		0.3μ	0.2900	0.3990	0.7510	0.7855	0.5985	0.6700
	0.5	0	0.6905	0.7715	0.9880	0.9915	0.9755	0.9825
		0.1μ	0.7380	0.8170	0.9975	0.9985	0.9795	0.9840
		0.3μ	0.8300	0.8975	0.9995	0.9990	0.9945	0.9985
0.7	0	0	0.0455	0.0535	0.0505	0.0450	0.0480	0.0440
	0	0.1	0.0560	0.0800	0.0805	0.1455	0.0645	0.0915
		0.3	0.1510	0.2495	0.4265	0.6125	0.2530	0.3145
	0.25	0	0.1835	0.2755	0.4780	0.5475	0.4265	0.5305
		0.1μ	0.2155	0.3170	0.5460	0.6105	0.4575	0.5555
		0.3μ	0.2545	0.3730	0.6135	0.6935	0.5105	0.6165
	0.5	0	0.5675	0.6810	0.9690	0.9750	0.9340	0.9530
		0.1μ	0.6230	0.7345	0.9720	0.9800	0.9535	0.9690
		0.3μ	0.7055	0.8100	0.9925	0.9950	0.9720	0.9860
0.9	0	0	0.0490	0.0505	0.0510	0.0470	0.0500	0.0465
	0	0.1	0.0525	0.0665	0.0620	0.0885	0.0510	0.0635
		0.3	0.0680	0.1135	0.1440	0.2590	0.0555	0.1020
	0.25	0	0.1005	0.1575	0.2365	0.3340	0.2130	0.3155
		0.1μ	0.1095	0.1760	0.2670	0.3545	0.2265	0.3315
		0.3μ	0.1210	0.1995	0.3075	0.4200	0.2465	0.3635
	0.5	0	0.2550	0.3675	0.6385	0.7385	0.5850	0.7210
		0.1μ	0.2855	0.3955	0.6830	0.7815	0.6155	0.7455
		0.3μ	0.3270	0.4490	0.7625	0.8560	0.6470	0.7725

REFERENCES

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7:177-188.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society, 4:(Supplement) 102-118.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. British Medical Journal, 327(7414), 557–560.

Han, B. & Eskin, E. (2011). Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. The American Journal of Human Genetics 88, 586–598.

Hardy, R. J. & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine, 17:841-856.

Cheng, K.F., Lee, J.Y., Zheng, W. and Li, C. (2014). A powerful association test of multiple genetic variants using a random-effects model. Statistics in Medicine, 33: 1816–1827.

CPS03:
Macroeconomic Statistics (1)

Evaluation of Inflation Targeting among ASEAN Countries, Does It Have Impact on Unemployment and Economic Growth

Chindy Saktias Pratiwi*
Statistics Indonesia, Jakarta, Indonesia – chindy.pratiwi@bps.go.id

Putu Wira Wirbuana
Statistics Indonesia, Jakarta, Indonesia – putuwiraw@bps.go.id

Abstract

In recent years, Inflation Targeting (IT) has emerged as the monetary policy framework in countries around the world. This policy firstly applied in 1990 and then adopted by many countries across the world. IT is a monetary policy framework with inflation rate as the operational target. By using this monetary policy, the inflation rate is expected to be low and stable. However, there are pros and cons among economist in applying IT. The pros side argued that by adopting IT, a country could rise the chances of attaining and maintaining low and stable inflation rate and reduce the uncertainties. On the other hand, the cons side argued that this monetary policy ignores the effort to reduce unemployment rate and rise economic growth. Therefore, this study aims to assessing the implication of IT implementation on unemployment and economic growth. The analysis of this study focused on ASEAN countries before and after adopting IT. Two models are constructed for each unemployment rate and economic growth using Panel Data, Fixed Effect Seemingly Unrelated Regression. The period of the study is from 1996 to 2014. Based on the model estimation result, it is showed that the implementation of IT had positive impact on unemployment and economic growth with different response on economic performance in each country. A lower rate of inflation could increase the unemployment rate by 0.08 per cent. However, when IT was implemented, the unemployment rate could be reduced by 1.17 per cent relative to pre-targeting period. Meanwhile, the impact of IT itself could reduce the unemployment rate by 1.2 per cent. Besides, when inflation rate decreased by 1 per cent, economic growth increased by 0.31 per cent. During the adoption of IT, this increment was boosted by 0.95 per cent. Therefore, Inflation Targeting could be an option for ASEAN emerging countries to make the economic better.

Keywords: Inflation Targeting (IT), Monetary policy, Unemployment, Economic Growth, Panel Data, Fixed Effect, Seemingly Unrelated Regression (SUR)

1. INTRODUCTION

In recent years, Inflation Targeting has emerged as the monetary policy framework in countries around the world. Inflation targeting was adopted as an alternative to the policy of targeting monetary aggregates or the exchange rate fixed (Dabousi, 2014). This policy firstly applied in 1990 by New Zealand. Then followed by many countries across the world such as Canada in 1991, the United Kingdom in 1992, Sweden in 1993, Finland in 1993, Australia in 1994 and Spain in 1994 (Mishkin, 1999).

The increasing of IT implementation is due to its successfulness in rising the credibility and transparency of central bank. Central bank could respond the short term fluctuation in economy without affecting its credibility (Lin & Ye, 2007). They would give public announcement of inflation targets along with their commitment to achieve and keep a low single digits level of inflation (Epstein and Yeldan, 2007). Mishkin (1999) stated there were several important advantages of IT implementation. IT enables monetary policy to respond to domestic economy's shock. It also allows central bank to use all available information to determine the best setting for monetary policy. Furthermore, by stabilizing the price it might lessen uncertainty and create a favourable economic environment for consumption and investment.

However, as this monetary framework growing, others start to concern about its effect on growth and unemployment. It is considered that IT will lead to low and unstable growth in output and employment

(Mishkin. 1999). There were trade-off between inflation and level of economic growth, which was keeping inflation on low level might cause reducing level of economic growth (Kovalchuck, 2012). Furthermore, some studies had failed to prove the impact of IT both on economic growth and unemployment. One of them showed by Ball and Sheridan (2005) revealed that IT did not affect either average level or volatility of output.

Despite the fact that this monetary policy became popular among countries, its influence economic performance is not clear yet. Furthermore, there is limited empirical study related to implementation of IT in Association of Southeast Asian Nation (ASEAN)[1]. Haque and Khan (1998) explained that there are several different methods in assessing the effects of policy implementation, two of them are the before-after approach and the with-without approach using control group. This study assess the impact of IT on economic growth and unemployment rate by comparing the condition period before and after this monetary policy implemented.

2. LITERATURE REVIEW

Many studies had been conducted in order to assess the performance of inflation targeting (IT). Clifton, Leon, and Wong (2001) examined the unemployment-inflation trade-off during IT implementation in OECD countries. The study compared pre and post-IT and showed that there was significant change enhancing effects of IT improve the unemployment-inflation trade off. A lower inflation rate were associated with smaller increases in unemployment.

According to Friedman (1977), studying on relationship between inflation and unemployment rate, inflation increased uncertainty due to the changes on relative prices and wages, leads to resource misallocation and lower levels of employment. Meanwhile, Furuoka and Munir (2011) assessed Phillips curve hypothesis in ASEAN Countries and found that there is no trade off between unemployment and inflation rate in ASEAN Countries.

On global financial crisis, the inflation targeter countries had less scares on deflation rather than Non-Inflation targeter countries. The inflation targeter countries had lower unemployment rate and relatively stronger growth performance than non inflation targeter countries (Filho, 2010).

On the other hand, the relationship between inflation and growth, according to the study of Anwar and Islam (2011) on ILO Publication, revealed that higher inflation is associated with higher growth. A recent study by the IMF, comparing inflation targeter and non-inflation targeter countries during the crisis, showed that though the advanced inflation targeter countries had higher GDP growth rates then their non-inflation targeter countries, there was no such difference for emerging countries (Filho, 2010). Svensson in Mitchell and Bill (2004) claimed that Inflation targeting not only reduces variability on inflation, but also reduces variability on growth.

Policy changes also affect the performance of IT. The effect of disinflation on unemployment and output growth will depend on how inflation expectations are formed and hoe expected inflation adjusts. A central bank that has established its credibility will be able to pursue a disinflation policy with a smaller decline in output or a smaller increase in unemployment because the expected rate of inflation will be lowered if disinflation is expected (Clifton, Leon, and Wong, 2011).

3. METHODS

Data

This study focused to evaluate the adoption of inflation targeting and its effect on unemployment and economic growth among ASEAN Countries. Data used in this study are based on annually frequency data from World Development Indicators The period study is from 1996 to 2014. Currently, three ASEAN countries are inflation targeter, which are Indonesia, Philippines and Thailand.

1 (a) Poon and Lee (2013) examined whether economies that do not systematically target inflation (non-inflation targeters) experience higher exchange rate volatility as compared with inflation targeters in 10 countries of the Association of Southeast Asian Nation (ASEAN) from 1990 to 2010; (b) Siregar and Goo (2009) evaluate IT policies in Indonesia and Thailand focusing on commitment of the monetary authorities and overall performances of IT regimes.

The variables that used in this study contain unemployment rate and economic growth as dependent variables and inflation targeting adoption as independent variables.

Model

The term panel data refers to the pooling of observations on a cross-section of households, countries, firms, etc. over several time periods (Baltagi, 2005). Gujarati (2004) explained that panel data gives more informative data, more variability, less collinearity among variables, more degrees of freedom, and more efficiency. Panel data can enrich empirical analysis in ways that may not be possible if we use only cross-section or time series data.

In general, there are three types of panel data models, which are pooled regression, fixed effect model, and random effect model. (i) Pooled Regression if Z_i contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common α and the slope vector β; (ii) Fixed Effects if Z_i is unobserved but correlated with x_{it}, then the least square estimator of β is biased and inconsistent as a consequence of an omitted variable. In instance, the model:

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it} \tag{1}$$

where $\alpha_i = z_i{}'\alpha$, embodies all the observable effects and specifies an estimable conditional mean. This fixed effect approach takes α_i to be a group-specific constant term in the regression model. The term "fixed" as used here indicates that the term does not vary over time, not that it is non-stochastic, which need not to be the case; (iii) Random Effects if the unobserved individual heterogeneity (u_i), however formulated, can be assumed to be uncorrelated with the included variables. The model may be formulated as:

$$y_{it} = x_{it}'\beta + E[z_i'\alpha] + \{z_i'\alpha - E[z_i'\alpha]\} + \varepsilon_{it} \tag{2a}$$

$$y_{it} = x_{it}'\beta + \alpha + \mu_i + \varepsilon_{it} \tag{2b}$$

That is, as a linear regression model with a compound disturbance that may be consistently, albeit inefficiently, estimated by least square. This random effects approach specifies that u_i is a group specific random element, similar to ε_{it} ε_{it} except that for each group, there is but a single draw that enters the regression identically in each period. To determine the best model among that three, we used Chow Test with F test statistics to choose between pooled regression and fixed effects with the null hypothesis is pooled regression is better than fixed effects model; Breush-Pagan Lagrange Multiplier to choose between random effects and pooled regression model with the null is pooled regression model is better than random effects model; and Hausman test to choose between fixed effects and random effects with the null hypothesis for Chow test is random effects model is better than fixed effects model. The next steps, after the model is selected, is determine the suitable estimation methods, Ordinary Least Square (OLS) or Feasible Generalized Least Square (FGLS), based on its residual variance-covariance structure, using Lagrange Multiplier (LM) test for homogeneity on variance-covariance structure and LM test for cross-sectional correlation on variance-covariance matrix.

This study uses before-after approach with two based model:

$$UNEMP_{it} = \beta_0 + \beta_1 INF_{it} + \beta_2 IT_{it} + \varepsilon_{it} \tag{3}$$

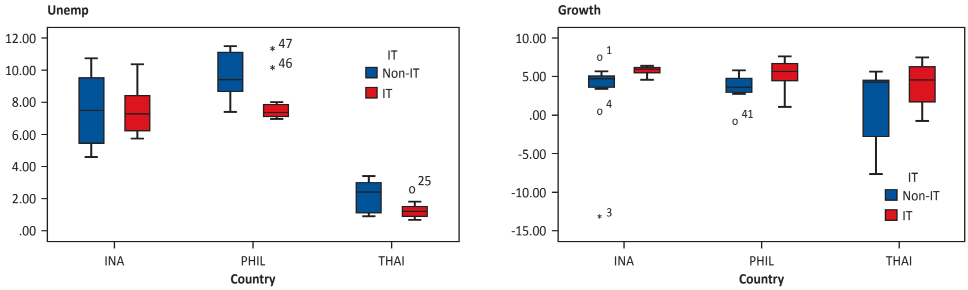
$$GROWTH_{it} = \beta_0 + \beta_1 INF_{it} + \beta_2 IT_{it} + \varepsilon_{it} \tag{4}$$

for t =1996, 1997, …, 2014 and i is each country

The dependent variable $UNEMP_{it}$ is an unemployment rate and $GROWTH_{it}$ is economic growth of country i at year t. While the explanatory variables comprise inflation that is represented by INF_{it} which is inflation rate of country i at year t; IT is dummy variable that value 1 if the country adopts inflation targeting and 0 if otherwise.

4. ANALYSIS

The situations of unemployment rate and economic growth are different for every country. Boxplots in Graph 1 showed the unemployment and growth conditions before and after implementing inflation targeting framework in each country.



Graph 1. Boxplot of unemployment rate (a) and economic growth (b) before and after inflation targeting adoption

Graph 1 (a) showed boxplot of unemployment rate pre and post-IT. After the implementation of IT, both Philippines and Thailand tend to have lower unemployment rate with lower dispersion than before adopting inflation targeting. It means that when inflation targeting implemented, it reduced variability on unemployment rate. This could be an indication that by adopting inflation targeting both countries could control unemployment rate. Meanwhile, Indonesia had similar unemployment rate before and after the adoption of inflation targeting but the unemployment rate was slightly higher after adopting inflation targeting. That condition followed with lower dispersion in post-IT. However, the dispersion of unemployment rate in Indonesia is still higher than other countries followed. Therefore it needs further examination to know how inflation targeting affected in Indonesia.

Graph 1 (b) showed boxplot of economic growth pre and post-IT. The economic growths in all countries tend to increase when the inflation targeting implemented. In Indonesia and Thailand, higher economic growth is followed by less variation on growth. However, in Philippines although it had higher growth after implementing inflation targeting the dispersion is slightly higher.

Table 1. Unemployment Rate and Economic Growth Before and After Implementing IT

Statistics	Indonesia		Philippines		Thailand	
	Pre - IT	Post - IT	Pre - IT	Post - IT	Pre - IT	Post - IT
Unemployment Rate						
Mean	7,45	7,56	2,16	1,27	9,69	8,04
Median	7,23	7,27	2,40	1,20	9,40	7,35
Std	2,27	1,62	1,12	0,51	1,63	1,67
Economic Growth						
Mean	2,86	5,72	0,86	4,14	3,50	5,40
Median	4,74	6,01	4,46	4,58	3,65	5,69
Std	5,87	0,59	5,81	2,71	2,10	1,82

Estimation models are constructed in order to evaluate performance of inflation targeting in ASEAN countries, Indonesia, Philippines, and Thailand. Using Fixed Effect Seemingly Unrelated Regression Model, the information on unemployment rate, growth, and inflation across countries over time are gathered.

Table 2. Fixed Effect Seemingly Unrelated Regression Model (Unemployment Rate Estimation Model)

Variable	Coefficient	Prob
C**	7.1726	0.0000
INF**	-0.0831	0.0135
IT**	-1.2495	0.0001
Fixed Effects (Cross)		
_INA—C	1.8186	
_THAI—C	-4.4913	
_PHIL—C	2.6727	
Test Summary	**Statistic**	**Prob.**
Chow Test**	232.9006	0.0000
Hausman Test**	219.9633	0.0000
LM Test Heteroscedastic**	27.5120	0.0000
LM Test Cross - sectional**	11.9860	0.0074
Normality Test		

	RESID_INA	RESID_THAI	RESID_PHIL
Jarque-Bera	0.477518	0.210997	3.113516
Probability	0.787605	0.899876	0.210818

Note: *:indicates statistical significance at the 10% level; **: indicates statistical significance at the 5% level;

According to the estimation model on Table 2, inflation rate statistically significant in affecting unemployment rate at significant value of 5 percent level. The coefficient of inflation variable means that every decreasing of inflation would increase employment rate by 0.08 percent. Moreover, at significant value of 5 percent level, the dummy variable of IT statistically significant, this indicates that when inflation targeting was adopted the disinflation policy could lower unemployment rate by 1.25 percent relative to pre-targeting period. This finding is in line with Clifton, Leon, and Wong (2001) revealed that inflation targeting improved the performance of unemployment-inflation trade-off which means that disinflations on average were associated with smaller increases in unemployment.

Table 3. Fixed Effect Seemingly Unrelated Regression Model (Economic Growth Estimation Model)

Variable	Coefficient	Prob
C**	5.4261	0.0000
INF**	-0.3110	0.0000
IT**	0.9540	0.0450
Fixed Effects (Cross)		
_INA—C	1.6953	
_THAI—C	-1.8950	
_PHIL—C	0.1996	
Test Summary	**Statistic**	**Prob.**
Chow Test**	41.6152	0.0000
Hausman Test**	21.5677	0.0000
LM Test Heteroscedastic**	27.5148	0.0000
LM Test Cross - sectional**	14.0274	0.0029
Normality Test		

	RESID_INA	RESID_THAI	RESID_PHIL
Jarque-Bera	0.994055	2.700664	2.056993
Probability	0.608336	0.259154	0.357544

Note: *:indicates statistical significance at the 10% level; **: indicates statistical significance at the 5% level;

The impact of IT on economic growth is showed by the estimation model on Table 3. According to the estimation model, inflation rate statistically significant at the 5 percent of significant level. The negative coefficient of inflation rate shows the reverse side relationship between growth and inflation rate. Based on Table (2), every increasing of 1 percent of inflation rate would decrease economic growth about 0,31 percent. Dummy variable of inflation targeting is also significant at significant level 5 percent. Differ with inflation rate, the coefficient of dummy variable of inflation targeting shows positive value. It means that when the countries adopt inflation targeting, the economic growth could increase 0,95 percent.

Indonesia, Thailand, and Philippines use flexible inflation targeting meaning that monetary policy aims to stabilise both inflation around its target and the real economy. Svensson (2009) explained that flexible inflation target is different from the strict inflation target where stabilizing the inflation rate is the only purpose without considering economic stability. Thus, under flexible IT gives a main role not only forecast inflation target but also other target variables such as output gap. These forecasts are conditional on the central bank's view of the transmission mechanism. The Bangko Sentral ng Pilipinas (2016), the Philippines central monetary authority, explained that by using flexible IT, it helps ensure that the design of the inflation target is more consistent with the country's economic circumstances, and safeguards the credibility of the IT. Moreover, according to Epstein and Yeldan (2012), the price stability will lead to sustained growth and employment creation.

5. CONCLUSIONS

The aim of this study is to assessing the implication of inflation targeting implementation on unemployment and economic growth in ASEAN countries by analysing the condition before and after implementing IT. This study covers targeter countries (Indonesia, Thailand, and Philippines) at the period 1996-2014. Using boxplot analysis our result of study shows the different tendency of unemployment rate and economic growth before and after inflation targeting implemented. Furthermore, using Panel Data, Fixes effect Seemingly Unrelated Regression, our study show that implementation of inflation targeting significantly effect on both unemployment and economic growth.

In term of labour market, the empirical result shows that IT could control unemployment rate. The variability of unemployment rate is also lower in post-IT period than pre-IT period.

Other finding in this study shows that implementation of IT influence economic performance. Trade of between inflation and economic growth did happened in pre-IT period. However, in post-IT period, if central bank keeps the inflation at low level, then economic growth will increase.

The lower unemployment rate and good performance of economic growth could be happened because Indonesia, Thailand, and Philippines use flexible inflation targeting meaning that monetary policy aims to stabilise both inflation around its target and the real economy. The price stability resulted from IT implementation could lead to sustained growth and employment creation.

This study is limited on focusing pre-post IT implementation and simply analyse its effect on economic performance. There are a lot of lacks in this study, especially on variables used in this study. However, this study could be a reference in choosing IT as an alternative policy framework for emerging countries. Therefore, it needs further examination on how inflation targeting policy works in ASEAN Countries and its effect on economic performance.

REFERENCES

Anwar, Sarah and Iyanatul Islam. (2011). Should Developing Countries Target Low, Single Digit Inflation to Promote Growth and Employment?. Employment Sector: Employment Working Paper No.87. Geneva: International Labour Office

Baltagi, Badi H. (2005). Econometric Analysis of Panel Data (third ed.). Chichester: John Wiley & Sons Ltd

Bangko Sentral ng Pilipinas. (2016). The BSP and Price Stability: Inflation Targeting September 2016. Department of Economic Research

Clifton, Eric V., Hyginus Leon, and Chorng-Huey Wong. (2001). Inflation Targeting and the Unemployment-Inflation Trade-off. IMF Working Paper WP/01/166

Daboussi, Olfa Manai. (2014). Inflation Targeting As a Monetary Policy Rule: Experience and Prospects. MPRA Paper No. 59336, posted 18 October 2014 13:56 UTC

Divino, Jose Angelo.2009. The Impact of Inflation Targeting on Unemployment in Developing and Emerging Economies. International Policy Centre for Inclusive Growth, United Nations Development Programme

Epstein, Gerald and A. Erinc Yeldan.(2007). Inflation Targeting, Employment Creation and Economic Development:Assessing the Impacts and Policy Alternatives. G-24 Policy Brief, 2009

Filho, Irineu de Calvaho. (2010). Inflation Targeting and the Crisis: An Empirical Assesment. IMF Working Paper WP/10/45

Friedman, Milton. 1977. Nobel Lecture: Inflation and Unemployment. Journal of Political Economy Volume 85, Number 3-June 1977

Furuoka, Fumitaka and Qaiser Munir. (2011). "Phillips Curve" in Selected ASEAN Countries: New Evidence From Panel Data Analysis. Sunway Academic Journal 6 page 89-102

Gujarati, Damodar N. (2004). Basic Econometrics [Fourth Edition. McGraw Hill, India.

Kovalchuk, Artur. (2012). Inflation Targeting in Developing Countries: Influence on Growth and Inflation Volatility. Thesis of Kyiv School of Economics

Lin, Shu and Haichun Ye.(2007). Does inflation targeting really make a difference? Evaluating the treatment effect of inflation targeting in seven industrial countries. Journal of Monetary Economics, 2007, vol. 54, issue 8, pages 2521-2533

Mishkin, Frederic S.. (1999). International Experiences with Different Monetary Policy Regimes. NBER Working Paper 6965

Mitchel, William and Anthea Bill. (2004). The Real Problem with Inflation Targeting. Working Paper No. 04-04 Centre of Full Employment and Equity July 2004 (Revised: September 2004).

Poon, Wai Ching and Yong Shen Lee. (2013). Inflation Targeting In Asean-10. South African Journal of Economics Vol. 82:1

Romdhane, Ikram Ben and Mensi, Sami. (2015). Assessing the Macroeconomic Effects of Inflation Targeting:Evidence from OECD Economies. International Journal of Economics, Finance and Management VOL. 4, NO. 1, January 2015 ISSN 2307-2466

Schmidt-Hebbel, Klaus and Martin Carrasco. (2016). The Past and Future of Inflation Targeting. Chile: Catholic University of Chile April 2016

Siregar, Reza Y and Siwei Goo. (2009). Effectiveness and Commitment to Inflation Targeting Policy:Evidences from Indonesia and Thailand. Munich Personal RePEc Archive 13 September 2009

Svensson, Lars E O. (2009). Flexible Inflation Targetng-Lessons From the Financial Crisis. Speech at the workshop "Towards a New Framework for Monetary Policy? Lessons from The Crisis" organized by the Netherlands

Between Hawks and Doves: Measuring Central Bank Communication

Stefano Nardelli*
European Central Bank, Frankfurt am Main, Germany – stefano.nardelli@ecb.europa.eu

David Martens
University of Antwerp, Belgium – david.martens@uantwerpen.be

Ellen Tobback
University of Antwerp, Belgium – ellen.tobback@uantwerpen.be

Abstract

Media scrutinise the ECB's communication very attentively to extract information on likely future moves of monetary policy rates, in particular after each press conference following monetary policy meetings. Assessing media's perception requires the translation of words into a quantitative indicator. In this paper, we propose the Hawkish-Dovish (HD) indicator which is computed out of a bulk of above 9,000 media reports on the ECB press conference (i.e. newspaper articles, newswires, etc.) and translates on a numerical scale the degree of "hawkishness" or "dovishness" of the ECB press conference tone perceived by media. We compare two different methods to calculate the indicator: one is based on the semantic orientation, while the other is constructed on a text classification with a Support Vector Machines classification model. We show that the latter method tends to provide more stable and accurate measurements of the perception on a labelled test set. Furthermore, we demonstrate the potential use of this indicator with some applications. We analyse the correlations with a set of interest rates and demonstrate the indicator's ability to anticipate monetary policy moves. Additionally, we use the Latent Dirichlet Allocation (LDA) algorithm to detect the dominant topics in the articles on ECB monetary policy decisions and conclude that the media's focus has shifted from the classic interest rate movements towards the non-standard monetary policy instruments. These findings provide decisive evidence in favour of using an advanced text mining classification model to measure more accurately how media perceive ECB communication.

Keywords: Central bank policies, Monetary policy, Communication, Big data, Data mining, Quantitative methods.

1. INTRODUCTION

During the financial crisis, monetary policy interest rates reached levels close to zero in most major central banks. As a result, communication has increasingly qualified as a powerful instrument able to influence financial market developments and drive expectations. Forward guidance, i.e. a form of verbal commitment to keep interest rates at a certain level for a period of time, possibly conditional to certain economic developments, became a genuine monetary policy instrument for many major central banks in addition to the standard toolkit of interest rates and refinancing instruments for the banking sector. Mirroring the increased role of communication in central banks, economic research has started to focus on how central bank communication adds information to that which is already contained in macroeconomic variables and how it may reveal policy makers individual preference functions. The overall objective is to enhance predictability of monetary policy decisions. The traditional approach focuses mainly on how information events affect both financial market developments and expectations of future policy moves.

Media usually analyse central banks' communication with a view to extract information about likely future moves of monetary policy rates (Hayo&Neuenkirch, 2015). Unlike market indicators, extracting relevant information on monetary policy from media reports is not straightforward and requires the translation of words into a numerical representation. Ideally, the resulting quantitative indicator should represent the media's perception of the central banks message and should be able to indicate whether expectations are more on the tightening side (hawkish perception) or rather on the loosening side (dovish perception).

In recent years, a number of studies have developed quantitative tools to measure (or, at least, to express on a numerical scale) the information contained in official central banks' statements (see e.g. KOF Swiss Economic Institute, 2007; Lucca&Trebbi, 2009; Hansen&McMahon, 2016; and Nechio&Regan, 2016). In other words, this approach can be characterised as an attempt to quantify information in communication, which is qualitative by nature. In this paper, we present an index measuring how the official monetary policy communications of the European Central Bank (ECB) are interpreted by the press, using data mining techniques. Concretely, we want to distinguish articles that perceive the ECB's statement as predominantly hawkish, i.e. pointing towards likely future policy rate increases, from those perceiving statements as dovish, i.e. hinting at declines or no increases in policy rates. We use two different techniques normally used in text analysis to measure the tone expressed in an article published after the press conference: Semantic Orientation (SO) and text classification using Support Vector Machines (SVM). The former measures how often the ECB is mentioned in a news article together with respectively hawkish and dovish words, the latter uses a classification model to predict the tone of a news article. We apply both methods to a data set of approximately 9,000 articles from the Dow Jones's Factiva global news database (which includes several million of items from nearly 33,000 sources), published between January 1999 and March 2016, in order to create the HD indicator (after the initials of hawkish and dovish). This indicator represents the average hawkishness or dovishness of the ECB as perceived by the media after each press conference.

This paper aims at offering on original contribution on various dimensions. First, it proposes a quantitative indicator computed from text sources, showing that data or text mining techniques can bring value in the process of monetary policy decision making (Section 2 and 3). Secondly, it compares two alternative methodologies and concludes in favour of the use of SVM (Section 4). Finally, it shows that one of the advantages of the SVM classification model there is its flexibility and the possibility to analyse in details terms most frequently used by the written media to describe monetary policy decisions (Section 4.3).

2. THE HD INDEX

The HD index was computed using two different methodologies: the first based on semantic orientation and the other on a support vector machine (SVM).

The first approach followed the methodology originally proposed by Lucca&Trebbi (2009), who analysed how Federal Open Market Committee (FOMC) monetary policy statements were reflected in press reports. Semantic orientation is a concept from computational linguistics and defines the position of a word or string of words between two opposite concepts. Turney (2002) has used Semantic Orientation (SO) to classify reviews as positive or negative. In practice, he defined the SO score as the difference between the strength of its association with a set of positive words and a set of negative words. In this way, the values taken on a numeric scale by the HD index reflects the relative frequency of two opposite concepts (in our case, dovish or hawkish) in a news article by counting the co-occurrences of strings referring to the ECB with words and expressions that are normally associated with these concepts, i.e.:

$$HD_t = \frac{\sum_{s(t)}(-1) \times I[s(t), R, D] + \sum_{s(t)}(1) \times I[s(t), R, H]}{\sum_{s(t)} I[s(t), R, D] + \sum_{s(t)} I[s(t), R, H]}$$

with *I*( ) being the indicator function that counts the co-occurrences of word *s(t)* with dovish (hawkish) words D (H) in for a given string of words R={European Central Bank, ECB, Mario Draghi, etc.}.

The alternative methodology applies text mining techniques, i.e. a Support Vector Machines (SVM) classification model following Provost&Fawcett (2013). This technique automatically looks for patterns in text documents to select the words with the highest discriminative power. As output, we get a linear model where each word is assigned a weight in favour of either class 1 (hawkish) or class -1 (dovish). A clear advantage is that this algorithm looks at every document as a whole and therefore tends to overcome the limitations of a predefined set of keywords.

To initialise the algorithm, a so-called training set was formed by selecting approximately 550 articles and pre-classifying them as hawkish or dovish. These articles were randomly selected from the available corpus. However, articles of uncertain classification were excluded not to introduce any bias and to enhance prediction accuracy. All articles were transformed into a 'bag-of-words' vector, i.e. $[t_0\, t_1 \ldots t_j \ldots t_n]$ containing all n unique words present in the training set, with t_j being the occurrences of word j in the article. From such vector a term-frequency matrix $tf(m,n)$ is built – with m being the number of articles and n the total number of words and – in which each cell (i,j) indicates the number of times a word j occurs in article i. To diminish the weight of the words that occur very frequently in the training set of articles, each term count is multiplied by the inverse document frequency (*idf*), which measures the frequency of a term across all documents (Weiss et al., 2010), i.e.:

$$idf(t,\ m) = \log \frac{\text{Number of articles in the training set } (m)}{\text{Number of articles in the training set where term } t \text{ occurs}}$$

The resulting *tf-idf* matrix is used as input to the SVM algorithm that searches for the "decision boundary" maximising the margin between the two classes, "dovish" and "hawkish" in this case. Linear SVM tries to solve the following optimization problem (Fan et al., 2008):

$$min_w \frac{1}{2} w^T w + C \sum_i max(1 - y_i w^T x_i, 0)^2$$

with w being the vector of the weights in the model, x_i and y_i representing respectively the input vector and the label of the i-th observation, while C is a cost parameter defined exogenously.

Articles are classified based on the following linear model:

$$f(x_i) = w_0 + w_1\, x_{i1} + w_2\, x_{i2} + \ldots + w_j\, x_{ij} + \ldots + w_n\, x_{in}$$

in which the weights w_j are estimated from above and x_{ij} the occurrence (frequency) of the j-th unique terms of the training set in the i-th article. The sign of the resulting decision value $f(x_i)$ is the predicted class the article belongs to (i.e. hawkish or dovish), whereas the value of $f(x_i)$ approximates an article's degree of hawkishness or dovishness. In this way, the larger the decision value is, the more certain the classifier is about the chosen class. Document of uncertain classification would therefore tend to have a decision value around 0, which can be interpreted as neutral tone.

3. DATA AND RESULTS

To compute our HD-index, we used articles extracted from Dow Jones Factiva. The selection was restricted to articles mentioning "ECB" or "European Central Bank" or the name of its President and limited to categories 'Major News and Business Publications: Europe' and 'Major News and Business Publications: US'. Only articles in English were included in the sample. Because the focus was the ECB press conference, a three-day window around the press conference (i.e. the day before, the very day and the day after) was used to extract articles. Using these criteria, we formed a corpus of slightly less than 9,000 articles, published between January 1999 and the latest press conference, covering therefore the whole history of the euro (in this paper, however, we included data until January 2016).

Figure 1 shows the news sources included in our data set and the respective proportion of articles in the corpus.

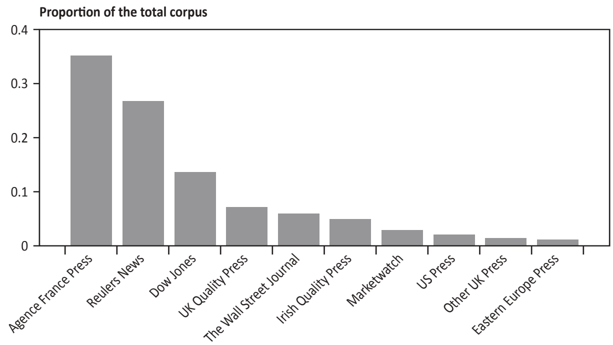


Figure 1: News Sources and Their Respective Proportion of Articles in the Corpus

Figure 2 shows the index computed using the two methodologies presented in Section 2 for the whole sample. The two series are contrasted with the evolution of the ECB monetary policy rate (namely, the interest rate on main refinancing operations) to contextualize them. Some relevant episodes are also indicated. Overall, the HD index generally appears to anticipate turning points in the ECB monetary policy stance and to be consistent with the different monetary policy cycles with a few exceptions. In other words, changes in the tone of communication that anticipate official interest rate movements generally appear to be understood appropriately. Specifically, the HD index rose prior to the two tightening monetary policy cycles started at the end of 1999 and at the end of 2005, but also HD declined ahead of the loosening cycles started in mid-2001 and the one caused by the deterioration of the financial crisis following Lehman's collapse in autumn 2008 (in particular the SVM index moved from a very hawkish to a neutral/dovish stance already in late 2007). During the crisis period the SVM HD index has consistently hovered around negative values thus indicating a prevailing dovish tone perceived in the ECB's communication. The decline to relatively low values in 2011 anticipated the marked loosening in monetary policy conditions taking the form of unprecedented standard and non-standard measures in order to bring inflation in line with its objective. While relatively low values were reached in the OMT phase, interestingly the HD index has become relatively less dovish in the period following the announcement of the extended APP, probably reflecting the belief that the ECB loosening policy reached a bottom or, at least, that what was done was sufficient.

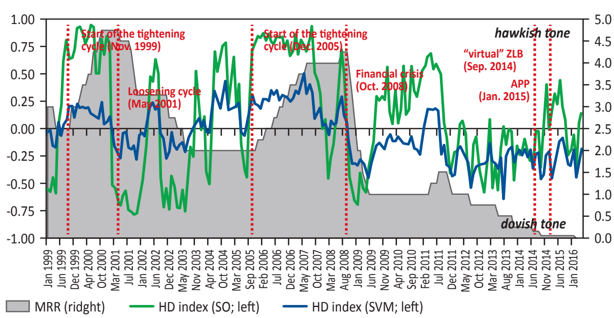


Figure 2: HD Index Computed with SO and SVM and ECB Official interest Rate

As regards differences in results between the two methodologies, although they broadly show a similar dynamics (the correlation between the two series is 0.78), the SO method tends to produce more marked volatile results and occasionally inconsistent results (for instance, towards the end of 2014).

4. VALIDATION CRITERIA

Whereas the SVM methodology overcomes some obvious limitations of the SO methodology (the preselection of a fixed set of words in the first place), there is no obvious validation method to compare the results obtained under the two methods. In this section we propose on three different criteria: (i) we measure how well both methods can predict whether the tone of an article is predominantly hawkish or dovish, (ii) we analyse in a deeper fashion the link of the HD index with some interest rates through correlations and (iii) we evaluate the indicators' performance qualitatively by extracting the most important topics mentioned in the news articles to investigate if they coincide with the indicators' values.

4.1. Performance Analysis

The classifications (scores) obtained from the two methodologies are compared with our own manual classifications using two performance metrics: Area under the receiver operating curve (*AUC*) and accuracy. *AUC* is a standard evaluation metric for classification models that represents a model's discriminative power by measuring to what extent positively labelled observations (hawkish) are ranked higher than negatively (dovish) labelled observations (Fawcett 2006). Unlike accuracy, which represents the percentage of correctly classified observations, *AUC* is able to deal with unbalanced distributions. Results are shown in Table 1.

Table 1: Performance Results of the SVM Model (with words) and SO to classify out-of-sample Articles

Method	AUC	Accuracy
SO	69.24%	62.00%
SVM	98.55%	92.00%

The SVM methodology outperforms SO when classifying the test articles as hawkish or dovish. We find a difference in accuracy of 30 percentage points in favour of SVM. The more advanced SVM-classification model is able to look at the broader context of an article, which results in a dovish classification for the example article.

4.2. Correlation Analysis

The visual inspection in Section 3 is substantiated here with the analysis of the correlation of the HD index and the ECB main refinancing rate. In addition, correlations were also computed with the Wu-Xia shadow rates to take into account the actual stance when official interest rates reached the zero lower bound and with the 12-month euro LIBOR rate, a money market interest rate reflecting expectations on official rates over the one-year horizon.

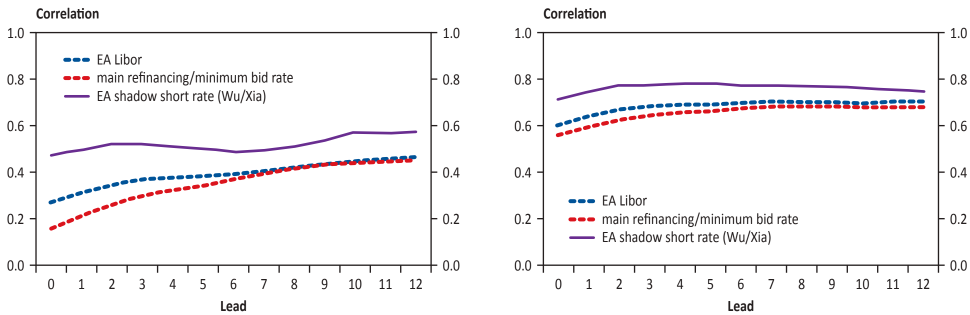


Figure 3. Correlations of the HD index (SO and SVM) with Various Interest Rates

Figure 3 shows correlograms from zero to twelve meetings ahead and can be interpreted as an indication on the lead properties of the HD index on actual monetary policy moves. It is important to stress that evidence of correlation should be interpreted here as a positive indication about the ability of the press to interpret correctly communication on future monetary policy stance (and of the index to represent significantly such interpretation). Positive correlation actually exists for the HD index and tends to be higher for the SVM methodology. The HD index tends to lead actual interest rates moves after 6 to 7 Governing Council meetings and then levels off for the SVM (8 to 9 meetings for SO). Furthermore, a relatively higher correlation (reaching a 0.8 peak) exists with the Wu-Xia shadow rates with a lead of 4 meetings in the case of SVM; also in this case SO tends to have a have a weaker correlation of only 0.6 and a lead of 11 meetings. A similar picture emerges from the correlation with the 12-month euro LIBOR rate with a lead of approximately 7 meetings and a peak correlation of about 0.7 for HD SVM and 0.45 for HD SO.

4.3. Topic Classification

In this case, a topic classification model was applied to the dataset divided into 13 periods selected on the basis of official interest rate developments (i.e. rising, stable, declining). Topics are derived using a Latent Dirichlet Allocation (LDA) as in Blei et al. (2003). This method assumes that each document can be represented by a mixture of topics, where the topic distribution follows a Dirichlet prior. Every topic has a probability of generating a set of words, which makes it possible to characterise the topic on the base of the related words. The topic subjects and a selection of the top 10 words for these topics can be found in Table 2 below:

Table 2. Top 10 words for the Most Frequently Topic

A	B	C	D	E	F	G	H	I	J	K
rate	rate	growth	Trichet	ECB	Greece	bond	price	cut	stock	rate
ECB	ECB	price	economy	Draghi	debit	market	growth	ECB	market	month
Trichet	cut	euro	market	bond	Greek	yield	remain	rate	share	Euribor
inflate	euro	economy	lend	crisis	ECB	debt	medium	point	index	market
price	bank	rate	record	cut	crisis	govern	risk	decision	bank	ECB
rise	zone	ECB	low	purchase	bond	Spain	inflate	move	expect	lend
bank	inflate	inflate	credit	monetary	default	purchase	stability	announce	rose	fixed
hike	expect	policy	recession	interest	bailout	crisis	expect	Thursday	trade	expect
interest	year	stability	financial	buy	plan	spread	develop	market	fell	overnight
increase	economy	recovery	crisis	debt	fund	Italy	continue	dollar	gain	EONIA

Figure 4 shows, along with the HD index, the periods and the three most dominant topics in each period. The topic classification confirm the hawkish peaks between May and July 2002 and between June 2004 and April 2005, where the articles mentioned both the rise in oil prices (topic C) and a possible rate hike (topic A).

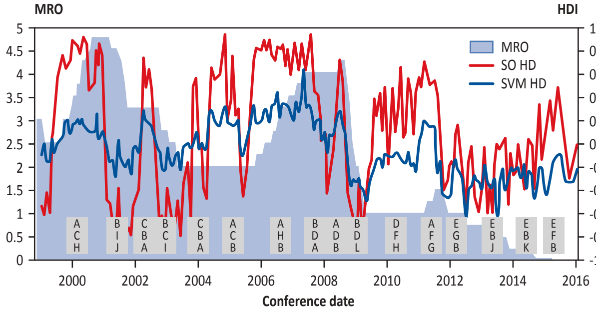


Figure 4. The HD index and Most Dominant Topics in 13 Periods

However, in the same periods topic B (rate cut) is discussed as well, indicating that there is a certain level of uncertainty or disagreement in periods of stable interest rates. During the tightening and easing cycles, the dominant topic is the one related to the interest rate decision, i.e. topic A or B. Interestingly, Figure 4 shows also an evolution in the topics discussed in articles from predominantly related to interest rate decisions (A and B) and price expectations (C) to crisis and bond purchases (E and F). For example, during the 2013-2014 easing cycle, the most discussed topic is the one on bond purchases, which are better captured by the SVM methodology while the SO would require the inclusion of expressions *ad hoc* to characterise monetary policy also in periods in which non-standard instruments are predominantly used.

5. CONCLUSIONS

In this paper, we present the development of a numerical indicator that represents the perceived degree of hawkishness or dovishness based on news articles reporting on ECB monetary policy decisions. The evidence provided indicates that the HD index can bring value to central bank communication. On a methodological ground, data mining techniques such as SVM prove to be superior to develop an index measuring the tone of communication as they are more flexible and may offer more insightful information. Specifically, we show that it is possible to extract and analyse the most frequently used topics in the text data, that for the HD index confirms a shift in media focus from the standard interest rate setting to non-standard monetary policy instruments. Apart from the results presented in this paper, the methodology can be easily applied to other communication outlets or extended to include other languages consistent with the multilingual nature of the Economic and Monetary Union to make it a fully-fledged tool to effectively monitor media reports and assessing effectiveness of ECB communication.

REFERENCES

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, 3, 993–1022.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern Recognition Letters*, 27, 861–874.

Hansen, S. & McMahon, M. (2016), "Shocking language: Understanding the macroeconomic effects of central bank communication", *Journal of International Economics,* 38th NBER International Seminar on Macroeconomics: S114–S133.

Hayo, B., & Neuenkirch, M. (2015), "Self-monitoring or reliance on media reporting: How do financial market participants process central bank news?", *Journal of Banking and Finance*, 59, 27–37.

KOF Swiss Economic Institute (2007), "KOF Monetary Policy Communicator for the Euro Area". URL: https://www.kof.ethz.ch

Lucca, D.O. & Trebbi, F. (2009), "Measuring central bank communication: an automated approach with application to FOMC statements". Technical Report National Bureau of Economic Research.

Nechio, F. & Regan, R. (2016), "Fed Communication: Words and Numbers", *FRBSF Economic Letters*, 2016–26.

Provost, F. & Fawcett, T. (2013), *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Turney, P. D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 417–424.

Weiss, S. M., Indurkhya, N. & Zhang, T. (2010), *Fundamentals of predictive text mining*, Springer Science & Business Media.

Wu, J. C., & Xia, F. D. (2016). "Measuring the macroeconomic impact of monetary policy at the zero lower bound", *Journal of Money, Credit and Banking*, 48, 253–291.

Cyclic Structural Equation Models and Their Identifiability

Mario Nagase*
Osaka University, Osaka, Japan – nagase@sigmath.es.osaka-u.ac.jp

Yutaka Kano
Osaka University, Osaka, Japan – kano@sigmath.es.osaka-u.ac.jp

Abstract

Causal analysis is a primary theme of science in all disciplines. There are some researches to address the problem. Among those, an approach is to apply cyclic Structural Equation Models. However, most of existing approaches are based on the premise that covariance between error variables often imply existence of unobserved confounders are mutually independent. In this paper, we study non-linear cyclic Structural Equation Models with normal errors allowing an error covariance between them so that more natural causal interpretation is delivered from the model.

Keywords: Causal analysis; Nonlinear function; Correlated error; Simulation study.

1. INTRODUCTION

Causal analysis is a primary theme of science in all disciplines. In traditional sense, it had been generally believed that it was impossible to elicit causal direction (and/or significance), namely "A causes B" or "B causes A" based on two-dimensional observed data on A and B. There are, however, some researches that have addressed such problems for decades, e.g., Spirtes (2000), Rechardson (1996). An approach to overcome the difficulty is to adopt models with non-normal error terms in Kano et al. (2003), and the model has been developed to be sophisticated ones called LiNGAM of Shimizu et al (2006), where a variety of extensions have been made including nonlinear functions with normal errors, e.g., Hoyer et al. (2009); Mooij (2011) and so forth. It should be noted that they always assume mutual independence between error variables. From the aspect of the basics of Structural Equation Models (SEMs) in Bollen (1989), however, covariance between error variables often imply existence of unobserved confounders that often get in the way of estimating true direct causal effects. Sometimes they try to detect causal direction by using Directed Acyclic Graph (DAG), but restricting causal graphs as DAG can result in significant misunderstanding on the nature of the phenomenon if cyclic causality is reasonable.

We consider a cyclic SEM as shown in Figure 1, and study the identification problem of the model, that is, under what conditions the variance and covariance parameters of error terms and parameters involved in the functions *f(x)* and *g(y)* are identified. Once the identifiability of the model is guaranteed, one can estimate the parameters and can obtain relevant information about determining the structure of the causal graph.

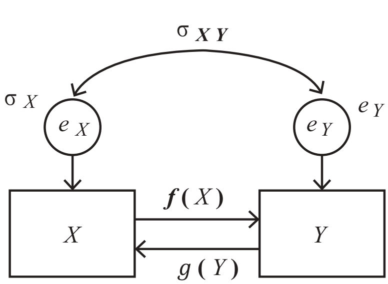


Figure 1. Bivariate cyclic SEM

Specifically we would like to study the non-linear cyclic SEM with normal errors allowing an error covariance as follows: $\begin{cases} Y = f(X) + e_Y \\ X = g(Y) + e_X \end{cases}, \quad \begin{bmatrix} e_Y \\ e_X \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$

where $f(x)$ and $g(y)$ are non-linear functions and $\mathbf{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}$.

Let us begin by showing how data are generated in the model and then proving that all the parameters in the model can be identified under certain conditions. Some simulation-based studies will be made as well.

2. DATA GENERATING PROCESS

Cyclic SEMs assume an equilibrium state for observed variables as shown in Bollen (1989). Suppose φ_0 and ψ_0 are arbitrary initial values for x and y, and consider the following iteration step

$$\begin{cases} \psi_{t+1} = f(\varphi_t) + e_Y \\ \varphi_{t+1} = g(\psi_t) + e_X \end{cases}, \qquad t = 0,1,2,\dots$$

An observed data x and y are then obtained as $\lim_{t\to\infty} \varphi_t$ and $\lim_{t\to\infty} \psi_t$, respectively. In order to guarantee their convergence, the necessary condition $|f'(x)g'(y)| < 1$ is required. Let $h(y) := y - f(g(y) + e_X) - e_Y$, and then $h'(y) = 1 - f'(x)g'(y)$. Under the necessary condition, $h(y)$ is strictly monotonically increasing. Thus, the solution in Eq.(1) is unique, if it exists.

3. IDENTIFIABILITY

It is well known that if the functions $f(x)$ and $g(y)$ in Eq.(1) are linear, the model can not be identified, as shown in Nagase et al. (2017). Hoyer et al. (2009) showed that if the functions are non-linear, the model is identified, under the condition that $\sigma_{XY} = 0$. However, identifiability of the models allowing error covariance between error variables still have not been proved. Then we shall show the identifiability of the simplest model as shown in Eq.(2):

$$\begin{cases} Y = af(X) + e_Y \\ X = bg(Y) + e_X \end{cases} \qquad \begin{bmatrix} e_Y \\ e_X \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}, \tag{2}$$

where $f(x)$ and $g(y)$ are known non-linear functions. The model includes 5 parameters $\theta = (a, b, \sigma_Y, \sigma_X, \sigma_{XY})$. It should be noted that the model has a fixed point (x_0, y_0) that satisfies

$$\begin{cases} y_0 = af(x_0) \\ x_0 = bg(y_0) \end{cases},$$

under the condition $|abf'(x)\, g'(y)|<1$.

Proposition

The model in Eq.(2) is identifiable, if $f''(x_0) \neq 0, g''(y_0) \neq 0, b\sigma_X^2 + ab\{2\sigma_{XY} f'(x_0) + a\sigma_Y^2 f'(x_0)^2\} \neq 0$ *and there exist certain data points* $x_1, \dots, x_k, y_1, \dots, y_k$ $(k > 3)$ *which satisfy*

$$\mathbf{0} \neq \begin{bmatrix} 1 & f'(x_1) & g'(y_1) & f'(x_1)g'(y_1) \\ 1 & f'(x_2) & g'(y_2) & f'(x_2)g'(y_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & f'(x_k) & g'(y_k) & f'(x_k)g'(y_k) \end{bmatrix} \begin{bmatrix} -\sigma_{XY} \\ -a\sigma_Y^2 \\ -b\sigma_X^2 \\ ab\sigma_{XY} \end{bmatrix}.$$

Proof.

The log-density function $\mathcal{D}$ of the model is expressed as

$$\begin{aligned} &\log \mathcal{D}\,(x, y; \theta | f, g) \\ &\quad \frac{1}{2}\log(\sigma_Y^2\sigma_X^2 - \sigma_{YX}^2) + [y - af(x) \quad x - bg\,(y)] \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}^{-1} \begin{bmatrix} y - af(x) \\ x - bg(y) \end{bmatrix} \\ &\quad + \log(1 - abf'(x)g'(y)). \end{aligned} \tag{3}$$

The model is identified if $\mathcal{D}(x, y; \theta|f, g) = \mathcal{D}(x, y; \theta^*|f, g) \Longrightarrow \theta = \theta^*$ holds. When we consider the first-order derivative of log D with respect to x and y at the fixed point (x_0, y_0), we have

$$\frac{\partial}{\partial x}\log \mathcal{D}(x_0, y_0; \theta|f, g) = \frac{-abf''(x_0)g'(y_0)}{1 - abf'(x_0)g'(y_0)} \tag{4}$$

and

$$\frac{\partial}{\partial y}\log \mathcal{D}(x_0, y_0; \theta|f, g) = \frac{-abf'(x_0)g''(y_0)}{1 - abf'(x_0)g'(y_0)}. \tag{5}$$

Eq.(4) and Eq.(5) imply that $ab = a^*b^*$, if $f''(x_0) \neq 0$ and $g''(y_0) \neq 0$ (These condition exclude the case where both *f(x)* and *g(y)* are linear). And when we take $\log \mathcal{D}\,(x_0, y_0; \theta|f, g)$ into consideration, the equality $\sigma_Y^2\sigma_X^2 - \sigma_{YX}^2 = {\sigma^*}_Y^2{\sigma^*}_X^2 - {\sigma^*}_{YX}^2$ can be obtained.

Next, we consider the second-order derivative of $\log \mathcal{D}$ with respect to x and y and we have

$$\frac{\partial^2}{\partial y \partial x}\log \mathcal{D}(x_0, y_0; \theta|f, g) = -\sigma_{XY} - a\sigma_Y^2 f'(x) - b\sigma_X^2 g'(y) + ab\sigma_{XY} f'(x)g'(y).$$

If there exist certain data points $x_1, \dots, x_k, y_1, \dots, y_k$ $(k > 3)$ which satisfy

$$\mathbf{0} \neq \begin{bmatrix} 1 & f'(x_1) & g'(y_1) & f'(x_1)g'(y_1) \\ 1 & f'(x_2) & g'(y_2) & f'(x_2)g'(y_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & f'(x_k) & g'(y_k) & f'(x_k)g'(y_k) \end{bmatrix} \begin{bmatrix} -\sigma_{XY} \\ -a\sigma_Y^2 \\ -b\sigma_X^2 \\ ab\sigma_{XY} \end{bmatrix},$$

the equalities $\sigma_{XY} = {\sigma^*}_{XY}, a\sigma_Y^2 = a^*{\sigma^*}_Y^2, b\sigma_X^2 = b^*{\sigma^*}_X^2$, and $b\sigma_X^2 = b^*{\sigma^*}_X^2$ can be obtained.

Finally, when we consider the second-order derivative of log D with respect to x

$$\frac{\partial^2}{\partial x^2}\log \mathcal{D}(x_0, y_0; \theta|f, g) = \sigma_X^2 + a\{2\sigma_{XY} f'(x_0) + a\sigma_Y^2 f'(x_0)^2\},$$

using the equality $ab = a^*b^*$, we can have

$$(a - a^*)[b\sigma_X^2 + ab\{2\sigma_{XY}\ \ '(x_0) + a\sigma_Y^2\ \ '(x_0)^2\}] = 0.$$

Then, with a condition $b\sigma_X^2 + ab\{2\sigma_{XY} f'(x_0) + a\sigma_Y^2 f'(x_0)^2\} \neq 0$, $a = a^*$ is obtained. Once $a = a^*$ is obtained, other parameters are obviously determined.

The proof would be extended to the model in Eq.(6).

$$\begin{cases} Y = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{f}(X) + e_Y \\ X = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{g}(Y) + e_X \end{cases}, \qquad \begin{bmatrix} e_Y \\ e_X \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}, \tag{6}$$

where

$$\mathbf{a} = [a_0 \ \cdots \ a_j]^{\mathrm{T}}, \qquad \mathbf{b} = [b_0 \ \cdots \ b_j]^{\mathrm{T}}$$
$$\boldsymbol{f}(X) = [f_0(X) \ \cdots \ f_j(X)]^{\mathrm{T}}, \qquad \boldsymbol{g}(Y) = [g_0(Y) \ \cdots \ g_j(Y)]^{\mathrm{T}},$$
$$|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{f}'(X) \cdot \boldsymbol{b}^{\mathrm{T}}\boldsymbol{g}'(Y)| < 1.$$

4. APPLYING TO TRIANGULAR-FUNCTION SERIES

We consider the functions $a^T f(X)$ and $b^T g(Y)$ are of following form :

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{f}(X) = a_0 + \sum_{n=1}^{l_f}\left[a_{2n-1}\cos\frac{2\pi nX}{T_f} + a_{2n}\sin\frac{2\pi nX}{T_f}\right],$$

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{g}(Y) = b_0 + \sum_{n=1}^{l_g}\left[b_{2n-1}\cos\frac{2\pi nY}{T_g} + b_{2n}\sin\frac{2\pi nY}{T_g}\right],$$

where l_f and l_g are orders of expansion, which determine the model complexity and the parameters T_f and T_f represent each of period same as what are in Fourier expansion. It should be noted that the model is also identifiable including T, and even when the values of l_f and l_g are small, i.e, $l_f = l_g = 2$, substantial variety of relationships can be in scope, if the linear approximation is considered to be good enough.

5. SIMULATION STUDIES

Here, we would like to consider the following model with $l_f = l_g = 2$,

$$\begin{cases} Y = a_0 + a_1 \cos\frac{2\pi X}{T_f} + a_2 \sin\frac{2\pi X}{T_f} + a_3 \cos\frac{2\pi 2X}{T_f} + a_4 \sin\frac{2\pi 2X}{T_f} + e_Y \\ X = b_0 + b_1 \cos\frac{2\pi Y}{T_g} + b_2 \sin\frac{2\pi Y}{T_g} + b_3 \cos\frac{2\pi 2Y}{T_g} + b_4 \sin\frac{2\pi 2Y}{T_g} + e_X \end{cases},$$

$$\begin{bmatrix} e_Y \\ e_X \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}.$$

5.1. Case 1

We set parameters as follows.

Table 1. Parameters set as true in Case 1

a_0	1.0	b_0	1.0	σ_Y	1.0
a_1	-0.3	b_1	-0.5	σ_X	1.0
a_2	0.3	b_2	0.6	σ_{XY}	0.5
a_3	0.2	b_3	-0.2	T_f	2π
a_4	0.2	b_4	-0.2	T_g	2π

We generated data set comprising n = 10,000 observations as shown in Figure 2, by means of forementioned iterative procedure starting from certain initial seeds φ_0 and ϕ_0 both of which are chosen from $\mathcal{U}[-\pi, \pi]$.

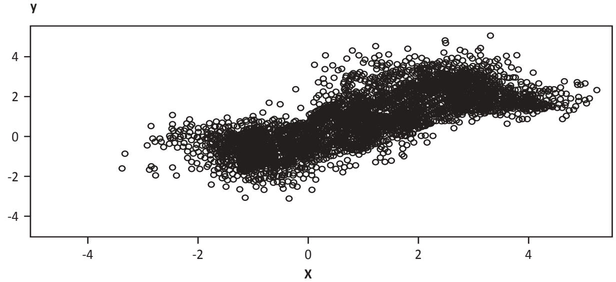


Figure 2. Data Generated

Each functions used in the model

$$\begin{cases} f(x) = a_0 + a_1 \cos\frac{2\pi x}{T_f} + a_2 \sin\frac{2\pi x}{T_f} + a_3 \cos\frac{2\pi 2x}{T_f} + a_4 \sin\frac{2\pi 2x}{T_f} \\ g(y) = b_0 + b_1 \cos\frac{2\pi y}{T_g} + b_2 \sin\frac{2\pi y}{T_g} + b_3 \cos\frac{2\pi 2y}{T_g} + b_4 \sin\frac{2\pi 2y}{T_g} \end{cases}$$

are also shown in Figure 3.

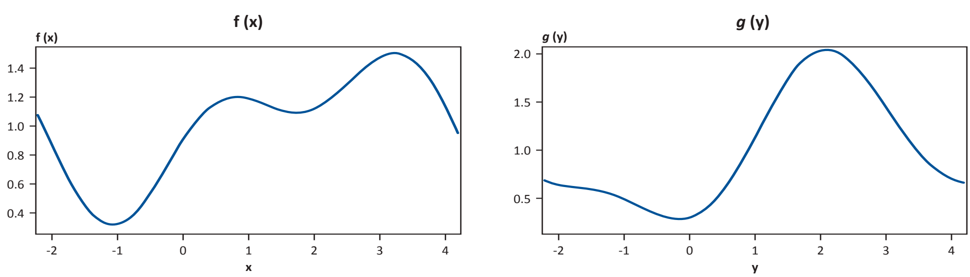


Figure 3. Functions

The loss function derived from the log-likelihood is as follows

$$
\begin{aligned}
&\mathcal{L}(\theta; x, y) \\
&\quad = const + \frac{1}{2}\log(\sigma_Y^2\sigma_X^2 - \sigma_{YX}^2) \\
&\quad + \frac{1}{2}[y - f(x) \quad x - g(y)]\begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{bmatrix}^{-1}\begin{bmatrix} y - f(x) \\ x - g(y) \end{bmatrix} \\
&\qquad - \log(1 - f'(x)g'(y)).
\end{aligned}
$$

We estimated parameters by using maximum log-likelihood estimation with quasi-Newton method. We set two sets of the following initial values for the iteration.

Table 2. Initial Values for quasi-Newton Method

a_0	a_1	a_2	a_3	b_4	b_0	b_1	b_2	b_3	b_4	σ_Y	σ_X	σ_{XY}	T_f	T_g
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6π 2.4π	1.6π 2.4π

Estimated results are shown in Table 3. It is found that the values of estimated parameters are almost the same over the choice of initial values and that they are enough close to the true values.

Table 3. Estimated Results in Case 1

	a_0	a_1	a_2	a_3	a_4	b_0	b_1	b_2	b_3	b_4
True	1.0	-0.3	0.3	0.2	0.2	1.0	-0.5	0.6	-0.2	-0.2
Est.	1.01	-0.32	0.31	0.21	0.20	1.02	-0.53	0.57	-0.18	-0.20

	σ_Y	σ_X	σ_{XY}	T_f	T_g	$\mathcal{L}$
True	1.0	1.0	0.5	2π	2π	27076.00
Est.	0.98	1.00	0.48	6.32	6.10	27070.65

Corresponding estimated functions are drawn in Figure 4.

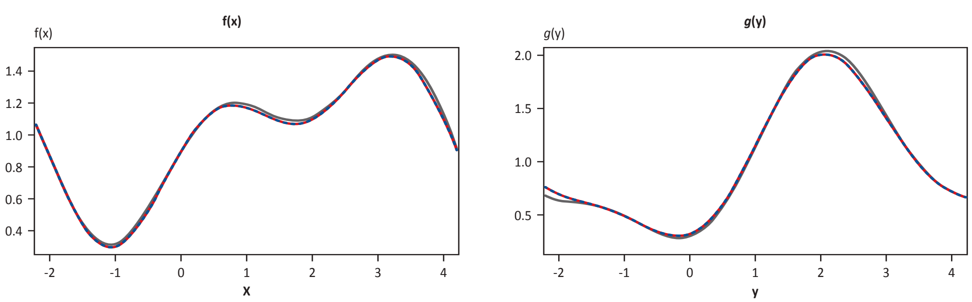


Figure 4. Estimated functions (black line: true, blue and red lines: estimates)

5.2. Case 2

Next, we set parameters as in Table 4, where no direct effects on Y → X are assumed.

Table 4. True Parameter values in Case 2

a_0	1.0	b_0	0.0	σ_Y	1.0
a_1	-0.3	b_1	0.0	σ_X	1.0
a_2	0.3	b_2	0.0	σ_{XY}	0.5
a_3	0.2	b_3	0.0	T_f	2π
a_4	0.2	b_4	0.0	T_g	2π

Estimated results are shown in Table 5.

Table 5. Estimated results in Case 2

	a_0	a_1	a_2	a_3	a_4	b_0	b_1	b_2	b_3	b_4
True	1.0	-0.3	0.3	0.2	0.2	0.0	0.0	0.0	0.0	0.0
Est.	1.06	-0.40	0.15	0.23	0.25	-0.02	-0.01	-0.03	0.01	0.06

	σ_Y	σ_X	σ_{XY}	T_f	T_g	$\mathcal{L}$
True	1.0	1.0	0.5	2π	2π	27080.90
Est.	1.04	0.99	0.55	6.73	16.03	27075.19

and corresponding estimated functions are drawn in Figure 5.

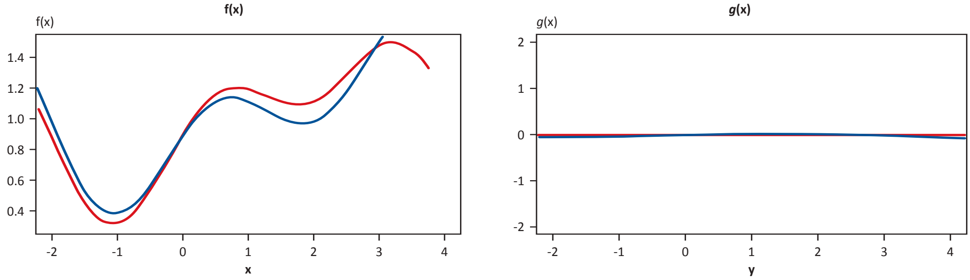


Figure 5. Estimated Functions in Case 2 (black line: true, blue line: estimates)

The estimated results might be felt bit different from true value, however, recall that the results are conducted based on the assumption of $l_g = 2$ whereas the true value of $l_g = 0$. Since l_f and l_g are just orders, one might want to tune the model from the perspective of model selection. In this paper, we compared the following cases in terms of AIC. The case #1 which corresponds to true setting is the lowest in AIC. It indicates that the model can completely detect true causal graph without taking the distinction of DAG and Directed Cyclic Graph (DCG) into account.

Table 6. Comparison of AIC in Case 2

Cases	l_f	l_g	b_0	Number of parameters	$\mathcal{L}$	AIC
#1	2	0	0	9	27077.19	54172.38
#2	2	0		10	27076.87	54173.74
#3	2	1		13	27075.29	54176.58
#4	2	2		15	27075.19	54180.38

6. CONCLUSIONS

It would be impressive that one can estimate all the parameters involved in the functions f and g as shown in Figure 3, by only using a data set in Figure 2. There are three aspects of advantage in the model. First, the model allows the covariance between error variables, and one will be able to identify it, whereas somehow almost all of existing methods presuppose mutual independencies between error variables. Then, it can be applied to many practical situations, especially for social science, e.g., Leenders (2002); Robbins (2012), where presupposing the existence of unobserved confounders would be better to capture the nature of the phenomenon. Second, one does not have to care about taking the distinction of DAG and DCG into consideration. In many cases, there is no way to rationally predispose whether the model is DAG or DCG beforehand. It could work for such situations. Third, assuming error terms as normal would make it easier to interpret the results, if one compares it to a model that use non-normal error.

In this paper we have dealt with only bivariate (p=2) cases, however, the model can apply to more complex models that include more variables (p>2), as long as computation is feasible.

ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Number 15H02669 and the Center for Mathematical Modeling and Data Science, Osaka University.

REFERENCES

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, 689-696.

Kano, Y., & Shimizu, S. (2003). Causal inference using nonnormality. In *Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion (AIC)*, 261-270.

Nagase, M., & Kano, Y. (2017). Identifiability of nonrecursive structural equation models. *Statistics & Probability Letters*, **122**, 109-117.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Leenders, R. T. A. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social networks*, **24**(1), 21-47.

Mooij, J. M., Janzing, D., Heskes, T., & Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, 639-647.

Richardson, T. (1996, August). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, 454-461. Morgan Kaufmann Publishers Inc..

Robbins, B. G. (2012). Institutional quality and generalized trust: A nonrecursive causal model. *Social Indicators Research*, **107**(2), 235-258.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **7**(Oct), 2003-2030.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag (2nd ed. MIT Press 2000).

Forecasting Inflation Rate in Sulawesi Using Generalized Space-Time Autoregressive (GSTAR) Models

Asrirawan*
Palopo Cokroaminoto University, Palopo, Indonesia – enalmantovani@gmail.com

Sumantri
Palopo Cokroaminoto University, Palopo, Indonesia – sumantri_sp@yahoo.com

Abstract

In this paper we apply a Generalized Space-Time Autoregressive (GSTAR) model to generate out of sample forecasts for the inflation rate in Sulawesi. We also apply VARIMA model and will be compared the forecast performance obtained from those two models. The forecast performance is evaluated using the mean squared errors (MSE) based on the last twelve actual data. The result showed the MSE of GSTAR ([4,1,0],) is smaller than VARIMA (4,1,0). Thus, the GSTAR ([4,1,0],) model might be considered as the best forecasting model for the inflation rate in Sulawesi.

Keywords: Forecasting, GSTAR, Inflation, VARIMA

1. INTRODUCTION

Inflation is a complex economic phenomenon that has still extensive attention of the macroeconomists, policymakers and the central bankers from both developing and developed world. It is also a main subject of the macroeconomics and one of the principal concerns of the policymakers and the public (Hussain and Malik, 2011). Also, one of the main views among economists and the government in Indonesia is the low and stable inflation. Economic theory assures us low and stable inflation is important for market-driven growth, and that monetary policy is the most direct tool for controlling inflation. Further, among all of the government's tools for influencing and directing the economy, monetary policy has proven to be the most flexible instrument for achieving medium-term stabilization objectives.

One of the ways to resolve inflation uses statistical tools, such as forecasting the future it. This is ever revealed by Lusia and Suhartono (2012) who had done that the current inflation based on the past inflation although it was just nebulous information. In addition, forecasting inflation is a key policy maker. In other side, Faizah and Setiawan (2013) proposed that a city in the process of filling the needs depends on another city especially at the city adjacent to each other. Generally, a city is very dependent of another city because not all the necessary requirements in the city would be met by itself but it could have been obtained from another city. Thus, the inflation movements have relevance at an earlier time, also they have been linked to other cities called by spatial relationships.

Spatial-temporal models arise when data are collected across time as well as space. An observation at a certain spatial location and time may be commonly influenced by previous-time observations in that location and its neighboring locations. In this paper we proffer a new procedure for space-time modeling, especially for generalized STAR (GSTAR) modeling. GSTAR was proposed by Borovkova, Lopuha and Ruchjana (2002). The GSTAR is one of space-time models characterized by autoregressive terms lagged in the p-th order in time and the order of $(p; \lambda_1, \lambda_{2,\dots,} \lambda_p)$ in space.

The purpose of this paper is to apply the GSTAR model for forecasting the inflation rate data in Sulawesi. This way has been generally shown by Wahyuningrum and Suhartono (2014). They compared two models between VARIMA and GSTAR model for forecasting inflation rate in East Java.

2. VARIMA MODEL

We introduce a more general class of vector time series models to describe relationship among several time series variables. A useful class of parsimonious models is the vector autoregressive moving average ARMA (p,q) process

$$\Phi_p(B)Y(t) = \Theta_q(B)e_t$$

where

$$\Phi_p(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$$

and

$$\Theta_q(B) = 1 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_p B^p$$

are the autoregressive and moving average matrix polynomials of order s p and q (Wei, 2006).

3. GSTAR MODEL

GSTAR model is a generalization of the space-time autoregressive (STAR) model which is also the specification of the vector autoregressive (VAR) model. The fundamental difference between GSTAR and STAR model rests on the parameter assumption. STAR model assumes that the locations used in the study are the same, so this model can only be applied at a uniform location. The GSTAR model assumes the heterogeneous locations signed in weighted matrices (Nurhayati, Pasaribu and Neswan, 2012).

If Let $\{Y(t): t = 0, \pm 1, \pm 2, \cdots, T\}$ is a multivariate time series and be an N-dimensional vector process, the p-order GSTAR model with spatial order $\lambda_1, \lambda_1, \cdots, \lambda_p$ can be written as follow

$$Y(t) = \sum_{s=1}^{p}\left(\Phi_{s0} + \sum_{k=1}^{\lambda_s}\Phi_{sk} W^{(k)}\right) Z(t-s) + e(t)$$

Where W=weighted matrix are chosen such that $w_{ii} = 0$ and $\sum_{i \neq j} w_{ij} = 1$. p is the autoregressive order, λ_s is the spatial order of the sth autoregressive term, $\Phi_{s0} = \mathrm{diag}\,(\phi_{s0}^1, \ldots, \phi_{s0}^N)$, $\Phi_{sk} = \mathrm{diag}\,(\phi_{sk}^1, \ldots, \phi_{sk}^N)$, and finally $\mathbf{e}(t)$ is an error vector at time t which is assumed to be i.i.d. (independent and identically distributed) with zero mean and constant variance. The following example is a GSTAR model using by differently three locations and one order as

$$Y(t) = \Phi_{10} Y(t-1) + \Phi_{11} W^{(1)} Y(t-1) + e(t)$$

The model can be exspressed as matrix notation

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} = \begin{pmatrix} \phi_{10} & 0 & 0 \\ 0 & \phi_{20} & 0 \\ 0 & 0 & \phi_{30} \end{pmatrix} \begin{pmatrix} y_1(t-1) \\ y_2(t-1) \\ y_3(t-1) \end{pmatrix} + \begin{pmatrix} \phi_{11} & 0 & 0 \\ 0 & \phi_{21} & 0 \\ 0 & 0 & \phi_{31} \end{pmatrix} \begin{pmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{pmatrix} \begin{pmatrix} y_1(t-1) \\ y_2(t-1) \\ y_3(t-1) \end{pmatrix} + \begin{pmatrix} e_1(t) \\ e_2(t) \\ e_3(t) \end{pmatrix}$$

4. APPLICATION OF GSTAR MODEL TO THE INFLATION RATE IN SULAWESI

In this section, we apply the GSTAR model to the inflation rate data in seven major cities in Sulawesi. The seventh major city in Sulawesi is Manado, Palu, Makassar, Palopo, Kendari, Gorontalo, and Mamuju.

4.1. *Dataset*

We have formed series of monthly data from Badan Pusat Statistik, South Sulawesi for periods 2008-2015. It consists of 96 observations of 7 dimensional vectors. For the purpose of forecasting the data was grouped into the training data set and test data set. The training data is the first 84 observations that will be used for model building and the test data is the last twelve data that will be used in forecasting performance comparison.

Table 1. The Monthly Average of City Major Inflation Rate in Sulawesi

Month	Manado	Palu	Makassar	Palopo	Kendari	Gorontalo	Mamuju
Jan	-0,037	0,610	0,910	0,457	0,650	0,359	0,409
Feb	0,489	0,061	0,237	0,304	0,039	0,581	0,093
Mar	0,671	0,220	0,091	0,334	0,299	0,846	-0,041
Apr	-0,200	-0,712	0,003	0,190	0,466	0,957	-0,267
May	0,067	0,446	-0,056	0,003	0,774	1,149	0,313
Jun	0,107	0,906	0,409	0,391	0,801	1,517	0,691
Jul	1,333	1,973	**1,319**	**1,260**	**2,091**	2,714	**1,221**
Aug	0,801	1,695	1,044	0,871	1,744	3,979	1,117
Sep	-0,499	0,207	0,190	0,009	0,520	3,331	0,414
Oct	0,399	1,160	-0,164	-0,241	0,256	3,456	0,507
Nov	0,644	1,013	0,216	0,313	1,039	4,099	0,943
Dec	**1,597**	**3,043**	1,039	0,767	1,289	**5,350**	0,850

Descriptive analysis of the inflation data of nine cities in Sulawesi by month can be seen in Table 1. Based on Table 1 it can be explained that the average rate of inflation was the highest inflation in July as the city of Makassar (1,319), Palopo (1,260), Kendari (2,091) and Mamuju (1,221). The other cities have relatively the highest average inflation in December (Manado 1,597; 3,043 Palu; and Gorontalo 5,350).

4.2. *Preliminary VARIMA Model Identification*

Preliminary model identification to choose the VARIMA parameters (p,d,q) is performed using MACF, MPACF and Akaike's Information Criterion (AIC) value. The model building of VARIMA is processed by supporting SAS package program. The steps of the model estimation is done through the following procedures: the identification by using MACF, MPACF, and AIC value, parameter estimation, check diagnostic to verify that the residual of the model satisfies the white noise requirement.

Schematic Representation of Cross Correlations

Variable/ Lag	0	1	2	3	4	5	6	7	8	9
y1	+++++++	...+++.	-......	...----	....	....	....	....	....	....
y2	+++++++	.++...+	.+.....	.+.....	.+.-...	.+....+	.+....+	......+	......+	......+
y3	+++++++	.+..++.	...---.	...--..	.+.....	.+.....	.+...+.	....	....	...--..
y4	+++++++	.+++++.	...---.	...---.	.+.....	.+.+...	.+.+++.	....	....	...--..
y5	+++++++	.+++++.	-......	...---.	.+.....	.+.....	...+++.	....	....	-..--..
y6	+++++++	.+++++.	-......	.+.---.	.+.....	++++++.	.+.+++.	....	....	...-...
y7	+++++++	.++..++	.+....+	.+..-..	.+.....	.+.....	.+.....	....	....	....

Figure 1. Sample Cross Correlation Matrices In Sulawesi

The 84 observations in the training data, depicted in figure 1, should be considered non-stationary. These data can be shown by summarizing these matrix elements using indicator symbols (+), (-), and (.) where (+) denotes value greater than twice the estimated standard error, (-) denotes a value less than twice the estimated standard error and (.) denotes an insignificant value. From SCCM, (+) and (-) are more than (.). After differencing at lag 1, SCCM is illustrated in figure 2.

Schematic Representation of Cross Correlations

Variable/ Lag	0	1	2	3	4	5	6	7	8	9
y1	+++++++	-......	-......	....--.	....	.+.....	....	....	....	....
y2	++.++++	-......	....	.....-.	----...	+......	....	....	....	....
y3	+.+++++	..-....	....---	....	....	....	.+.....	.-.....	....	....
y4	+++++++	....	-.-----	....-..	....	....	...+.+.	....	....	....
y5	+++++++	....	-.---.-	....--.	....	....	.....+.	....	....	....
y6	+++++++	....	-------	....-..	....	...+...	....	....	....	....
y7	+++++++	....	..-.---	....	....	....	.....+.	....	....	....

Figure 2. Sample Cross Correlation Matrices after Differencing In Sulawesi

Figure 2 clearly indicates that inflation rate data have already qualified stationary. These patterns indicate cut off after lag 4. Additionally, we have to consider AIC value to determine VARIMA order supporting information in Table 2. Using AIC as the selection criterion, we select the VARIMA model with the lowest value of the AIC. From the table, we get the minimum AIC value as AR (2) order. So, initial identification of VARIMA order for the best forecasting inflation rate in Sulawesi suggest VARIMA (4,10).

Table 2. AIC Values of Dataset of Inflation Rate for Different Models

Lag	MA (0)	MA (1)
AR (0)	-1,972	-2,385
AR (1)	-2,961	-2,283
AR (2)	**-3,405**	-2,350
AR (3)	-3,205	-1,854
AR (4)	-2,373	-0,929
AR (5)	-1,148	0,884

A summary of the parameter estimates and their standard errors for the fitted VARIMA (4,10) model after restriction using uniform weighted process is displayed in Table 3. The estimated parameter coefficient of all location is statistically significant at the 5% level. VARIMA model equation for inflation Manado city shows that inflation in these locations are affected by inflation at the same location at 1 month, 2 months, 3 months, and 4 months earlier. In addition, the city of Manado also affected by inflation in Kendari at 1 month and 2 months earlier respectively. VARIMA model for inflation in Palu is as follow

$$y_{1,t} = y_{1,t-1} - 0,712 y_{1,t-1} + 0,712 y_{1,t-2} + 0,498 y_{5,t-1} - 0,498 y_{5,t-2} - 0,4 y_{1,t-2} + 0,4 y_{1,t-3} - 0,201 y_{1,t-3} + 0,201 y_{1,t-4} + e_{1,t}$$

Tabel 3. Estimated Parameters for VARIMA (4,10) for Dataset of Inflation Rate

Location	Parameter	Estimate	SE	Location	Parameter	Estimate	SE
Manado	ϕ_{11}^{1}	-0,712	0,083	Kendari	ϕ_{57}^{3}	-0,113	0,032
	ϕ_{15}^{1}	0,498	0,105				
	ϕ_{11}^{2}	-0,400	0,066				
	ϕ_{11}^{3}	-0,201	0,064				
Palu	ϕ_{22}^{1}	-0,243	0,088	Mamuju	ϕ_{73}^{1}	-0,304	0,089
	ϕ_{23}^{1}	0,345	0,150				
Makassar	ϕ_{33}^{1}	-0,491	0,056	Gorontalo			
	ϕ_{33}^{2}	-0,216	0,059				
	ϕ_{33}^{3}	-0,185	0,060		ϕ_{66}^{1}	-0,278	0,053
	ϕ_{32}^{4}	-0,132	0,026		ϕ_{65}^{2}	0,215	0,060
Palopo	ϕ_{43}^{1}	-0,241	0,032		ϕ_{66}^{2}	-0,486	0,071
	ϕ_{43}^{2}	-0,137	0,035		ϕ_{66}^{3}	-0,239	0,052
	ϕ_{43}^{3}	-0,183	0,037		ϕ_{66}^{4}	-0,194	0,048
	ϕ_{42}^{4}	-0,037	0,014				
	ϕ_{43}^{4}	-0,138	0,028				

VARIMA model equation for inflation in Palopo is as follows

$$y_{4,t} = y_{4,t-1} - 0{,}241 y_{3,t-1} + 0{,}241 y_{3,t-2} - 0{,}137 y_{3,t-2} + 0{,}137 y_{3,t-3} - 0{,}183 y_{3,t-3} + 0{,}183 y_{3,t-4} - 0{,}037 y_{2,t-4} + 0{,}037 y_{2,t-5} - 0{,}138 y_{3,t-4} + 0{,}138 y_{3,t-5} + e_{4,t}$$

VARIMA model in Palopo inflation rate shows that inflation in these locations are affected by inflation at the same location at 1 month earlier. Moreover, this location is also influenced by inflation in Makassar at 1 month, 2 months, 3, 4 and 5 months earlier and inflation in Palu at 4 months and 5 months earlier.

4.3. *GSTAR Model Building for Inflation Rate*

To identify of the GSTAR model order, we can use same method from VARIMA models. MPACF and AIC are presented by determining GSTAR order. In practice, of course, GSTAR ($[4,1,0]_1$) can be applied modeling inflation rate in Sulawesi directly and used one spatial order. The uniform space weight are generally taken as follow

$$W_{ij} = \begin{bmatrix} 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \end{bmatrix}$$

For parameter estimation, GSTAR model is done by regressing variables $y_{i,t}$ where it is the result of the differencing dataset at lag 1, with 56 variables as prediction. By using a significance level of 0,05, the results of parameter estimation use uniform space weight significantly can be seen in Table 4.

Table 4. Parameter Estimation for Model GSTAR ($[4,1,0]_1$) using Uniform Space Weight

Locations	Parameter	Estimate	*Standard Error*	*t-value*	*p-value*	Variable
Manado $(y_{1,t})$	ϕ_{10}^{2}	-0,24	0,13	-1,81	0,04	$y_{1,t-3}$
Palu $(y_{2,t})$	ϕ_{20}^{1}	0,425	0,13	3,35	0,0012	$y_{2,t-1}$
	ϕ_{20}^{2}	0,29	0,14	2,05	0,04	$y_{2,t-3}$
	ϕ_{21}^{4}	0,76	0,35	2,18	0,03	$v_{2,t-4}$
Makassar $(y_{3,t})$	ϕ_{30}^{1}	0,392	0,15	2,61	0,01	$y_{3,t-1}$
	ϕ_{30}^{2}	-0,35	0,16	-2,24	0,03	$y_{3,t-3}$
Palopo $(y_{4,t})$	ϕ_{41}^{1}	0,23	0,13	1,82	0,04	$y_{4,t-2}$
Kendari $(y_{5,t})$	ϕ_{50}^{1}	0,697	0,125	5,56	0,0001	$y_{5,t-1}$
	ϕ_{51}^{4}	0,35	0,17	2,06	0,04	$v_{5,t-4}$
Gorontalo $(y_{6,t})$	ϕ_{60}^{1}	0,765	0,116	6,6	0,0001	$y_{6,t-1}$
	ϕ_{61}^{4}	0,52	0,23	2,22	0,03	$v_{6,t-4}$
Mamuju $(y_{7,t})$	ϕ_{70}^{1}	0,892	0,106	8,39	0,0001	$y_{7,t-1}$
	ϕ_{70}^{3}	0,29	0,14	2,04	0,04	$v_{7,t-1}$

a. GSTAR ($[4,1,0]_1$) model in Manado

$$y_{1,t} = y_{1,t-1} + 0{,}24 y_{1,t-3} - 0{,}24 y_{1,t-4}$$

b. GSTAR ($[4,1,0]_1$) model in Palu

$$y_{2,t} = y_{2,t-1} + 0{,}425 y_{2,t-1} - 0{,}425 y_{1,t-2} + 0{,}29 y_{2,t-3} - 0{,}29 y_{2,t-4} + 0{,}127 y_{1,t-4} - 0{,}127 y_{1,t-5} + 0{,}29 y_{2,t-4} - 0{,}29 y_{2,t-5} + 0{,}127 y_{3,t-4} - 0{,}127 y_{3,t-5} + 0{,}127 y_{4,t-4} - 0{,}127 y_{4,t-5} + 0{,}127 y_{5,t-4} - 0{,}127 y_{5,t-5} + 0{,}127 y_{6,t-4} - 0{,}127 y_{6,t-5} + 0{,}127 y_{7,t-4} - 0{,}127 y_{7,t-5}$$

c. GSTAR ($[4,1,0]_1$) model in Makassar

$$y_{3,t} = y_{3,t-1} + 0{,}392 y_{3,t-1} - 0{,}392 y_{3,t-2} - 0{,}35 y_{3,t-3} + 0{,}35 y_{3,t-4}$$

d. GSTAR ($[4,1,0]_1$) model in Gorontalo

$$y_{6,t} = y_{6,t-1} + 0{,}765 y_{6,t-1} - 0{,}765 y_{6,t-2} + 0{,}52 y_{6,t-4} - 0{,}52 y_{6,t-5} + 0{,}087 y_{1,t-4} - 0{,}087 y_{1,t-5} + 0{,}087 y_{2,t-4} - 0{,}087 y_{2,t-5} + 0{,}087 y_{3,t-4} - 0{,}087 y_{3,t-5} + 0{,}087 y_{4,t-4} - 0{,}087 y_{4,t-5} + 0{,}087 y_{5,t-4} - 0{,}087 y_{5,t-5} + 0{,}087 y_{7,t-4} - 0{,}087 y_{7,t-5}$$

e. GSTAR ($[4,1,0]_1$) model in Mamuju

$$y_{7,t} = y_{7,t-1} + 0{,}892 y_{7,t-1} - 0{,}892 y_{7,t-2} + 0{,}29 y_{7,t-1} - 0{,}29 y_{7,t-2} + 0{,}048 y_{1,t-1} - 0{,}048 y_{1,t-2} + 0{,}048 y_{2,t-1} - 0{,}048 y_{2,t-2} + 0{,}048 y_{3,t-1} - 0{,}048 y_{3,t-2} + 0{,}048 y_{4,t-1} - 0{,}048 y_{4,t-2} + 0{,}048 y_{5,t-1} - 0{,}048 y_{5,t-2} + 0{,}048 y_{6,t-1} - 0{,}048 y_{6,t-2}$$

Based on these models created by GSTAR models, we can be explored that inflation rate in Manado is only affected by the inflation rate one, three and four months earlier and is not affected by the inflation rate of the other cities. Inflation in Palu is respectively influenced by itself one, two, three, four and five months earlier. Besides, this inflation location is also influenced by the inflation rate of the other location such as Manado, Makassar, Palopo, Kendari, Gorontalo and Mamuju city respectively four and five months earlier.

For inflation in Makassar, it is influenced by the inflation rate of itself is one, two, three and four months earlier and is not influenced by the inflation rate of the other location. Almost equal to Makassar, the inflation rate in Palopo also only influenced by itself one, two and three months earlier. Kendari has a pattern similar to Palu inflation value depends on the other locations four and five months earlier and also depends on itself one, two, four and five months earlier. This pattern also is applied in Gorontalo and Mamuju inflation rate.

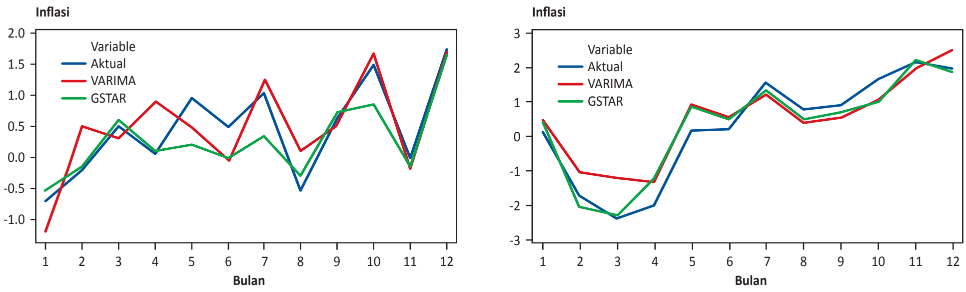


Figure 3. Forecasting Performance for Inflation Rate (a) Manado (b) Palu

Forecasting performance for the dataset taken by two locations as sample is displayed in figure 3. Based on analysis, the forecasting by using GSTAR ($[4,1,0]_1$) model is better than the VARIMA (4,10) models. We compared both using by mean square error (MSE) where GSTAR ($[4,1,0]_1$) and VARIMA (4,10) models have MSE 0,291 and 0,358.

5. CONCLUSIONS

We have shown that time series VARIMA and GSTAR models can be used to model and forecast the inflation rate in Sulawesi. Based on the result, we proposed VARIMA (4,1,0) and GSTAR ($[4,1,0]_1$) models for forecasting dataset of inflation rate in Sulawesi. The forecast performance is evaluated using the mean squared errors (MSE) based on the last twelve actual data. The result showed the MSE of GSTAR ($[4,1,0]_1$) is smaller than VARIMA (4,1,0).

REFERENCES

Borovkova, S.A.; Lopuha, H.P, & Ruchjana, B.N. (2002). The Space Time Autoregressive Models. Workshop on Space Time Models and Its Applications. Bandung, 2-4 Agustus 2005.

Faizah & Setiawan, 2013. Pemodelan Inflasi di Kota Semarang, Yogyakarta, dan Surakarta dengan pendekatan GSTAR, Jurnal Sains Dan Seni Pomits Vol, 2, No,2, (2013) 2337-3520, Surabaya

Hussain, S. & Malik, S. 2011. Inflation and Economic Growth: Evidence from Pakistan. International Journal of Economics and Finance. Vol. 3, No. 5; October 2011.

Lusia & Suhartono, 2012,Peramalan Inflasi dengan Metode Weighted Fuzzy Time Series, Jurnal Statistika ITS.

Nurhayati, N., Pasaribu, U. S., & Neswan, O., 2012. Application of Generalized Space-Time Autoregressive Model on GDP Data in West European Countries. Journal of Probability and Statistics Volume 2012, Article ID 867056.

Wahyuningrum, S. R. & Suhartono. (2014). GSTARX-GLS Model For Spatio-Temporal Data Forecasting. Disertation. Surabaya: ITS.

Wei, W.W.S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. United State of America : Addison-Wesley Publishing Co., USA.

Food Price Information System Application: Survey Based Data Towards Regional Inflation Management Protocol

Noor Yudanto*
Bank Indonesia, Jakarta, Indonesia – noor_y@bi.go.id

Handri Adiwilaga
Bank Indonesia, Jakarta, Indonesia – adiwilaga@bi.go.id

Maxmillian T. Tutuarima
Bank Indonesia, Jakarta, Indonesia – maxmilian@bi.go.id

Abstract

High food price volatility and disparity among regions have been observed as persistent source of inflation in Indonesia. As an archipelago country, inflation in Indonesia in many cases caused by the weakness of supply side such as under potential food production, inefficient food distribution, above normal food prices, and public assymetric information on prices that sparks uncontrolled price expectation. Bank Indonesia (BI), the central bank, by Act has mandate to achieve and manage price stability, by optimizing monetary instruments to acheive the single objective. However, the complexity of issues on controlling inflation, both from demand and supply side, urged BI to maintain a high intense policy coordination with the government, both in central and regional level. Since 2008, this role of policy coordination has been institutionalized in the form of a flexible regional inflation controlling team, or TPID. Currently there are 507 of 542 autonomy regions spreadout the country have set up TPID. TPID policy recommendations to the regional government on preventing potential goods and services price pressures have been praised for its contribution to lower the level of regional inflation.

On the context to monitor and strengthening policy on controlling regional inflation, Bank Indonesia tooks step forward by establishing Food Price Information System Application in mid of 2016. The application, that has already available on website and mobile application, has been sources of reliable daily food price data. The high frequency data was built based on Bank Indonesia high standardized food price daily survey methodology on 82 CPI cities throughout the country. Gradually, the cities surveyed will gradually be expanded to all TPID cities. By the end of 2018, this application would expected to features on data of food stocks on producers level, virtual chat for decision makers, and e-commerce for more transparant, direct, and efficient distribution. As a 'handy; application with full comprehensive data and information relates to regional inflation issues, the application is hopefully provide valuable inputs for BI, Goverment and other relevant institutions for better and stronger policy coordination on controlling regional inflation.

Keywords: Inflation; Food price; Central bank; Policy coordination

1. INTRODUCTION

Indonesia, in term of GDP, currently ranked at 16 among the globe (World Bank, 2016). As the largest archipelago country in the world, but still constrained by under develop connectivity infrastructures, to achieve and manage low and stable inflation rate is a big challenge. Historically, the country highest peak of inflation rate was 635%, back in 1966 due to the economic alongside with politics turmoil. For the last ten years, the annual inflation rate has been successfully kept on 4-5% averagely. Volatility on food price has been almost persistent source of the inflation in the country that closely mainly caused by supply side factors such as production capacity gap, inefficient logistics and distribution system, seasonal factors, and assymetric information. On CPI basket, the weight of food component is currently around 30%, almost similar to peers countries. However, Indonesia inflation rate remains the highest among the peers. Government policies imposed on price of fuel, public transport, electricity, and taxes on cigarette were also noted as significant sources of the rise of general prices. As general practices, the central bank (Bank Indonesia) has a mandate to achieve and maintain the stability of price by optimizing its monetary policy. Bank Indonesia currently implemented light inflation targeting framework to carry out the mandate. Under this framework, Bank Indonesia is optimizing its policy instruments to achieve the inflation target[1]. However, since the inflation often triggered by supply side factors, mainly related to food, and government policy related to tariffs, therefore controlling inflation requires close coordination with the government. In this context, Bank Indonesia and the government has set up a team known as Inflation Controlling Team (TPI) in national level and Regional Inflation Controlling Team (or known as TPID[2]) in regional level.

The main purpose of TPI and TPID is to set a coordinated policy and joint program in order to address supply side issues in inflation. The TPIDs innovated programs on controlling inflation in regions had been claimed for its contribution to bring volatile food inflation rate on the 'new' pattern. For comparison, the average of volatile food inflation on 2011-2015 was at 8-9%, or lower than previous period (2006-2010) was at 12-13% annually. Furthermore, TPID begin to start aiming to address more structural issues in inflation. One of the structural issues is assymetric information problem which hinder widely public access to a credible food price data. Aiming at the purpose, the Task Force started to launch the National Food Price Information System (PIHPS[3]) program in 2012. A need for a high accurate food price data lead Bank Indonesia, as a member of the Task Force, to initiate strengthen PIHPS by using Bank Indonesia daily food price survey. Therefore PIHPS can provide near-real-time access to food price data which can be accessed through website or using mobile application. PIHPS has also supported features of early warning system and virtual meeting purposed for coordination and immediate policy action. In broader perspectives, data on PIHPS provide valuable information for larger aspects of government policy, such as, national food security strategy, agriculture infrastructure development planning, connectivity infrastructure development planning, etc. The challenges on developing PIHPS lay on how to keep its continuity, accuracy the data, keep up with user need, and resources.

2. INFLATION IN INDONESIA: CHARACTERISTICS, DETERMINANTS, AND POLICY STRATEGY

Inflation Characteristic

Inflation in Indonesia is measured using Consumer Price Index (CPI), which is calculated using weight of a basket of commodity goods and services consumed by the population in 82 cities[4]. In general the pressures of Indonesia's prices were contributed by shocks, especially shocks in food prices and the government's strategic policies affected the administered prices. Indonesia has been experience hyperinflation during 1966 up to 635%, due to the economic alongside with politics turmoil. In the period of 2000-2016, spike on inflation movement occurs at the time of a supply disruption of food. Similarly, in the event of changes in energy-related price set by the government (Chart 1). On the short term basis, shocks on the demand side also gave impact on the volatility of prices, especially on the holiday events (Ramadhan, Christmas). Moreover, another factor that needed to be concerned was the backward and forward inflation expectations, which inclined to be dominated by backward expectations[5].

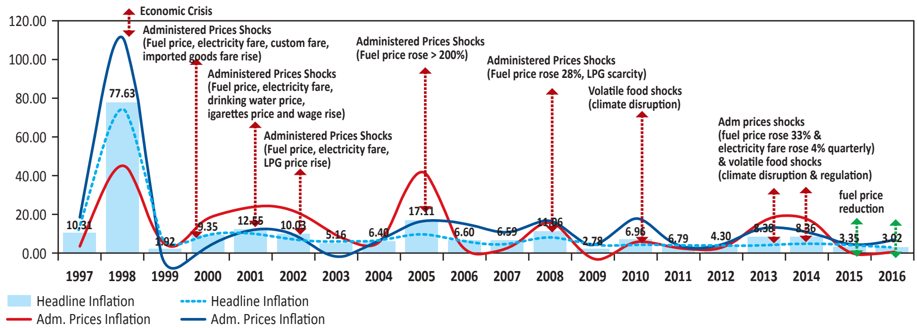
1 Inflation target is set by the government after coordination with the Bank Indonesia.

2 TPID: Tim Pengendalian Inflasi Daerah (in Bahasa).

3 PIHPS: Pusat Informasi Harga Pangan Strategis, (in Bahasa).

4 Based on the Cost of Living Index (SBH) 2012, the CPI is measured based on classifications known as disaggregation of inflation, including volatile foods commodities (weighs16,62% of the CPI), administered prices commodities (weighs 18,02% of the CPI), and core inflation commodities (weighs 65,36% of the CPI).

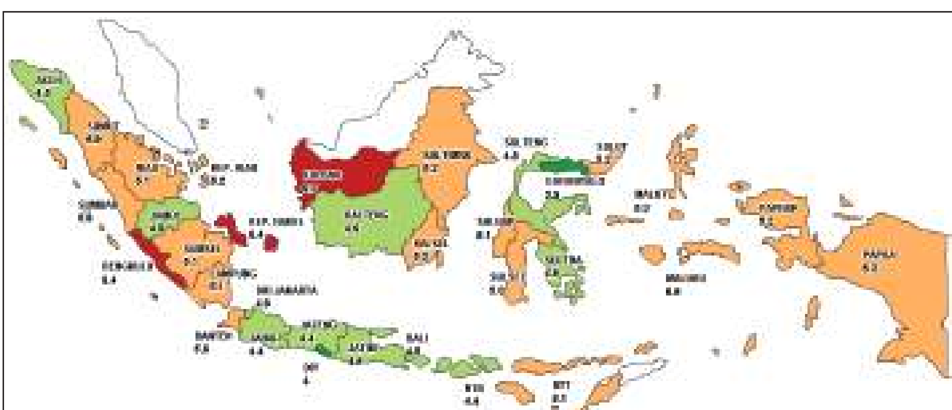
5 Wimanda, Rizki E., Turner, Paul M., & Hall, Maximilian J.B. (2011), 'Expectations and The Inertia Of Inflation: The Case of Indonesia' Journal of Policy Modelling No. 22 pp. 426-438



Source: BPS and Bank Indonesia

Chart 1. National Inflation Rate

Spatially, inflation in many regions outside Java tends to be higher than in Java and even more fluctuating due to the the high transportation (distribution) and logistical costs of goods. Much lower inflation pressures in Java mainly part because of role of Java as the center of distribution of goods with large scale retail chain. On the other hand, inflation rates in the Sumatera and Eastern Indonesian regions are relatively higher. Higher inflation rate in those two regions is mostly caused by the factor of price volatility of a few food commodities that are imported from Java and other factors such as differences in sectoral specialization, asynchronous business cycles, degree of openness, or nominal wage rigidities, and local specific factors[6]. This non-monetary factor is driven by the cost-push of goods and services due to the availability, adequacy and affordability of supplies. This supply-side inflation has been an issue in controlling regional inflation in Indonesia in recent years. Inflation disparity also remains a challenge in Indonesia, which requires policymakers to factor in the heterogeneities among regions and also the implication of their policies that would have an impact on regional inflation.



Source: BPS and Bank Indonesia

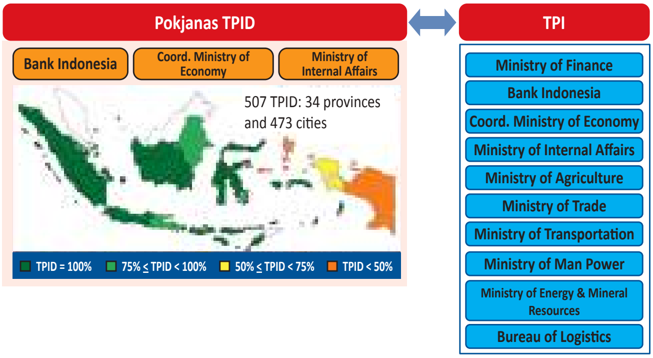
Diagram 1. Regional Inflation Rate (average 2014-2016)

6 M. Ridwan, (2015), Inflation Differentials, Determinants, and Convergence: Evidence from Indonesia Subnational Data, Proceedings of the Asia Pacific Conference on Business and Social Sciences 2015, Kuala Lumpur.

Policy Strategy: Managing Inflation in Indonesia

Bank Indonesia (the central Bank of Indonesia) adopting Inflation Targeting Framework (ITF) in July 2005 to achieve inflation target. The framework employed Bank Indonesia (BI)-policy interest rate as reference rate, and introduced anticipated monetary policy formulation mechanism, transparant communication strategy, and empowered the coordination with government. The volatility of core inflation in Indonesia has experienced a decrease from time to time since the implementation of Inflation Targeting Framework. The average year on year core inflation tends to be lower that the CPI inflation rate. On the other hand, the volatile food inflation tends to fluctuate frequently and has always been higher than the CPI inflation rate. Meanwhile, the administered price inflation rate has always been less volatile that the volatile food inflation, although it is always higher than the CPI inflation rate in general. This dynamics indicate that relying only on interest rate to maintain inflation is not sufficient, considering the inflation characteristic often triggered by non-monetary factors, such as food price volatility and administered prices policy.

Efforts to lower inflation effectively required a strong coordination between the Government, as the authorities for fiscal and sectorial policies, and Bank Indonesia, as the regulator for monetary policies. Under the Bank Indonesia Law, the Government should establish the inflation target by coordinating with Bank Indonesia. As a follow-up to the Law, the government and BI established a coordination mechanism by forming the Inflation Targeting, Monitoring and Controlling Team (TPI) in a Decree of Minister of Finance 2005[7]. The team is not only enacted in concern of establishing the national inflation target, but also to monitor and evaluate the policy undertaken by the Government and Bank Indonesia in achieving the target[8]. The coordination mechanism was also expanded in 2008 to the regional level by the set up of the Regional Inflation Controlling Team (TPID)[9]. The main purpose of TPID is to set a coordinated policy in order to address supply side issues in inflation and also initiate a joint program or activities to mitigate inflation pressure. The establishment of TPID was based on the facts that controlling inflation in the regional level would have significant impact in achieving the low and stable national inflation rate[10]. By the end of 2006, there are 507 regions has established TPID out of 542 regions in Indonesia. To optimize the effectiveness of inflation controlling effort nationwide, the national and regional policy need to be synchronize, therefore in national level Bank Indonesia, with the government also set up a TPID National Working Group (known as Pokjanas TPID, in Bahasa).



Source: Pokjanas TPID report

Chart 2. Coordination Mechanism

7 GBI No.88/KMK.02/2005 dan No.7/9/KEP.GBI/2005.

8 Membership structure of TPI is available on appendix.

9 The establishment of TPID was governed by the Instruction of Minister of Internal Affairs which required all regional levels (provinces and cities) to form TPID in a Decree of Regional Head (Governor for Provinces, Mayor for Cities).

10 Regional inflation rate contribute 80,77% to national inflation rate (exclude Jakarta).

Assymetric Information and Food Price Volatility in Indonesia

Economists have long emphasized that information is critical for the efficient functioning of markets. When information is limited or costly, agents are unable to engage in optimal arbitrage. Excess price dispersion across markets can arise, and goods may not be allocated efficiently. According to Jensen (2007), information became highly importance for the functioning of markets because information makes markets work. Another insight from Labonne and Chase (2009) show that asymmetric information provide traders better market information, that will be their advantage when negotiating with farmers over the price of what they produce. Providing price information to farmers could be beneficial by increasing their bargaining power when they negotiate with traders. Furthermore, better access information allows farmers to strike better price deals within their existing trading relationships and to make better choices in their markets.

Regarding the issue, it is important to establish a food price information system which provides accessible, accurate and reliable price information as a reference for market players. Furthermore, providing a transparent and reliable information can support the decision-making mechanism by the governments, both in national and regional levels. In this context, monitoring food price movements becomes highly important for Indonesia which can help to better anticipate any potential of abnormal changes food price in the country across regional. Given the geographic characteristic of Indonesia, where market players are spread in its archipelago, an efficient approach in monitoring food price movement is needed, but should be based on real market price. At this point, it is important to have information that are commonly used to explain price dynamics (trend, volatility and spike).

3. STRATEGIC FOOD PRICE INFORMATION SYSTEM

Attempts to address asymmetric information has become part in an effort to control inflation in Indonesia. The characteristic of Indonesia's inflation which frequently caused by the volatility in food prices indicates that the availability of a reliable and timely data is essential part for maintaining price stability. From both consumer and producer perspective (and farmers), the availability of food price data can be used as a reference for transaction. Furthermore, the rapid development of internet and information technology have evolved into powerful means of rapid information exchange, enabling public to easily access information. Exploiting these technologies to gather and disseminate real time food price data provides enormous potential for policy decisions, managing expectation, and thus achieving inflation target.

One of the effort to address asymetric information problem is the initiative by the central bank in collaboration with the government to develop information system to provide a reliable and timely food prices data accessible to public. The initiative aims to provide public with access to a near-real-time food prices data based on a credible survey, employing the technology and internet. This initiative is referred to as Strategic Food Prices Information System (in Bahasa is Pusat Informasi Harga Pangan Strategis - PIHPS). The system is design to address some drawbacks in the existing food price data available, such as time delay between survey and data publication, validation process of the data collected, dissemination methods, and ease of use of the data by public. Furthermore, the system is design for tracking inflation and as as a complement to the price index data published by the statistical agencies.

Overview of PIHPS Design/Framework

The development of tear-real time food price data monitoring through the PIHPS are designed to play an important role in the area as of follows: (i) Ensure the availability of food price information for all economic agents, both producers (farmers), traders and consumers; (ii) Optimizing data or information of strategic food commodity prices that have been developed by the relevant institutions at central and local levels; (iii) Facilitating the role of regions in the supply, processing and dissemination of information of strategic food prices; (iv) Improving the effectiveness of information dissemination and communication of food prices at national and regional level; (v) Strenghten the economic analysis mainly tracking of inflation forecast across regions; (vi) Providing early warning indicators for the policy makers at the national and local level, in particular through national and regional controlling inflation team (TPI and TPID).

The Survey

On the first stage of PIHPS development mainly focus on food price data collection from markets and development of dissemination media. This is in line with the principle of ensuring the availability of information in food prices for all economic agents as well as increasing the utilization of existing data. The strength of PIHPS lays on the the Bank Indonesia daily food price survey conducted on 82 cities/municipals that are represent all provinces in Indonesia and included in the calculation of national inflation by BPS-Statistics Indonesia. The survey method has been mimicking the CPI survey of statistical agency completely in order to keep the standardized and quality of data resulted. The main different between BI and statistical agency survey was on its frequency. This survey is also considering that data from PIHPS could be used as an analysis tool for tracking inflation and as early warning indicators for policy-makers.

The Survey conducted on retailers in main market – traditional market and modern market - in each cities/municipals. Data and information on PIHPS consists of minimal 10 major food commodities for each cities/municipal that are considered to have considerable influence in inflation and available across regions[11]. However, selection of type quality for each commodity depend on the (consumer) consumption characteristics in each region. The mechanism of survey is designed so that there is minimum lag from data collection, data validation, data uploading to the system, and data dissemination. There are two types of media dissemination developed in PIHPS that public easily accessed, that is web based and mobile app (Android and iOS basis). Both of media dissemination are designed to be able to display the data up to the level of the market with the name of the market, both traditional markets and modern markets. The level of ease to access data through web and mobile app is take into highly consideration on the development.

On the next stage of development, PIHPS data will include wholesaler price, producer price, stock data, and public engagement modul. Some of them will be utilizing data or information that have been developed by the relevant institutions at central and local levels e.g. National Logistic Agency (Bulog), statistic agency, and commercial company data. Moving forward, the nex stage of PIHPS development is also to begin utilizing social media to gather information about current food price and the risks entail across regions.

Challenges

In the implementation of PIHPS, the challenge is mainly related to the scope of data and data processing. There is a still a limited number of commodity surveyed on PIHPS. This number of commodity has fairly shows a good indications to be used for tracking food inflation and thus for early warning indicators. However, from the public view, there is are need for PIHPS to cover more commodity for reference. In certain cases the supply of commodities is not consistently available in the market e.g. shocks of supply from production centers, shocks in distribution caused by natural disaster. Coverage of cities in PIHPS which is still limited to 82 cities lead to the limited public access to PIHPS. In terms of data processing, factor of quality internet network that has not been widely available accross regions hinder process of data input and that causing delay in some cases.

Future Development

PIHPS future development will be focused on the expansion of the city that will be included on the survey, the expansion of commodity coverage, and strengthening the early warning indicators. Expansion of the city on the PIHPS will gradually include all cities across the nation[12]. There is also need to expand the commodity coverage in PIHPS which inlcude more foodstuffs based on the characteristics of local consumption. PIHPS also have potential to not only collecting food data, but also other commodity that significantly affecting local inflation movement e.g. building materials. This is very important in

11 Rice, sugar, cooking oil, broiler chicken, beef, broiler chicken egg, red onions, red peppers, soybeans, and wheat flour.

12 There are total 509 cities and municipals in Indonesia in 2015.

order to strenghten its role as an early warning system. PIHPS future development potential is still wide open, especially to take advantage of crowdsourcing information. Utilization of crowdsourcing for data collection allows to expand data coverage in PIHPS. This is supported by improvements of the network quality across regions along with increasing mobile phones users. In addition, by optimizing the use of crowdsourcing will have a positive impact on cost required for data collection

4. CONCLUSIONS

This paper showed that on battling inflation rate in Indonesia, whose characteristically high sensitive of supply shocks factor, empowered institutional both in central and regional level for a stronger coordination among government agents proved as correct strategy to support the conventional monetary and fiscal policies. The works of the coordination forum, known as TPI and TPID, has successfully contribution to bring volatile food inflation down gradually. Administered price inflation, the inflation impact of Government policy imposed good price, also performed more manageable. Under the national controlling inflation team (TPI) coordination, the timing of policy imposed on good prices, e.g., electricity tariffs became more right as the calculation impact was considered the least. Under regional controlling inflation team (TPID) coordination, the shortage of rice, cooking oil, chicken eggs, and other basic food in regional market had been anticipated correctly. Overall, the success works of TPI and TPID on managing potential price pressure had been the background of the introduction of the Food Price Information System (PIHPS) on late 2016.

The paper also shared, how PIHPS, the system on data and information of prices combined with the coordination system, has been an accepted as new instrument to monitor the daily food price movement in easy and efficient way. Data reliability of PIHPS that resulted from Bank Indonesia survey was also the main reason for public trust to utilize the system. The feature of early warning system information considerably provided fundamental information for national or local government to set up policy response. While, the feature of virtual meeting accommodated as media of discussion and coordination to make quick policy response in the road. In the midst of its challenges, the high commitment of inflation task force to develop PIHPS for further steps showed how the importance of the system. The system is not only for controlling and managing food price inflation but also as provide basic information for the next larger aspect government policies such as national food security, spatial agricultural development strategy, national logistics and distribution system, and improving connectivity infrastructures.

REFERENCE

Wimanda, Rizki E., Turner, Paul M., & Hall, Maximilian J.B. (2011), 'Expectations and The Inertia Of Inflation : The Case of Indonesia' Journal of Policy Modelling No. 22 pp. 426-438

M. Ridwan, (2015), Inflation Differentials, Determeninants, and Convergence: Evidence from Indoonesia Subnational Data, Proceedings of the Asia Pacific Conference on Business and Social Sciences 2015, Kuala Lumpur

Nugroho Joko Prastowo, dkk (2008), Pengaruh Distribusi dalam Pembentukan Harga Komoditas dan Implikasinya terhadap Inflasi

Robert Jensen, (2007), The Digital Provide: Information (Technology), Market Performance, and Welfare in South India Fisheries Sector, The Quarterly Journal of Economics, Vol. CXXII

Julien Labonne, Robert S. Chase (2009), The Impact's of Mobile Phones on Farmer's Welfare in the Philippines, Policy Research Working Paper 4996

CPS04: Residential & Commercial Property Statistics

An Application of Hedonic Price Models in Property Markets: Consumer Willingness to Pay and Property Price Index Estimation

Masarina Flukeria*
Statistics Indonesia, Jakarta, Indonesia – masarina@bps.go.id

Dewi Agita Pradaningtyas
Statistics Indonesia, Jakarta, Indonesia – dewiagita@bps.go.id

Abstract

This paper considers the application of hedonic price models in property markets as an initial study to estimate the Property Price Index. The basic premise of this method is that the price of heterogeneous goods as property is related to its attributes or characteristics. The property price index is intended to monitor the development of house prices which are considered important for determinants or indicators for goods price inflation and economic activity. Unfortunately, it is not a simple task to construct the property price index. Due to the limited data availability, we utilize the data of Sales Value of Taxable Object (SVTO) by districts in Jakarta as a proxy of property prices and the period 2010 and 2014 as time-varying variable. Other data of property characteristics that are expected to affect property prices are also utilized, including area densities, accessibility characteristics that affect prices, such as distances to business centers, shopping centers, and availability of public transportation infrastructure, as well as environmental characteristics that affect prices such as risk level of flood. We employ regression techniques to estimate hedonic price function of the property that relates property price to the property characteristics. The hedonic price function is used to determine how much more money must be paid for a property with an extra unit of a particular property characteristic. The result of this study gives the implicit price of a property characteristic that is the marginal price of a characteristic indirectly revealed by the amounts of money that people are willing to pay for the whole property in order to obtain better quality or quantities of the characteristic. Moreover, based on the hedonic price model, property price index of Jakarta in 2014 with the base year of 2010 is estimated to yield higher increase than that of the Residential Property Price Index (RPPI) at the same period calculated by the Central Bank of Indonesia.

Keywords: Property price index; Hedonic price function; Implicit price.
JEL Classification: C43, C51, E30, E31, R31

1. INTRODUCTION

A house is often the largest and most important asset of households and therefore accounts for a major share of household wealth and can be used as collateral for loans. House price fluctuations may therefore have a major effect on economic activity and the soundness of the financial system. As a result, house price fluctuations may significantly amplify the effects of macroeconomic shocks, like supply, demand, or monetary policy shocks, and non-fundamental movements or bubbles in house prices may give rise to imbalances in the economy and in the financial system. The housing price index is deemed necessary. However, comprehensive house price data does not exist in Indonesia. The role of house prices has not been explored to the large extent by price statisticians including those in Statistics Indonesia (BPS). The underlying reasons why these happen are direct indicators of housing prices are not easy to find in large and diverse country like Indonesia. Moreover, it is difficult to construct price index by comparing the house prices of like with like on a regular basis. The compilation of price indices typically relies on *matching* the prices for identical items over time but in the housing context, each property has a unique location and usually a unique set of structural characteristics. Therefore, the matched model methodology will be difficult or impossible to apply. In other words, the matched model method to calculate overall price change is highly problematic for housing. The objective of this paper is to construct property price index especially residential property price index by utilizing property tax as a proxy of property prices. Since property prices are essentially a bundle of performance characteristics, we construct Hedonic Property Price Index using regression technique. Furthermore, we specifically investigate how much more money must be paid for a property with an extra unit of a particular property characteristic. We expect that the empirical result of this study gives the implicit price of a property characteristic that is the marginal price of a characteristic indirectly revealed by the amounts people are willing to pay for the whole property in order to obtain better quality or quantities of the characteristic. The remaining of the paper is organized as the following. In section 2, we will briefly discuss theoretical considerations of how property price index is estimated. Section 3 will discuss data sources and empirical strategies, followed by discussion on empirical results of an application of the hedonic method to property price index in Section 4. Finally, Section 5 will contain conclusions and the policy implications.

2. THEORETICAL CONSIDERATIONS OF PROPERTY PRICE INDEX

The property price index is intended to monitor the development of house prices which are considered important for determinants or indicators for goods price inflation and economic activity. Theoretically, property can be divided into residential and commercial properties. In this study, we only focus on the residential property. The Residential Property Price Index is expected to measure the price changes of all residential properties, both newly built and existing, purchased by households. However, the residential property price index is difficult to calculate with the typical ways by using the matched model. Generally, price statisticians use matched model methodology to calculate price index in which prices of exactly the same item are compared over time. It is the natural starting point for the construction of any price index. But, in the residential property context, due to the low incidence of transactions, and also the quality of houses which continually changes, the standard matched model methodology cannot be applied straightforwardly (Haan & Diewert, 2013). For instance, the newly houses provide a particularly interesting application since they are not homogenous goods. The house prices depend on the location, accessibility, environmental characteristics such as risk of flood, pollution, etc. From these points of view, it can be said that house prices can be viewed as aggregates of individual characteristics. To be specific, residential property has uniqueness of being heterogeneous. Hence, in order to compare the value of transaction prices at different points in time, one should know the attributes or characteristic which can affect house prices. Information on specifications and characteristics is needed to be taken into account in the house prices. House prices are essentially a bundle of performance characteristics. The Hedonic Price Index approach gives insight to measure property price index. This approach recognizes that heterogeneous goods can be described by their attributes or characteristics including house prices using regression technique. The regression technique which is also known as hedonic price model can be used to estimate the property price index including residential property price index. The underlying reason of using hedonic regression is that each house has different attributes or characteristics concerning the location, accessibility, environmental characteristics, etc. In addition, the quality of houses changes rapidly due to technological innovation. In a hedonic regression, the economist attempts to consistently estimate the relationship between prices and product attributes in a differentiated product market. The regression coefficients are commonly referred to as implicit (or hedonic) prices, which can be interpreted as the effect on the market price of increasing a particular product attribute while holding the other attributes fixed. Given utility-maximizing behavior, the consumer's marginal willingness to pay for a small change in a particular attribute can be inferred directly from an estimate of its implicit price; moreover, Bajari, Patrick, et al. (2010) stated that these implicit prices can be used to recover marginal willingness to pay functions for use in valuing larger changes in attributes. Consumer willingness to pay is the maximum amount that a consumer will pay for a property. It measures how much consumer values the property.

3. DATA SOURCES AND EMPIRICAL STRATEGIES

Data used for compiling the Residential Property Price Index are the data on the basis of property tax in Jakarta. Jakarta is a good example to exercise the construction of property price index because Jakarta is the largest and most populous city in Indonesia which attracts many people to find a place to live. We consider Sales Value of Taxable Object (SVTO) by districts/sub-districts of the period 2010 and 2014 in Jakarta to be utilized as a proxy due to the limited data availability. SVTO means the average value that is obtained from sale purchase transaction that is occurred reasonably, and if there is no sale purchase transaction, the SVTO shall be determined with the price comparison of the other equal object. In addition to measure quality improvements for property, other data of property characteristics that are expected to affect property prices are also utilized, including area densities, accessibility characteristics that affect prices, such as distances to business centers, shopping centers, and availability of public transportation infrastructure, as well as environmental characteristics that affect prices such as risk level of flood. We demonstrate the role of a number of characteristics which reflect measures of qualities in decision to buy a house in Jakarta. In the case of residential property, since knowledge of residential property prices are easily known by people, choosing measures of qualities are determined by the subjective and rationale judgment of important considerations for many buyers of house.

Table 3.1. Definition of Variables, Source of Data, Mean, Standard Deviation and Expected Sign

Independent Variable	Definition	Source of Data	Mean	Stdev	Exp. Sign
1. FLOODZ (categorical variable)	Geographical area in DKI Jakarta sub-districts which is vulnerable to flood 1: if the sub-district has low vulnerability to flood 2: if the sub-district has medium vulnerability to flood 3: the sub-district has high vulnerability to flood	Map of Vulnerability to Flood, Disaster Management Board of DKI Jakarta	1,94	0,73	-
2. GREENZ (categorical variable)	Amount of area in DKI Jakarta sub-districts which is functioned as green zone, an open space where the plants grow naturally or deliberately planted. 1: if <40% of the sub-district area are green zone 2: if 40-60 % of the sub-district area are green zone 3: if > 60 % of the sub-district area are green zone	Provincial Regulation of DKI Jakarta No. 1 2012 on spatial plans in 2030, Figure 17. Map of Spatial Pattern Plan of DKI Jakarta Mainland	1,45	0,76	+
3. OFFICEZ (categorical variable)	Amount of area in DKI Jakarta sub-districts which is functioned as building for offices, trades, and services, tend to be centered as other activities related with convenient access to major roads of the city and outside the city, thus ensuring the flow of goods and close to other functions related, such as warehousing, terminal/station. 1: if <40% of the sub-district area are offices, trade, and services area 2: if 40-60 % of the sub-district area are offices, trade, and services area 3: if > 60 % of the sub-district area areoffices, trade, and services area	Appendix 1 of Provincial Regulation of DKI Jakarta No. 1 2012 on spatial plans in 2030, Figure 17. Map of Spatial Pattern Plan of DKI Jakarta Mainland	1,48	0,76	+
4. MALL (numerical variable)	Number of building or a collection of several buildings in each sub-district in DKI Jakarta which assembles a number of independent vendors or variety store with a variety of brand, not merely a supermarkets.	Wikipedia Encyclopedia about lists of mall/shopping center in Jakarta, the location, and the year of the opening of the mall	2,03	2,08	+
5. MARKET (numerical variable)	Number of building in each sub-district in DKI Jakarta consisting of shops or kiosks both permanent and semi-permanent that sells goods traditionally, managed by the local company Pasar Jaya.	Lists of traditional markets and the locations in the website of Pasar Jaya	3,48	2,56	+

Table 3.1. Definition of Variables, Source of Data, Mean, Standard Deviation and Expected Sign *Lanjutan*

Independent Variable	Definition	Source of Data	Mean	Stdev	Exp. Sign
6. BUSWAY (numerical variable)	Number of busway/transjakarta lane which pass througheach sub-districtsin DKI Jakarta.	Transjakarta Network Map in the website of PT. Transportasi Jakarta	2,51	2,57	+
7. TOLL (numerical variable)	Number of toll gate in each sub-district in DKI Jakarta.	Wikipedia Encyclopedia about lists of Jakrata inner and outer ring-road	1,39	1,43	+
8. RAILWAYS (numerical variable)	Number of railway station in each sub-district in DKI Jakarta.	Commuter Line Route Map in the website of PT. KAI Commuter Jabodetabek	1,05	1,13	+
9. BUSINESSD (categorical variable)	Distance ofeach sub-district in DKI Jakarta to Setia Budi sub-district. We assumed Setia Budi sub-districts as bussiness center because there is a high concentration of business activity indicated by large percentage of offices, trade, and services area in there. 1: if the distance of sub-districts to bussines center is < 10 km 2: if the distance of sub-districts to bussines center is 10-20 km 3: if the distance of sub-districts to bussines center is > 20 km	Google maps	1,80	0,70	-
10. DENSITY (numerical variable)	The number of people per one kilometer square area in each sub-district ``in DKI Jakarta.	DKI Jakarta in Figures	4,14	0,26	-

The hedonic model relates property price to the property characteristics. These characteristics reflect measures of quality. Due to the limited data availability, we utilize the data of SVTO by sub-districts in Jakarta as a proxy of property prices and the period 2010 and 2014 as time-varying variable. In the case of residential property characteristics. In empirical strategies, to calculate the Residential Property Price Index in Jakarta, the SVTO of housing of n sub-district/*kecamatan* at period t is set to be P_{tn} . In our strategy, we allow housing prices in Jakarta to be determined by hedonic price function of period 2010 and 2014 as time-varying variable as follows:

$$\ln P_{tn} = \alpha_o + \alpha_t \sum_{t=0}^{1} D_t + \sum_{k=1}^{K} \beta_k Z_{tnk} + \varepsilon_{tn} \tag{1}$$

where:

P_{tn} =	Sales Value of Taxable Object (SVTO) of Housing of n sub-district/*kecamatan*at period t	β_k =	Parameter for characteristic k of house.
α_o =	Constant term	Z_{tnk} =	Characteristic k of housing n at period t.
α_t =	Time dummy parameter at period t.	ε_{tn} =	Error term, the residual from a hedonic regression contains information that we can use to price home attributes that we do not directly observe.
D_t =	Time dummy variable ('1' at the period of 2014 and "0" at the period of 2010).		

To create estimates of the property price index for 2010 and 2014, we take anti logarithm as follows:

Quality-adjusted price index for 2014 = $\dfrac{P_{2014}}{P_{2010}} = \exp(\alpha_t)$ (2)

In order to calculate the consumer willingness to pay, we constraint the derivative of ln P_{tn} with respect to each characteristic to be constant over time. This constraint allow us to recover an implicit price of a property characteristic that is the marginal price of a characteristic indirectly revealed by the amounts people are willing to pay for the whole property in order to obtain better quality or quantities of the characteristic.

4. APPLICATION OF THE HEDONIC METHOD TO PROPERTY PRICE INDEX

Regression estimation results are listed in Table 2 Since heteroskedasticity is a typical problem in the estimation of housing hedonic equations, it was feared that a genuine heteroskedasticity would arise. Surprisingly, the log transformation was sufficient to eliminate heteroskedasticity and passed the commonly used heteroskedasticity tests (Breusch-Pagan/Cook-Weisberg's and White's tests were used). Tests for collinearity resulted in the culling of some variables (not included in Table 2). Table 2 shows that the coefficients of variables YEAR, OFFICEZ, BUSSINESD, TOLL, MALL, MARKET, BUSWAY are significantly different from zero while the coefficients of variables FLOODZ, GREENZ, RAILWAYST, and LN_DENSITY are not statistically significant. All significant coefficients have the expected sign, except MARKET.

Table 2. Implicit Price of The Property Characteristic: Hedonic Model Estimates

LN_SVTO	Coef.	Std. Err.	t	P>t	Exp(Coef.)	WTP4)
D.YEAR	0.4541 1)	0.1463	3.10	0.0030	1.5747	2,571,799
D.FLOODZ						
med	0.1161	0.1770	0.66	0.5140	1.1231	551,084
high	0.2601	0.2087	1.25	0.2170	1.2971	1,329,568
D.GREENZ						
med	0.0768	0.2112	0.36	0.7170	1.0799	357,389
high	0.0503	0.2242	-0.22	0.8230	0.9509	-219,606
D.OFFICEZ						
med	0.7970 1)	0.2023	3.94	0.0000	2.2189	5,454,547
high	0.7988 1)	0.2749	2.91	0.0050	2.2229	5,472,280
D.BUSSINESD						
med	-0.5991 1)	0.1889	-3.17	0.0020	0.5493	-2,016,824
high	-1.2916 1)	0.2893	-4.46	0.0000	0.2748	-3,245,098
D.TOLL	0.3212 3)	0.1757	1.83	0.0720	1.3787	1,694,782
D.RAILWAYST	-0.1373	0.1535	-0.89	0.3740	0.8717	-574,042
BUSWAY	0.0742 2)	0.0338	2.20	0.0310	1.0770	344,754
MALL	0.1459 1)	0.0417	3.50	0.0010	1.1571	703,065
MARKET	-0.0588 3)	0.0350	-1.68	0.0970	0.9429	-255,660
LN_DENSITY	0.2256	0.1604	1.41	0.1640	0.4610	-2,412,137
_CONS	12.3984 1)	1.5059	8.23	0.0000		

Note: 1), 2),3) stands for 1%, 5%, 10% levels of significance respectively, 4) Willingness to pay calculated for marginal 1 unit change in independent variable for median SVTO 2014 of Rp4,474,946

Based on our data, there was not enough statistical evidence that the risk level of flood affect the property price in Jakarta. It is reasonable that the risk level of flood are not becoming a major consideration for the Jakarta's homebuyers since most of the Jakarta area is vulnerable to flood. The concentration of green zone area is not significant as the concentration of office, trade, and services area. The property price of a house in an area near the offices, trade, and services center in Jakarta would be higher than one in an area with high concentration of green zone. The number of railway station is also proven to be not as significant as the existence of toll gate and busway in affecting property price in Jakarta. Population density in Jakarta might not significantly affect the property price due to the relatively high population density in all sub-districts in Jakarta. The property price in a densely populated area in Jakarta would remain high if the area is close to business and shopping centers and also has good transportation infrastructure. The exponentiated coefficient for a certain dummy variable is the ratio of the expected geometric mean for the other groups over the expected geometric mean for the base group, when other variables held at some fixed value. In our data, exponentiated coefficient for area with medium concentration of office, trade, and services is exp (.7970) = 2.2189. We can say that the property price (represented by SVTO) will be 122% higher for area with medium concentration of office, trade, and services than area with low concentration of office, trade, and services. There is not much difference between elasticity of area with medium and high concentration of office, trade, and services. Property price in area in Jakarta which is a bit far from the bussines center will be 45% lower than one which is close to the business center, while the property price in an area which is far from the bussines center will be 73% lower than in the close one. The existence of toll gate access in an area would increase the property price by 38% and an extra busway lane would increase property price by 8% in Jakarta. The existence of mall/shopping center and traditional market affect the property price in Jakarta in different way. For a one-unit increase of mall/shopping center, we expect to see about 16% of an increase in property price. However, one-unit increase of traditional market might make the property price fell by 6%.

The result of this study also gives the implicit price of a property characteristic that is the marginal price of a characteristic, or in other word, the amounts people are willing to pay for the whole property in order to obtain better quality or quantities of the characteristic.The marginal implicit price of the variable of interest is calculated by differentiating the hedonic price function with respect to the variable. Therefore, in our case, each unit-increment in the number of mall/shopping center concentration results in an estimated increase in the property value (represented by SVTO) times the property value. Since the SVTO data is the price in a meter square area, for example, a sub-district whose SVTO is Rp4,474,946 per meter square would have a marginal benefit of a unit increase in the number of mall/shopping center equivalently Rp703,065 per meter square. In other words, the homebuyer is expected to be willing to pay Rp703,065 per meter square to enjoy the benefit of a one unit-increase in mall/shopping center. Using the same way, from the regression estimation coefficient, we can say that the homebuyer is expected to be willing to pay about 5.4 million rupiahs per meter square to have property near the offices, trades, and services area, Rp1,694,782 per meter square to have property with toll gate access, and Rp344,754 per meter square to have property with an extra busway lane in the area. For variable distance to bussiness center, the homebuyer is expected to be willing to pay no more than 2-3 million rupiahs to avoid the disadvantage to property value associated with the increasing distance from the business center. In this study, traditional market has negative effect to the property value. The existence of traditional market is associated with a noisy environment and tends to be seedy. Thus, the homebuyer is expected to be willing to pay no more than Rp255,660 to avoid the disadvantage to property value associated with a one unit increase of the number of traditional market in the area.

Table 3. Property Price Comparison Intercity in Jakarta

LN_SVTO	Coef.	Std. Err.	t	P>t
CITY				
island	-3.5681	0.4497	-7.93	0.0000
south	-0.3981	0.2698	-1.48	0.1440
east	-1.3616	0.2698	-5.05	0.0000
west	-0.7400	0.2844	-2.60	0.0110
north	-0.7048	0.3072	-2.29	0.0240
_CONS	15.8706	0.2011	78.91	0.0000

By assuming that Central Jakarta is the base city, it is known that of 6 regions in Jakarta, all regions have coefficients with negative sign meaning that Central Jakarta is the most expensive city in terms of SVTO (Table 3). Meanwhile, property price index estimation could be computed by taking anti logarithms of the estimated coefficients of the time dummy variables. Normalizing the base year value to unity in 2010, hedonic price index in 2014 is 157.47 (Table 4). It means that the property price increases about 57.47 percent over the period 2010-2014. The increase of hedonic property price index is relatively higher than that of Residential Property Price Index (RPPI) at the same period calculated by the Central Bank of Indonesia (37.97 percent). In turn, the relatively high difference between Hedonic Property Price Index and RPPI can give different policy implication to monitor the possibility of asset bubbles.

Table 4. Hedonic and Residential Property Price Index

Property Price Index	Year	Estimated Coefficient	Base Year	Price Index	% Change (2010-2014)
Hedonic Property Price Index	2010		2010=100	100.00	
	2014	0.4541		157.47	57.47
RPPI (Central Bank of Indonesia)	2009 QIV		2002=100	144.89	37.97
	2014 QIV			199.91	

5. CONCLUSIONS AND POLICY IMPLICATIONS

Our paper demonstrates an estimation of residential property price index using hedonic price model. Having identified the property characteristics that affect the property prices in Jakarta, it is known that areas of offices, trades, and services, distance to business center, toll gate access, malls, traditional markets, and busway/transjakarta lane significantly influence the residential property prices in Jakarta. Nevertheless, geographical area which is vulnerable to flood, green zone, number of railway station, population density are statistically insignificant to influence the property prices based on our data.

The information of consumer willingness to pay is important for the Local Government of DKI Jakarta when they judge whether to determine appropriate residential property tax or not when a certain property characteristic has already improved by them. In other words, the consumer willingness to pay can be used as valuation price for taxation for local services. Also, the Central and Local Government could have an idea as to what extent the housing payment affordability is necessary for citizen. Based on empirical results, it is known that consumer willingness to pay for marginal 1 unit change in offices, trades, and services areas is the highest in comparison with other variables, followed by the toll gate access, and busway lane area. It means that the consumer will enjoy the benefit to have property associated with those variables. Meanwhile, the homebuyer is expected to avoid the disadvantage to property value associated with the increasing distance from the business center. In this study, traditional market has negative effect to the property value. The existence of traditional market is associated with a noisy environment and tends to be seedy. Thus, the homebuyer is expected to avoid the disadvantage to property value associated with a one unit increase of the number of traditional market in the area.

The empirical results show that the growth of residential property prices increase rapidly during the period of 2010-2014. The already rising residential property prices lead to increases in aggregate national wealth. In turn, it can be used as an indicator of economic activity and inflation rate as well.

Further research work is encouraged to investigate accurate indicator of residential property price index at various geographic levels, not only Jakarta region. Hopefully, the property price index can provide useful information for estimating changes in the rates of mortgage defaults, and housing payment affordability in specific geographic areas.

REFERENCES

Bajari, Patrick, et al. (2010). Y-based approach to hedonic price regression with time-varying unobserved product at the price of pollution. *Working paper 15724 The National Bureau of Economic Research.*

Berndt, E. R. (1991). The measurement of quality change: constructing an hedonic price index for computers using multiple regression methods. *The practice of econometrics: Classic and contemporary*, 127-131.

Bhattarai, G. R., Pandit, R., & Hite, D. (2004). Willingness to pay for public goods: A hedonic demand model for neighborhood safety, school and environmental quality. In *2004 Annual Meeting, February 14-18, 2004, Tulsa, Oklahoma* (No. 34628). Southern Agricultural Economics Association.

Bishop, K. C., & Timmins, C. (2015). Estimating the Marginal Willingness to Pay Function Without Instrumental Variables. *Available at SSRN 2748657*.

Chay, K. Y., & Greenstone, M. (2004). *Does air quality matter? Evidence from the housing market*. National Bureau of Economic Research.

Melser, D. (2005). The hedonic regression time-dummy method and the monotonicity axioms. *Journal of Business & Economic Statistics*, *23*(4), 485-492.

Haan, J. D., & Diewert, W. E. (2013). Handbook on Residential Property Price Indices.

Alsherfawi Aljazaerli, M. (2005). Hedonic valuation of marginal willingness to pay for air quality in metropolitan Damascus.

Van Ommeren, J. N., & Van der Vlist, A. J. (2016). Households' willingness to pay for public housing. *Journal of Urban Economics*, *92*, 91-105.

A Micro-Level View of Housing Affordability in Malaysia Using an Age Cohort-Housing Type Analysis[1]

Gary John Rangel
Universiti Sains Malaysia, Georgetown, Malaysia – gary.rangel@usm.my

Jason Wei Jian, Ng[2]
Monash University Malaysia, Bandar Sunway, Malaysia – jason.ng.wj@monash.edu

Thangarajah@M. Thiyagarajan, Murugasu
Monash University Malaysia, Bandar Sunway, Malaysia – thangarajah.thiagu@monash.edu

Wai Ching, Poon
Monash University Malaysia, Bandar Sunway, Malaysia – poon.wai.ching@monash.edu

Abstract

Housing affordability has been a troubling issue for both developed and developing countries. Much of the priori literature has used the price-income ratio as the main standard in assessing housing affordability, with the median disposable household income of the sample population commonly used. As such, conclusions drawn based on such macro-level analyses are devoid of useable policy recommendations. This paper attempts to address the housing affordability issue at a greater level of detail by considering age cohorts and different house types. Using household monthly income data by age cohort provided by the Malaysia's Department of Statistics (DOS), and house price data by type of housing obtained from the National Property Information Centre (NAPIC), we construct a more in-depth view of housing affordability via a descriptive analytics approach. We dissect the household monthly income data by age cohort at the 20th, 40th, median, and 75th percentile. As DOS collects household monthly income on a biennial basis, we interpolated the data to obtain the household monthly income on a yearly basis from 1995 to 2014 using the cubic spline method. Monthly mortgage payments for the different house types over the corresponding years are calculated based on the base lending rate (BLR) data obtained from one of the prominent local banks. It assumes a 10 per cent down-payment upon house purchase, with varying tenure periods dependent on the borrower's age, with monthly compounding of the BLR. Using the common benchmark by banks, a cut off ratio of 0.35 for the mortgage instalment to income ratio is used as a measure of housing affordability. The results show that for those aged 20-24 and 60-64, housing is severely unaffordable across all house types over the sample period, except for the 75th income percentile group. Semi-detached and detached housing types, which are the priciest, remain unaffordable to all groups except for the 75th income percentile group aged 30 years and above. Terrace and high rise housing types are the most affordable across each income percentile. Concerning the affordability issue over time, the mortgage payment-income ratio was extremely high in the early years of the sample, averaging at about one. However, it has steadily reduced over time but has largely remained stagnant in the past decade across all income percentiles and age cohorts. This shows a lack of housing affordability improvement even with recent government interventions through affordable housing schemes specifically targeted at the lower income groups. This paper makes several contributions. By analysing the affordability problem from a more granular perspective, strategies to develop a thriving rental market should be targeted to those in the 20-24 age cohort. Concerted effort in information dissemination by the government needs to be undertaken to convince would-be house purchasers that the best time, on average, to enter the housing market would be when they reach their mid-30s. Emphasis should also be placed on housing starts in the high-rise and terrace housing types. This analysis also helps commercial mortgage lenders segment their clienteles in a more refined manner and thus helps reduces information asymmetry.

Keywords: Malaysia; Housing affordability; Age cohort; Cubic spline.
JEL Classification: R21; R31; R38

---

1 We gratefully acknowledge funding support from the Fundamental Research Grant Scheme (FRGS/1/2014/SS07/MUSM/03/2) provided by Malaysia's Ministry of Education.

2 Corresponding author.

1. INTRODUCTION

The off-late hotly debated issue of housing affordability in Malaysia stems from the exponential increase in house prices across the country which have outpaced the increase in household incomes (Lee & Lye, 2014, The Star, 2014). The debate has also been exacerbated by the variants in definitions provided by the government, industry and academic scholars, sometimes resulting in contentious conclusions made about the overall extent of affordability.

Several proposals have been implemented to bridge the affordability gap. The demand-driven proposals seek to increase the ability of households to purchase property. They include the proposal for developers to be loan providers, commercial banks extending tenure of housing loans from 35 years to 40 years, and the Employee Provident Fund (EPF) allowing contributors to withdraw more of their retirement funds to buy homes (Mahalingam, 2016). Other proposals are supply-driven which involve providing affordable housing according to incomes of certain target groups, and new methods of construction such as the implementation of the industrial building systems (IBS) that lowers construction costs and ultimately lead to lower house prices. Apart from private sector initiatives, the government has also stepped up its role to alleviate the problem of affordability. These initiatives include the establishment of a National Housing Council (NHC) to develop appropriates measures and actionable plans to build affordable housing ranging from RM150,000 to RM450,000, the incorporation of Perbadanan PR1MA Malaysia (PR1MA) with the goal to plan, develop, construct, and maintain high-quality housing with a lifestyle concept that is conducive to the middle-income group in urban centres (The Star, 2013b, Surendran, 2016).

Notwithstanding the progress that has been made by the initiatives underlined above, the goal of providing affordable housing to all across the specific target groups, such as the first-time homebuyers, remains elusive (The Star, 2013a). Even government initiatives like PR1MA have "lost its focus" and have even allowed 'second-time house buyers' a chance to ballot and buy PR1MA developments even though there is a waiting-list of 1.3 million registrants (Surendran, 2016). The lack of affordable housing can lead to serious economic and emotional implications for households, especially for those in the low-to-moderate income group. For example, a longitudinal analysis of household movements in Australia found significant mental health deterioration of individuals living in low-to-moderate income households when housing costs exceed 30% of household income (Bentley et al., 2011).

This paper presents a micro-view of housing affordability for Malaysia, addressing the issue at a greater level of detail by considering age cohorts and different house types, and subsequently provide case-specific policy recommendations. Firstly, the micro-view stems from the premise that the issue of affordable housing should not only emphasize affordability at the median household income level but also whether house prices are out of reach for those below the median household income. This takes added importance as the Malaysian government has put great emphasis in uplifting the well-being of the bottom 40% household income group (B40 household) (Economic Planning Unit, 2015). Secondly, the level of housing affordability stress could be unevenly distributed, depending on the demographics of age. By dissecting household income into various age groups, this paper provides a thorough analysis on which age groups are more vulnerable to housing affordability stress. This motivation stems from prior literature that different age groups experience differing levels of stress related to affordable housing (Bujang et al., 2015, Zielinski, 2016). Lastly, analysing the differences in affordability by type of dwelling can be an input to the design of various housing supply strategies initiated by both the federal and state governments in Malaysia.

The rest of the paper is as follows. Section 2 describes the key definitions of housing affordability and related prior literature examining the extent of the housing affordability issue. Section 3 describes the data used in the analysis with explanations on the assumptions used to calculate the mortgage-income ratio. The following section after provides a thorough analysis of housing affordability stress faced by Malaysian households and the policy implications that authorities need to address. Section 5 concludes the paper.

2. LITERATURE REVIEW

The literature on housing affordability indicates a close connection between financial institutions and the housing market. Affordability is determined by house price, financial standing of the house buyer, and the financial institution (Gan & Hill, 2009, Yuen et al., 2006). The mismatch between the three may lead to unaffordability in a home purchase. Several approaches have been outlined in the literature to measure housing affordability.

- Price or expenditure-to-income ratio is the most commonly used in housing affordability studies across many geographies (Bogdon & Can, 1997). It is calculated as the ratio of the mean/median price to the mean/median household income. Two measures of affordability can be derived. Using the price-to-income indicator would measure access to housing whereas the housing expenditure-to-income ratio measures affordability after the household owns a home (Chen et al., 2010). This measure states that the household should not spend more than 30%-35% of monthly income on housing loan monthly payment (Carter, 1997, Murray, 1998).
- An increasingly popular approach tackles the issue of housing affordability through the lens of residual income. The residual income approach generally postulates that housing costs become an issue only when the residual income after housing expenditure fails to support a socially acceptable non-housing expenditure. This is then considered as housing induced poverty. This approach relies on identification of non-housing expenditure and how taxes are derived (Stone, 2006). The drawback with this approach stems from the lack of consensus as to what level of socially acceptable housing expenditure would be deemed as desirable (a normative approach) (Chen et al., 2010).

This study subscribes to the mortgage-income approach whereby for housing to be considered affordable, the monthly mortgage payments should not exceed 30% of gross monthly income (Tan, 2013). Notwithstanding its limitations, the mortgage-income ratio is easy to compute and comprehend. Being in ratio form, it can also be compared over time, allowing for a descriptive analytics approach in determining the overall view of housing affordability in Malaysia, but at a micro-level. We use this approach with the assumption that households are able to make good the necessary upfront costs of owning a home. This is the typical 10% up-front payment made by first-time homebuyers.

3. METHODOLOGY

The calculation of the annual mortgage-income ratio first involves obtaining data from two data sources – the CEIC database which provides data on house price indices, and the Malaysian Department of Statistics which provided data for household incomes.

The annual Malaysian House Price Index (MHPI) from 1988 to 2015 was obtained from CEIC, with year 2000 indexed as the base year. The MHPI is a measure of overall house prices, calculated as the weighted average of four sub-indices measuring the house prices for the four housing types in Malaysia: terrace, high-rise, semi-detached, and detached housing. Therefore, corresponding indices for these four housing types from 1988 to 2015 were also obtained. In order to calculate the annual house prices corresponding to these five price indices, we utilized the house prices published in Malaysia's National Property Information Centre (NAPIC) for year 2003, and imputed a series of average house prices for the overall and four housing types for the years 1988-2015.

Using the five annual house prices (i.e. overall house price and prices for the four housing types) calculated for years 1988-2015, we proceeded to derive the associated monthly mortgage payments. For this paper, we assume that the starting present value of the mortgage would be 90% of that year's house price, as the minimum down payment is usually at 10% of the house price, especially for first-time homebuyers. In recognizing that homebuyers from different age groups are able to apply for loans of different tenures, we calculate the monthly mortgage payments based on loan tenures of 35 years, 30 years, 25 years, 20 years, and 15 years using a normal annuity formula. Table 1 below lists the assumed loan tenures for the respective age groups. These assumed loan tenures are based on (i) the setting of the maximum loan tenure at 35 years and (ii) the maximum allowable age by which the mortgage should be repaid is 75 years old. The interest rate used was the Base Lending Rate (BLR) also obtained from CEIC.

Table 1: Assumed Loan Tenure Periods for The Respective Age Groups

Age Group	Loan Tenure (Years)
20-24	35
25-29	35
30-34	35
35-39	35
40-44	30
45-49	25
50-54	20
55-59	15
60-64	15

Household income data was obtained from the Malaysia's Department of Statistics (DOS). As the DOS does not conduct the household income survey on an annual basis, household income data by age group was only available for the following survey years: 1995, 1997, 1999, 2002, 2004, 2007, 2009, 2012, and 2014. In particular, the DOS provided for us monthly incomes at the 25th, 40th, 50th and 75th percentiles for the respective age groups in those survey years. The age groups considered are also those listed in Table 1. As our interest is in tracking the affordability of housing over time from 1995 to 2014, we obtained the monthly household incomes for the missing years by interpolation via the cubic spline to obtain smoothed values of monthly household incomes for the years that DOS did not conduct the survey.

Finally, using the monthly mortgage payments derived from the prices of the overall housing and four dwelling types, combined with the monthly incomes, we are able to compute the associated mortgage-income ratios for households at the different income percentiles, categorized by their age groups, for years 1995 to 2014.

4. RESULTS & POLICY IMPLICATIONS

4.1. Results

For brevity of the paper, we only report results pertaining to households at the 40th income percentile corresponding to the B40 households. Figure 1 plots the associated mortgage-income ratios from 1995 to 2014, for homebuyers in the respective age groups, taking into account the differing loan tenures as listed in Table 1. The mortgage-income ratios reported are for the overall house prices and four housing types.

The plots in Figure 1 also include two horizontal lines at the ratio marks of 0.30 and 0.35. These values represent the affordability threshold and are set based on the basic rule of thumb that any single monthly loan repayment should not exceed a third of the borrower's monthly income. Therefore, any ratio values above these threshold lines indicate that more than a third of the borrower's monthly income is used to service the monthly mortgage, rendering the house unaffordable when measured against the household's income.

Several insights can be gleaned from Figure 1. Firstly, housing affordability has generally improved over the past 2 decades for all age groups, as shown by the overall downward trend in the mortgage-income ratios. However, housing affordability has remained stagnant over the past decade as evidenced by the plateaued ratios since 2005. Secondly, households from the youngest and oldest age groups are faced with the worst housing affordability situation, with even the cheapest housing types (i.e. terrace and high-rise housing) out of their reach. Thirdly, since 2010, the terrace and high-rise housing are considered affordable for households in the 30-34 age group, with the monthly mortgage payment constituting less than 35% of their monthly incomes. Lastly, the high end properties (i.e. semi-detached and detached housing) are out of reach for all households at the 40th income percentile, regardless of age groups.

We also briefly state the following findings for households at the 25th, 50th and 75th income percentiles:

- All housing types are unaffordable for households at the 25th income percentile, across all age groups. In particular, the affordability challenge is worst for households in the youngest and oldest age groups.

- For households at the 50th income percentile, terrace and high rise housing are affordable for those aged 25 to 59. All housing types remain out of reach for households in the youngest and oldest age groups.
- For households at the 75th income percentile, terrace and high rise housing are more than affordable for all age groups, with the mortgage-income ratio averaging at about 0.20. The high-end properties are affordable only for households aged 30 years and above.
- Mortgage-income ratios show an overall downward trend from 1995 to 2014, but have plateaued in the last 10 years.

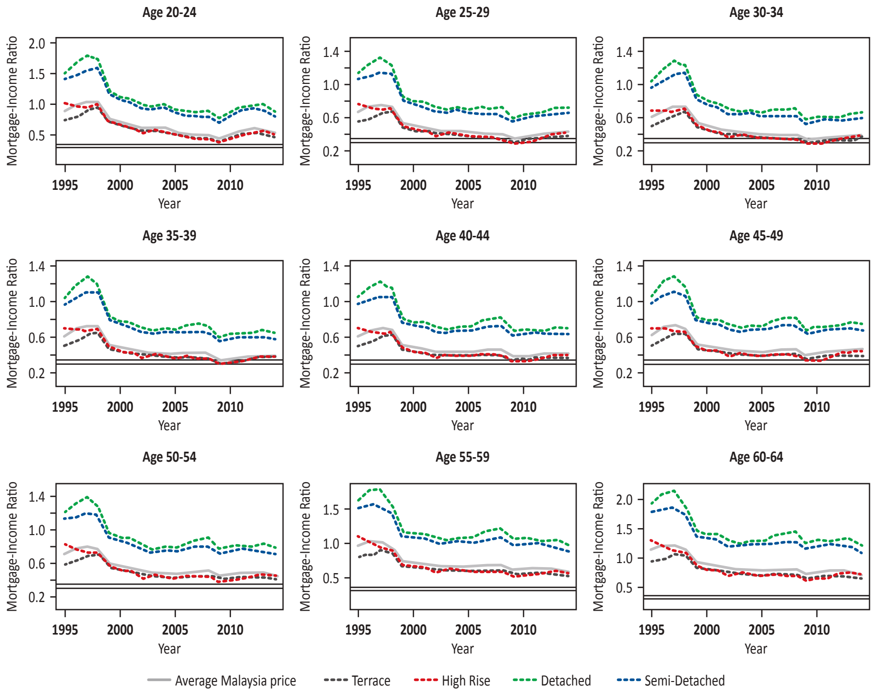


Figure 1: Mortgage-Income ratio from 1995 to 2014 Across Different House Types, for 40th Income Percentile Households in Respective Age Groups

4.2. Policy Implications

The results in Section 4.1 provides several important insights in terms of policy implications. Firstly, greater housing affordability stress is observed from the younger age cohorts. It is therefore recommended that these age cohorts enter the rental market rather than going through undue stress of owning a home. Support measures by the federal and state governments to create a vibrant rental market would help alleviate the pressures on these cohorts. The idea of creating a vibrant rental market is not new as Malaysia's central bank, in its latest annual report, mooted the idea of the rental market being accorded equal weightage in national housing policies (Bank Negara Malaysia, 2015). Countries with severe housing affordability issues such as Switzerland, Germany, and Australia have taken the lead in developing a vibrant rental market. Such an initiative would gather greater impetus provided that the appropriate rental data be collected and made available for dissemination to the public at large. Malaysia has yet to collate data on rentals to develop rental indices for the country and corresponding states, compared to data on house price indices. The development of timely and frequently updated rental indices is paramount in ensuring would-be renters have the necessary decision-making information before committing to a rental contract. An example for Malaysia to emulate is the rental index as published by the Urban Redevelopment Authority of Singapore (URA) on a quarterly basis.

Secondly, frequent and timely dissemination of the MHPI and sub-indices would also help reduce information asymmetry for would-be homebuyers. The current update by the National Property Information Centre (NAPIC) is until quarter 3, 2015, with finalized data for quarter 4, 2015 as preliminary data. This translates into a 5-quarterly period lag. In contrast, Singapore's URA updates its property price index and rental index on a quarterly basis with a one-quarter lag.

Thirdly, rising migration from the rural to urban states, and foreign purchases have undoubtedly enhanced the housing demand and pushed up the house prices. Due to the rapid increase in house prices, the gap between house prices and income levels is widening. Residents who stay in the more developed states are facing difficulties to afford homeownership. People are increasingly relying on the private real estate market for housing demand, but the problem is the shortage of the supply of reasonably priced houses in more developed states. As our analyses indicates that certain house types are more affordable, government initiatives should therefore focus on increasing supply of the high rise and terrace segment and correspondingly reduce housing started for semi-detached and detached housing types. As these housing types are mass produced, the use of the Industrialised Building System (IBS) as suggested by Ismail et al. (2015) would reduce the overall cost of delivery.

Fourthly, there is a need to reassess government policy at all levels, to ensure that only the genuine intended first-time buyers are granted the opportunity to buy the property out of need. For example, PR1MA, as an initiative to assist young adults to own their first house, should strictly not allow second-time house buyers qualify for PR1MA to purchase second homes. Otherwise, this would fail to satisfy the most deserving category when the supply is not enough for first-time buyers.

5. CONCLUSION

We analysed the housing affordability of house buyers by comparing the mortgage-income ratios vis-à-vis a standard that a ratio of 0.3-0.35 would be deemed as affordable. The results showed that although housing affordability in general has increased since the beginning of the sample period, it has stagnated in the last 10 years. As such, the various policies by the federal and state governments to provide affordable housing have not improved housing affordability especially for the B40 households. As such, other measures such as the development of a vibrant rental market would suffice, coupled with the dissemination of timely information on both prices and rents to reduce information asymmetry.

REFERENCES

Bentley, R., Baker, E., Mason, K., Subramanian, S.V. & Kavanagh, A.M. (2011). Association between housing affordability and mental health: A longitudinal analysis of a nationally representative household survey in Australia. *American Journal of Epidemiology,* **174,** 753-760.

Bogdon, A.S. & Can, A. (1997). Indicators of local housing affordability: Comparative and spatial approaches. *Real Estate Economics,* **25,** 43-80.

Bujang, A.A., Jiram, W.R.A., Abu Zarin, H. & Md. Annuar, F.H. (2015). Measuring the Gen Y housing afforability problem. *International Journal of Trade, Economics and Finance,* **6,** 22-26.

Carter, T. (1997). Current practices for procuring affordable housing: The Canadian context. *Housing Policy Debate,* **8,** 593-631.

Chen, J., Hao, Q. & Stephens, M. (2010). Assessing housing affordability in post-reform China: A case study of Shanghai. *Housing Studies,* **25,** 877-901.

Economic Planning Unit (2015). Elevating B40 households towards a middle-class Society. Ed. E.P. Unit. Putrajaya, Malaysia: Economic Planning Unit.

Gan, Q. & Hill, R.J. (2009). Measuring housing affordability: Looking beyond the median. *Journal of Housing Economics,* **18,** 115-125.

Ismail, S., Jalil, I.N. & Megat Muzafar, P.M. (2015). Making housing affordable. Kuala Lumpur: Khazanah Research Institute.

Lee, P. & Lye, A. (2014). Salaries not going up as fast as prices of homes. in *The Star* Penang, Malaysia: Star Publications (M) Bhd.

Mahalingam, E. (2016). Budget needs to address affordable housing issue. in *The Star* Penang, Malaysia: Star Publications (M) Berhad.

Malaysia, B.N. (2015). Annual Report. Kuala Lumpur: Bank Negara Malaysia.

Murray, M.S. (1998). Housing conditions of nuclear and extended households in the Miami-Fort Lauderdale area. *Population and Environment,* **20,** 55-76.

Stone, M.E. (2006). What is housing affordability? The case for the residual income approach. *Housing Policy Debate,* **17,** 151-184.

Surendran, S. (2016). 'Serious mismatch' in housing supply. in *The Edge Financial Daily* Selangor, Malaysia: The Edge Communications Sdn. Bhd.

Tan, W.K. (2013). Housing the nation: Policies, Issues and Prospects. Kuala Lumpur, Malaysia: Cagamas Holdings Berhad.

The Star (2013a). 1.7 million households yet to own houses. in *The Star* Penang, Malaysia: Star Publications (M) Bhd.

The Star (2013b). Budget 2014: Initiatives a big boost for affordable housing. in *The Star* Penang, Malaysia: Star Publications (M) Bhd.

The Star (2014). Key issues facing homeownership in Malaysia. in *The Star* Penang, Malaysia: Star Publications (M) Bhd.

Yuen, B., Kwee, L.K. & Tu, Y. (2006). Housing affordability in Singapore: Can we move from public to private housing? *Urban Policy and Research,* **24,** 253-270.

Zielinski, C. (2016). Housing affordability a growing issue for older Australians. http://www.domain.com.au/.

Property Cycle of Indonesia[1]

Justina Adamanti*
Bank Indonesia, Jakarta, Indonesia – justina@bi.go.id

Muhammad Sahirul Alim
Bank Indonesia, Jakarta, Indonesia – msahirulalim@gmail.com

Arlyana Abubakar
Bank Indonesia, Jakarta, Indonesia – arlyana@bi.go.id

Abstract

Understanding property cycle is important in macroprudential policy formulation for property sector. This study aims to construct property cycle of Indonesia, which focused on housing market, using spectral analysis and frequency based filter. Data included are Residential Property Price Index (RPPI), CPI:Housing and mortgage loan. The result shows the evidence of common cycle between RPPI and CPI:Housing in frequency domain for duration around eight years. Therefore, the property cycle is constructed from both of those variables, which result can explain the event analysis well.

Keywords: Property cycle; Spectral analysis; Frequency based filter; RPPI.

JEL: E32, E58

1. INTRODUCTION

As one of important economic sector, property can have cycle as in business cycle since it can also experiences boom and bust. As an option for asset placement, property is influenced by the procyclicality in business cycle. The increasing of property price in expansion phase can lead to risk taking behavior on economic agent, which in turn can harm the financial system.

In Indonesia, property is still interesting for asset placement. Based on the survey for Residential Property Price Index (RPPI) in primer market conducted by Bank Indonesia on 3rd quarter of 2016, 74.77% people use mortgage as main source of property financing. Recently, the share of mortgage is around 10% of total credit. However, the rapid increasing in property price can influence economic agent perception about the recent economy. Therefore, as the macroprudential regulator, Bank Indonesia has a respobility to manage the property price using its policy, Loan to Value ratio (LTV). This policy controls portion of property price that can be financed by mortgage, based on the recent development of property price, macro economic, and also financial system condition.

LTV is a countercyclical policy in the property sector. The portion of financing from the loan is decreased in expansion phase, in the hope that it will help reducing the growth of mortgage. While in the contraction phase, the portion of financing from the loan is decreased, in the hope that it will contribute to boost the growth of credit by increasing the growth of mortage. In line with that, the proper policy has to be supported by much information. In case of LTV, property cycle can be one of the references for understanding the phase of property price, in addition to information from business and financial cycle.

This study is purposed for constructing property cycle of Indonesia, as one of the references of policy implementation in property sector. The scope of property in this study is focusing on housing property/ residential. Furthermore, the variable discuss in this study is restricted to RPPI, mortgage (House Ownership Loan – KPR) and Consumption Price Index component related to housing (CPI : Housing).

This study is organized as follows: section 2 reviews some related recent studies, section 3 describes methodology and data used ini this study, section 4 discusses the results, and section 5 concludes the study.

---

1 The views expressed in this study are those of the authors and do not necessarily represent those of Bank Indonesia.

2. RELATED LITERATURE

The cycle in property price come from the lags between demand and supply. If the property can be produced quickly to meet new demands, the market will always reach equilibrium, so that property cycle does not exist. But in reality, market mechanisms sometime fail to balance supply and demand.

According to Gover & Gover (2013), property cycle has four phases, namely recovery, prosperity, recession, and depression. The phase in property cycle is similar to phase in financial and business cycles, since all of them are related. At recovery stage, the property demand begins to increase along with availability of credit from the financial sector. The property rental price slowly begins to increase and some property projects start to begin. In the next stage, prosperity, the increasing in demand causes property developer to be more aggressive and speculative in offering new property products. The property price also increases very high as the developer needs more funding for new property projects. If it continues, the property cycle can turn to recession phase which is indicated by decreasing in demand. At this stage the sector experiences excess supply which causes the property price and rent cost begin to decline. Lastly, the property cycle will experience depression phase in which many developers fall on bankruptcy due to a significant drop in demand.

One of the issues in modeling the property cycle is the existing of non-fundamental fluctuations in time series data that can lead to misspecification in property cycle structure. Therefore, it is necessary to decompose the main cycle component from the data. A common method for constructing the property cycle is spectral analysis which can determine the dominant frequency of a variable.

Some studies use spectral analysis in constructing property cycle are Gray (2015), Liow (2007), Wilson & Okunev (1999), and Wang (2003). These studies discuss the property cycle length using several data such as property prices, returns on real estate stocks, or property securitisation such as Real Estate Investment Trust (REIT). The use of returns on stocks data for constructing property cycle is still a debate, since it contains risk taking behavior that sometimes do not reflect the physical property market. However, this data is often used because it can reflect the performance of the property company that can affect the property cycle, and also available in length data.

Meanwhile, there are several methods that can be used to separate cycle component from time series data, such as the first-difference, moving-average, Hodrick-Prescott (HP) filter and Band Pass Filter.

Based on several studies, the length of the property cycle is varied, depends on the economic characteristics of the country/region. Developed countries generally have longer cycle length than developing or small countries. Man & Chau (2007) used the index of total return of various segments in real estate market, rental return, and property prices in Hong Kong. The result shows the property cycle length ranges from 36-50 months. Wilson & Okunev (1999) use property stock data and data from securitized property market, which is believed can reflect the cycle of physical properties. Using data from securitized property market, the study proves that the developed countries tend to have longer duration, such as in US and UK which is ranged between 85 and 96 months compared 32 months for Australia. Nevertheless, the result is different when using the property stocks data. Property cycle has shorter cycle length, which is ranged between 33 to 36 months for US, 58 months for UK, and 42 months for Australia.

Liow (2007) uses the return of real estate stocks as a proxy of real estate performance. The study shows similar results with study of Wilson & Okunev (1999) when using property stock data, the duration of the property cycle is ranged between 30 to 44 months.

3. METHODOLOGY AND DATA

This study uses spectral analysis method to determine the dominant frequency of a variable. The analysis was performed for each stationary variable individually and in pairs. Data must be stationary, since trends can lead to misspecification in spectral analysis results. Stationary data is obtained by de-trending data using several methods, such as yearly growth, first difference, moving averages of four data points (MA4), and moving average of six data points (MA6). For every individual variable, spectral analysis is performed using parametric and non-parametric approaches. The parametric approach is performed using ARMA model, while the non-parametric approach is performed using Tukey-Hamming, Danielle, and Bartlett.

After finding dominant frequency for every individual variable and variables in pairs, the main cycle is filtered using Christiano-Fitzgerald Band Pass Filter, which is allowed the user to filter desired frequency range from time series data. For variables in pairs, common cycle is constructed from the average of two cycles which are already filtered on its dominant frequency, as in Drehmann et al.(2012). Finally, property cycle will be chosen based on comparison to event analysis of crises and the implementation of LTV.

Spectral Analysis

A time series data can have low or high frequency as the dominant frequency. Low frequency means a variable has a longer duration that can be categorized as medium or long-term cycle, while high frequency is vice versa. Low frequency variable requires longer duration to recur back to similar pattern, so it can be called as medium or long-term cycle. While high frequency variable requires shorter duration to recur back to similar pattern, so it can be called as short-term cycle.

Frequency identification is important in the decomposition of cycle components from a time series data. A priori or predefined selection of frequency can produce misleading cycles that do not represent the actual cycle (Gonzalez, Lima, & Marinho, 2015). Spectral analysis is a method that can analyze the dominant frequency of time series data. Basically, spectral analysis transforms time series data from time domain to frequency domain. After that, this method analyses the contribution of a frequency to the total variance of time series data.

Spectrum of a stationer series $\{Y_t\}$ can be written as:

$$f(\omega) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k} \qquad (1)$$

Where $\gamma_k = cov(Y_t, Y_{t-k})$ is an auto-covariance of $\{Y_t\}$, and ω is the frequency.

Equation (1) is equivalent to:

$$f(\omega) = \frac{1}{\pi}\left[\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k cos(\omega k)\right] \qquad (2)$$

Spectrum, also called as spectral density, can be seen as the relative strength of each frequency, the contribution to the total variation of time series data. In this study, the spectrum is estimated using non-parametric and parametric approach.

Cross Spectral Analysis

The purpose of the cross spectral analysis is to determine the existence of a common cycle between two variables based on the co-spectral density, coherence, and phase.

Cross spectrum of the two stationary series $\{X_t\}$ and $\{Y_t\}$, can be written as follows:

$$f_{xy}(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_{xy} e^{-i\omega k} = C_{xy}(\omega) - iq_{xy}(\omega) \qquad (3)$$

Where γ_{xy} is cross-covariance function between two series.

The cross spectrum can be partitioned into a cross amplitude spectrum, the relationship between the magnitudes of the components in the series at different frequency:

$$a_{xy}(\omega) = |f_{xy}(\omega)| = \sqrt{C_{xy}(\omega)^2 + iq_{xy}(\omega)^2} \qquad (4)$$

And the phase of the spectrum is:

$$\varphi_{xy}(\omega) = tan^{-1}[-q(\omega)/c(\omega)] \qquad (5)$$

The coherence between $\{X_t\}$ and $\{Y_t\}$ at frequency ω:

$$C_{xy}(\omega) = \frac{a_{xy}(\omega)^2}{f_{xx}(\omega) f_{yy}(\omega)} \tag{6}$$

The value of coherence is between 0 and 1. Higher coherence value indicates stronger relationships between the two series.

Data

This study uses Residential Property Price Index (RPPI) as proxy for housing price, mortgage loan, and CPI:Housing in quarterly format from 2000Q1 to 2016Q2. RPPI is constructed from Bank Indonesia regular survey for housing price in primer market in 16 big cities in Indonesia. CPI: Housing is a part of Consumer Price Index, which its components consist of materials for house including wage for manpower. All variables are in seasonal adjusted form, transformed into log, indexed to 2012Q2, and detrended to be stationer. Indexing is intended for data comparison and data integration in common cycle. 2012Q2 is selected as indexing point because LTV have been implemented since that time.

4. RESULTS

Before performing spectral analysis, the data have to be detrended to eliminate noise. Detrending methods used in this study are yearly growth, first difference (FD), moving averages of four data points (MA4), and moving average of six data points (MA6). Table 1 presents stationary results for all detrended data.

Table 1. Stationarity Test Results for Detrended Data

Data	Stationarity Test	Growth YoY		First Difference		MA4		MA6	
		t-stat	prob	t-stat	prob	t-stat	prob	t-stat	prob
CPI: Housing	ADF	-2.39	0.15	-6.89	0.00	-3.38	0.02	-2.57	0.10
	PP	-2.89	0.05	-6.97	0.00	-3.43	0.01	-2.66	0.09
RPPI	ADF	-2.03	0.27	-3.75	0.01	-3.22	0.02	-2.96	0.04
	PP	-2.59	0.10	-3.70	0.01	-2.35	0.16	-2.41	0.14
Mortgage	ADF	-3.50	0.01	-3.53	0.01	-2.76	0.07	-2.23	0.20
	PP	-2.48	0.13	-6.66	0.00	-2.90	0.05	-2.49	0.12

ADF = Augmented Dickey Fuller, PP = Phillips Perron

Based on the stationarity tests, the data are stationer when de-trended using FD and MA4. Wilson and Okunev (1999) stated that using first difference for detrending is not appropriate for the study of cyclical behavior since it can eliminate low frequency (long cycles) of spectral density and amplify high frequency (short cycles). However, the detrended data using FD will be included in this study, since visually similar to detrended data using MA4 (see Appendix 1).

For every individual variable, spectral analysis is performed using parametric and non-parametric approaches. The parametric approach is performed using ARMA model, while the non-parametric approach is performed using Tukey-Hamming, Danielle, and Bartlett. Table 2 shows the results of spectral analysis. RPPI and mortgage are dominated by low frequencies, while the CPI: Housing is dominated by high frequency. RPPI and mortgages have cycle length around 15 – 16 years, while the CPI: Housing has cycle length around 4.5 years. All individual variables become the candidate of Property Cycle, and will be selected based on the event analysis of crises and the implementasion of LTV.

Table 2. Spectrum Peaks for Individual Variables

Data	MA4			FD		
	Frequency	Cycle Length		Frequency	Cycle Length	
RPPI	Low	63 quarters	16 years	Low	62 quarters	15 years
Mortgage	Low	61 quarters	15 years	Low	62 quarters	15 years
CPI:Housing	High	18 quarters	4.5 years	High	12 quarters	3 years

Table 3 shows the cross spectral analysis results. RPPI and CPI:Housing are coherence and similar in phase, where the best result is when data are detrended using first difference. Both variables are coherent around 62% on the main peaks in the period of 8 years. In addition the phase difference between the two variables is not too large, which is around 7.62 degrees. Similar results are shown by the common cycle between IHPR and CPI: Housing detrended using MA4. However, the main peaks are in shorter co-cycle length, which is around 3 years, and the phase difference is larger, which is 12.72 degrees.

Based on the results of cross-spectral analysis, the common cycle candidates are IHPR and CPI: Housing, de-trended using FD and MA4.

Table 3. Cross Spectrum Peaks

Data		Co-cycle Length		Proportion of covariance at main peak (%)	Coherence		Phase Difference	
		Quarters	Years		Main Peak	Total Average	Radians	Degrees
FD	IHPR & KPR	30.5	8	20.9	0.20	0.18	0.14	7.94
	IHPR & CPI:Housing	30.5	8	24.9	0.62	0.29	0.13	7.62
	KPR & CPI	61.0	15	10.2	0.26	0.26	0.50	28.84
MA4	IHPR & KPR	62.0	16	19.7	0.19	0.33	-3.01	-172.52
	IHPR & CPI:Housing	12.4	3	13.7	0.52	0.33	0.22	12.72
	KPR & CPI	61.0	15	13.7	0.20	0.21	0.43	24.40

Based on the results of the spectral analysis for each individual variable and the cross-spectral analysis, the property cycle will be constructed based on the dominant frequency. Appendix 2 shows the list of property cycle candidates.

The event analysis used to evaluate the property cycle are crises and the implementation of LTV policy. Crises that occurred in Indonesia in the range of data used in this study are the mini crisis (2005Q3 - 2006Q1) and the Global Financial Crisis (2008Q4 - 2009Q4). LTV has been implemented in Indonesia since 2012Q2, and subsequently was adjusted three times, namely in 2013Q3, 2015Q2 and 2016Q3. The first and the second LTV are tightening policy, while the third and the forth are loosening policy.

Based on the event analysis, the most proper cycle is the common cycle between RPPI and CPI:Housing detrended by first difference, and filtered in frequency 30 – 31 (RPPI_CPI_FD_30_31). It can be a leading indicator for the GFC, describes LTV implementation, and describes the peaks on the yearly growth of RPPI, as one of the references of LTV policy. The first peak in the cycle can be a leading indicator for the GFC, since it occurs 3 years before (chart 1, second crises bar). Therefore, this cycle has common characteristic of property price as leading indicator. Related to the second LTV implementation (tightening policy), it occurs before the second peak (chart 1, second LTV line). It means that the cycle can show the decreasing in property price after the LTV implementation. Furthermore, compare to the yearly growth of RPPI, the selected cycle can describe the peak in 2004Q1 and 2013Q3, even though with lags (chart 2).

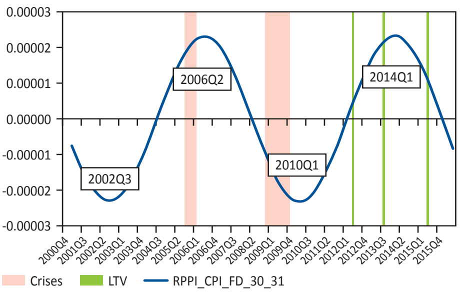


Chart 1. Common Cycle of RPPI and CPI:Housing, detrended by FD

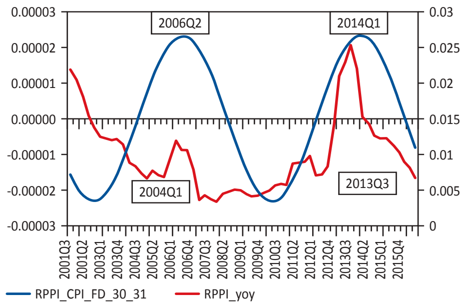


Chart 2. Common Cycle of RPPI and CPI:Housing, Compared to RPPI (YoY)

5. CONCLUSION

Based on this study, the length of the property cycle of Indonesia is 30.5 quarters or around 8 years. This cycle is constructed from a common cycle of RPPI and CPI: Housing which are detrended using first different. Both variables are coherent around 62% on the main peak and have small phase difference, around 7.62 degrees. In addition, the cycle can be a leading indicator of the GFC, describes LTV implementation, and describes the peaks on the yearly growth of RPPI.

REFERENCES

Drehmann, M., Borio, C., & Tsatsaronis, K. (2012). Characterising the financial cycle: don't lose sight of the medium term!. BIS Working Papers No. 380.

Gonzalez, R. B., Lima, J., & Marinho, L. (2015). Business and Financial Cycles: an estimation of cycles' length focusing on Macroprudential Policy. Banco Central do Brasil.

Gray, D. (2015). Are price of New dwelling different? A spectral analysis of UK property vintages. Congent Economics and Finance.

Grover, R., & Grover, C. (2013). Property Cycles. Journal of Property Investment and Finance, pp. 502-516.

Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.

Liow, K. H. (2007). Cycles and Common Cycles in Real Estate Markets. International Journal of Managerial Finance, pp. 287 - 305.

Man, K., & Chau, K. (2008). Hong Kong Property Cycle – a frequency domain analysis.

Wang, P. (2003). A Frequency Domain Analysis of Common Cycles in Property and Related Sectors. JRER Vol.25 no.3.

Warner, R. M. (1998). Spectral Analysis of Time Series Data.

Wilson, P., & Okunev, J. (1999). Spectral Analysis of Real Estate and Financial Assets Markets. Journal of Property Investment and Finance, 17(1), pp. 61-74.

APPENDIX 1 – DETRENDED DATA

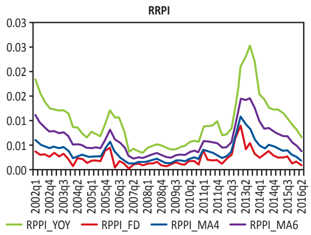


Chart 1.1. RPPI

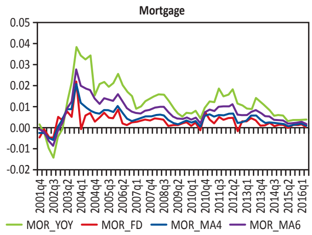


Chart 1.2. Mortgage

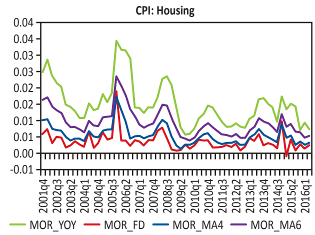


Chart 1.3. CPI:Housing

APPENDIX 2 - THE CANDIDATES OF PROPERTY CYCLE BASED ON MAIN PEAK

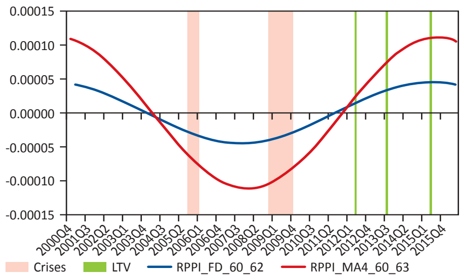


Chart 2.1. RPPI main cycle

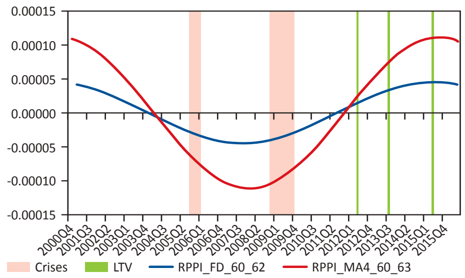


Chart 2.2. Mortgage main cycle

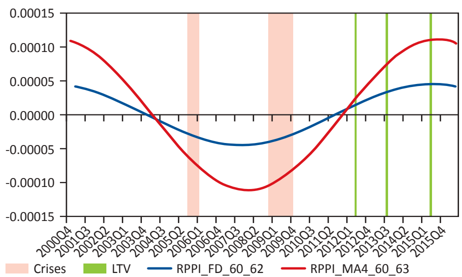


Chart 2.3. CPI:Housing main cycle

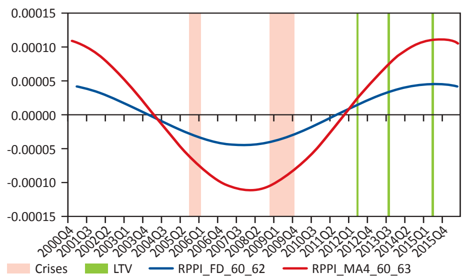


Chart 2.4. Common main cycle of RPPI and CPI, detrended by MA4

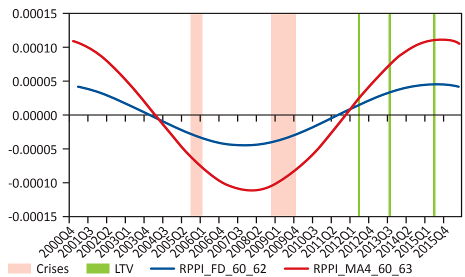


Chart 2.5. Common main cycle of RPPI and CPI, detrended by first different

A Hedonic Residential Property Price Index: The Case of Indonesia (A Preliminary Study)

Herina Prasnawaty Dewayany
Bank Indonesia, Jakarta, Indonesia– herinapd@bi.go.id

Listyowati Puji Lestari*
Bank Indonesia, Jakarta, Indonesia– listyowati_p@bi.go.id

Ahmad Rasyid
Bank Indonesia, Jakarta, Indonesia– ahmad_rasyid@bi.go.id

Abstract

Considering the important role of Residential Property Price Index (RPPI) in the economic policy making process, adequate techniques to measure RPPI are essential. Based on several studies, the measures using the chain-index methods were outdated. In order to improve international comparability of the Indonesia's property price indices, this paper concern on analysing the RPPI construction using Hedonic Regression Method, namely the Time Dummy Variable and Imputation Method using the limited data available. It finds that the existing chain index method does not significantly different from the hedonic – imputation method. In addition, according to the correlation analysis between the resulting indices and some economic indicators shows that the indices coming from all methods could well describe the residential property price in Indonesia. One of the hedonic method advantage is that it could explain the share of individual house characteristic on the price level. Both hedonic models estimate that the effect of building size on house price surpass the effect of lot size. An implication of the results is that these two methods could be complementary with each other, especially chain index method and hedonic – imputation method.

Keywords: RPPI; Chain index; Hedonic method; Imputation; Time dummy.

1. INTRODUCTION

As monetary authority, Bank Indonesia's single objective is to maintain the stability of inflation and currency exchange rates (Inflation Targeting Framework – ITF). One of the monetary policy transmission channel is through the asset prices, such as property prices. Bank Indonesia conducts Residential Property Price Survey (RPPS) in order to monitor the property prices. RPPS held quarterly in the primary market for newly built houses, starting from the first quarter of 1999 and currently covers 16 major cities in Indonesia. The result of RPPS, especially Residential Property Price Index (RPPI) is used for monetary policy formulation, such as macroeconomic indicator of inflation, financial stability or soundness indicator, and a measurement of asset price.

RPPI measures the price changes of residential property (especially houses) and plays some important roles in the decision-making process of economic policy-making. First, RPPI is an indicators of expenditure which used to calculate the Consumer Price Index (CPI). Second, RPPI is one of the macroeconomic indicators, where property prices increase are often linked to economic growth and indicates revenue increase of the economic actors involved in the property market as well as gives a positive impact on national income through tax revenues. In addition, home prices increase affects the behaviour of households on their financial planning, such as savings and investment (Eurostat, 2013). Third, RPPI is a measure of citizen capacity to participate and contribute to economic growth. Fourth, monitoring the residential property prices is considered important, especially in times of economic turbulence. These are some reasons RPPI needs to be evaluate, especially in terms of methodology used to generate the index.

Considering the important role of RPPI in the economic policy making process, adequate techniques to measure RPPI are essential. Otherwise, decision based on poorly constructed indices can be biased. On the other hand, Indonesia as a member of the G-20 is required to improve data quality of property indicators, as stated in the 19th Recommendation, concerning the provision of RPPI data for dissemination on the Bank for International Settlements (BIS) website (IMF and FSB, 2015).

The existing RPPI – published by Bank Indonesia – is a transaction based index and utilize the chain index methods, which almost the same as the stratification method. Based on several studies, the measures using the chain-index methods were outdated and there are some other alternatives of methods used by other country. In order to improve international comparability of the Indonesia's property price indices, this paper concern on analysing the RPPI construction using Hedonic Regression Method. There are several alternatives for calculating hedonic price indices, namely time dummy variable method and imputation method. This paper will explore the possibility to generate Indonesia's property price indices by utilizing the Hedonic Regression Method and limited data available from RPPS.

Given several methodologies recommended by the Eurostat, it will be instructive to investigate the differences in the outcomes as other methodology used to construct the RPPI, specifically the Hedonic Regression Method – Time Dummy Variable Method and Imputation Method. Alternative indices are generated using the existing datasets that comprises transactions data for private residential properties in Indonesia. The resulting indices are then compared with the existing quarterly RPPI and other indicators.

The rest of the paper is organized as follows. The next section provides a concise literature review of methodologies commonly adopted for constructing house price indices. Our focus is on the methodology recommend by Eurostat in the Handbook on Residential Property Price Indices (RPPIs), namely Hedonic Regression Method. Pursuant to the literature review, the following section discusses the possible methodology adopted. The result of theoretical model of hedonic price indices are described. Details of the data sources are discussed and in addition the computed indexes are reported and compared to the indices constructed using the existing methodology and other economic indicators. The next section concludes.

2. LITERATURE REVIEW

2.1. Hedonic Regression Methods

There are some alternatives of the methodology commonly used by other countries to generate RPPI, namely Hedonic Regression Method, Stratification Method, Repeat Sales Method and Appraisal-Based Method. This paper analyse the RPPI using Hedonic Regression Method since this method has some advantages which could complete the existing method (Chain Index Method). First, the Chain Index method could provide a price index but not information on the value of individual house characteristics on price level (Eichholtz, 1996). Second, the Hedonic Regression method could be adjusted for non-quality characteristics, while the other method could not accommodate those characteristics. Third, the Hedonic Regression method also considered as the most efficient methods for making use of the available data (Eurostat, 2013).

The Hedonic Regression Method recognizes that heterogenous goods can be described by their attributes or characteristics. That is, a good is essentially a bundle of characteristics. In the housing context, this bundle may contain attributes of both structure and the location of the properties which contributes to the price of the properties. There are some purposes of the Hedonic Method. First, to obtain estimates of the willingness to pay for, or marginal cost of producing the different characteristics. Second, to construct the quality-adjusted price indices.

2.1.1. Time Dummy Variable Method (TDVM)

One advantage of this approach is its simplicity, since the price index follows immediately from the estimated time dummy regression equation. This index measures the effect of "time" on the logarithm of price, specifically house price changes between the base period "0" and each comparison period "t".

Run a pooling cross-section data need to assume a fixed characteristics parameters, then we choose to estimate the model separately. In addition, pooled cross-section data estimation will have a problem with the change of the characteristics coefficient due to the additional of new sample data. Hence, the pooled cross-section or multiperiod time dummy method appears to be of limited use for the production of official RPPI due to the revision problem.

One way to deal with the problem of revision and overcome the flawed assumption of fixed parameters faced by the pooled cross-section time dummy method would be to estimate time dummy indices separately. First, estimate time dummy indices for adjacent period *t-1* and *t* and then multiply them to obtain a time series which is free of revision. Second, estimate time dummy indices between base period and the last period.

2.1.2. Imputation Method

Since the period t prices of properties sold in base period cannot be observed and are "missing" because a greater part of those properties will not be resold in period *t* and vise versa, to apply standard index number formula e these "missing prices" must be imputed (Hill and Melser, 2008; Hill, 2011). Hedonic imputation indices do this by using predicted prices, evaluated at fixed characteristics, based on the hedonic regression for all time periods.

3. EXISTING INDEX CONSTRUCTION AND TWO ESTIMATION MODELS

To demonstrate the existing index construction in Indonesia, this section analyses the construction of Indonesia's RPPI based on Chain-Index method. Two alternative index construction models using the Hedonic Regression Methods are then presented that are based on the time dummy variable method and imputation method discussed earlier.

3.1. Existing Chain-Index Method

The existing housing price index (RPPI) in Indonesia which is based on sales prices obtained from Residential Property Price Survey (RPPS) held quarterly. More specifically, the indices were determined as follows :

1. Transactions are first grouped by property type and provinces. The property type grouping is selected on the bases of the building width (sq m).
2. The missing data is imputed with replacement values and treating these as if they were observed (last observation carried forward).
3. The average (mean) quarter to quarter (qtq) price changes for every type of house is computed.
4. Then, a price index for each property type on a province basis is constructed. The overall price index for every province constructed from the average of each property type.
5. The national price index is essentially a weighted average of the price created for each province. The weight assigned to each province is the value of the cost of living (obtained from a survey).

3.2. Alternative Index Computation Models

Given the existing RPPI construction using Chain-Index Method, we attempt to construct other indices using the hedonic approaches as alternatives to the existing approach. However, we present only two sets of indices constructed from the time dummy variable method and imputation method. To a large extent, our methodologies choices are constrained by the dataset available from the RPPS conducted quarterly by Bank Indonesia. The main shortcoming of the data collected is the lack of potential explanatory characteristics for hedonic price index construction. Therefore our hedonic models are parsimonious and quite simple. However, such a reduced rank hedonic may still be reasonably reliable for price index construction (Melpezzi et al., 1998). Another shortcoming is the relatively short period over which micro-data on transaction are available. The database dates back to the fourth quarter of 2006 only. Prior data is not reliable.

Taking into account the data limitations, the time dummy variable hedonic equation becomes :

$$lnp_n^t = \beta_0 + \delta^t D_n^t + \beta\ G\ [ln\ x_n^t] + \varepsilon_n^t$$

Where D_n^t is dummy variable of time which is equal to 1 for each quarterly observation period t and 0 for base year, $exp\ \delta$ captures the changing mean in housing prices compare to the base year, β denotes a vector of parameters, G is a function of the attributes, and ε is the error term. The index n is runs from one to the total number of home sales in a given quarter t, where t runs from 2006:4 to 2016:3. The x(.) variable in the sample uses the lot size and the size of building measured in square meters. The hedonic equation is estimated for each time period separately.

The next hedonic regression approach is the hedonic imputation method which uses the implicit prices of the characteristics of the model (the regression coefficients) as the basis for constructing the price index. The imputation hedonic model is:

$$lnp_n^t = \beta_0 + \beta\ G\ [ln\ x_n^t] + \varepsilon_n^t$$

The next step is to compute a hedonic price index from the regression results by utilizing the Laspeyres, Paasche, and Fisher formula.

4. THE RESULTS

Since the existing RPPI uses 2002 as base year but the available database only began in 2006, it was necessary to compute RPPI indices with the fourth quarter of 2006 as the base period. In other words, the existing RPPI are computed by duplicating the chain-index technique of index construction. While the hedonic models runs on about 4000 observations in 2006 and growing to about 9700 observation in 2016. The result calculation using chain index and estimations based on both hedonic models can be seen in Figure 1 and Table 1.

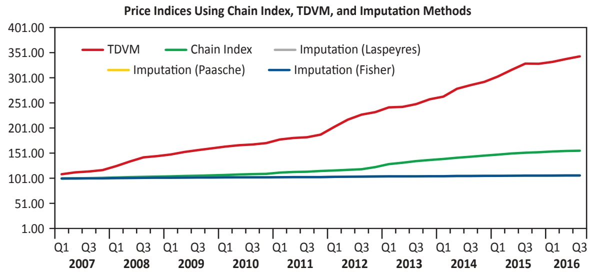


Figure 1. The Price Index Using Chain Index and Hedonic Models (Q42006=100)

Table 1. The Price Index Using Chain Index and Hedonic Models (Q42006=100)

Periods	Index				
	Existing (Chain Index)	TDVM	Imputation		
			Laspeyres	Paasche	Fisher
Q4 2006	100.00	100.00	100.00	100.00	100.00
Q1 2007	100.25	109.05	100.46	100.46	100.46
Q2 2007	100.93	112.78	100.65	100.65	100.65
Q3 2007	101.25	114.57	100.73	100.74	100.73
Q4 2007	101.74	117.32	100.86	100.86	100.86
Q1 2008	102.50	125.27	101.20	101.20	101.20
Q2 2008	103.36	134.29	101.57	101.57	101.57
Q3 2008	103.85	142.48	101.89	101.89	101.89
Q4 2008	104.35	145.16	102.00	101.99	102.00
Q1 2009	104.84	148.27	102.11	102.11	102.11
Q2 2009	105.56	153.27	102.28	102.29	102.29
Q3 2009	106.03	156.95	102.41	102.42	102.41
Q4 2009	106.76	160.52	102.53	102.54	102.53

Table 1. The Price Index Using Chain Index and Hedonic Models (Q42006=100) Lanjutan

Periods	Index				
	Existing (Chain Index)	TDVM	Imputation		
			Laspeyres	Paasche	Fisher
Q1 2010	107.50	163.91	102.64	102.67	102.66
Q2 2010	108.61	166.68	102.73	102.76	102.75
Q3 2010	109.13	168.21	102.78	102.81	102.79
Q4 2010	109.86	170.86	102.86	102.89	102.87
Q1 2011	112.32	178.05	103.07	103.11	103.09
Q2 2011	113.54	181.24	103.16	103.20	103.18
Q3 2011	114.09	182.72	103.21	103.24	103.22
Q4 2011	115.40	187.85	103.36	103.40	103.38
Q1 2012	116.35	203.20	103.79	103.74	103.76
Q2 2012	117.71	217.57	104.15	104.08	104.11
Q3 2012	118.93	227.55	104.39	104.30	104.34
Q4 2012	123.21	232.72	104.51	104.41	104.46
Q1 2013	129.15	242.10	104.71	104.60	104.66
Q2 2013	131.97	243.01	104.73	104.62	104.68
Q3 2013	135.00	248.64	104.85	104.75	104.80
Q4 2013	137.39	258.16	105.05	104.94	105.00
Q1 2014	139.38	263.52	105.16	105.08	105.12
Q2 2014	141.74	279.22	105.47	105.37	105.42
Q3 2014	143.81	286.22	105.60	105.50	105.55
Q4 2014	146.02	292.68	105.72	105.63	105.67
Q1 2015	148.12	303.49	105.91	105.82	105.87
Q2 2015	150.17	316.61	106.13	106.05	106.09
Q3 2015	151.65	329.02	106.33	106.25	106.29
Q4 2015	152.76	328.88	106.34	106.26	106.30
Q1 2016	154.27	332.53	106.39	106.32	106.36
Q2 2016	155.26	338.38	106.48	106.41	106.45
Q3 2016	155.82	343.55	106.57	106.49	106.53

Generally, the result of imputation methods almost the same for Laspeyres, Paasche, and Fisher formula. In theory, Laspeyres index is known to yield measurement biases by tending to overstate the rise of capital prices by not allowing any substitution between goods to occur (Diewert, 1998). Conversely, a Paasche index tends to understate the rise in the price. An implication of this is that some average of the Laspeyres and Paasche price indices should provide a reasonably good approximation to the true price changes. Such an average would be represented by the Fisher index, which is the geometric average of Laspeyres and Paasche indices (Meese and Wallace, 1991). Therefore, the Fisher formula is taken as the best approximation to a theoretical true price index and will be used in further discussion on this paper.

Furthermore, the one-way ANOVA test result indicates that there is a statistically significant difference between the mean index of three different methods outputs. The Tuckey post hoc test was utilized to determine which groups differed from each other. The result indicates that there is no statistically significant difference of index between the imputation method and the chain index method ($p = 0.160$), but there is statistically significant difference to the time dummy variable method. According to the result, the time dummy variable method seems to overstate the price movement. This result might occur since the hedonic models equation are quite simple. We can only include two characteristics in the model because there are only those two variables available from the RPPS.

The next step would be comparing the resulting indices with other indicators, namely the Gross Domestic Product (GDP) – real estate sector, housing loans disbursement, and property stock price index. The analysis would utilize the correlation analysis by estimating the correlation coefficient between the resulting index of each method and the control indicators. The result could be seen in Table 2.

Table 2. Correlation between Resulting Indices and Economics Indicators

Pearson Correlation Coefficient		Existing (Chain Index)	TDVM	Imputation (Fisher)
GDP – Real Estate Sector	Corr. coef	0.85	0.87	0.91
	p-value	0.00	0.00	0.00
Housing Loans Disbursement	Corr. coef	0.99	0.99	0.98
	p-value	0.00	0.00	0.00
Property Stock Price Index	Corr. coef	0.95	0.93	0.89
	p-value	0.00	0.00	0.00

According to the above results, all indices having positive and strong correlation to all indicators, indicated by the values of Pearson correlation coefficient ($r > 0.5$). Furthermore, the level of statistical significance (*p*-value) of the correlation coefficients in above table are 0.00, which means that there are statistically significant relationships between indices and all of the indicators. Therefore indices coming from all three methods could describe the residential property price in Indonesia.

However, taking into account the small differences among correlation coefficient values, we can say that the indices from the chain index and imputation method have the highest correlation to all indicators, while the indices from time dummy variable method only excel in one indicator. Therefore indices coming from chain index and imputation methods better describe the residential property price in Indonesia compare to indices coming from the time dummy variable method. Since the result of one-way ANOVA test indicates that there is no significant difference between indices coming from chain index and imputation method, then these two indices could be complementary each other.

In addition, one of the hedonic methods advantage is the ability to explain the share of individual house characteristic on the price level. Both hedonic models estimate that the effect of building size on house price surpass the effect of lot size. The results of estimation for all periods describes that an extra percentage of lot size and building size will increase the price of house by 0.15% and 1,18% respectively, ceteris paribus (all variables other than the observed variables are assumed to be constant). This information could be used as a complement to the existing chain index method.

Both existing chain index and hedonic regression method could well explain the residential property prices. But the constructed hedonic models are parsimonious and quite simple due to limited data available from the RPPS. These data limitations make the hedonic regression method could not accommodate the effect of characteristics other than lot size and building size. Currently, Bank Indonesia is making an enhancement on the RPPS by adding some variables, such as number of floors, bedrooms, bathrooms, garages, swimming pools, and so on. This enhancement is starting from the first quarter of 2016, in order to gain more characteristics for further hedonic regression analysis in the future.

6. CONCLUSIONS

This paper was motivated by the urge to improve data quality of property prices indicators by examining other methods besides existing chain index method. The option goes to hedonic methods, namely imputation and time dummy variable methods. The one-way ANOVA test result indicates that there is no statistically significant difference of index between the imputation method and the existing chain index method, but there is statistically significant difference with the time dummy variable method. Meanwhile, the correlation analysis result shows that all indices having positive, strong, and significant correlation to all control indicators. In addition, one of the hedonic methods advantage is the ability to explain the share of individual house characteristic on the price level. Both hedonic models estimate that the effect of building size on house price surpass the effect of lot size and an extra percentage of lot size and building size will increase the price of house by 0.15% and 1,18% respectively, ceteris paribus. Therefore the Hedonic – imputation method result could complement the result of the existing chain index method.

REFERENCES

Diewert, W.E. (1998), "Index number issues in the consumer price index", *Journal of Economic Perspectives*, Vol. 12, pp. 47-58.

Eurostat (2013), Handbook on Residential Property Prices Indices (RPPIs). Luxembourg: Publications Office of the European Union.

Eichholtz, P.M.A. (1996), "A long run house price index: The Herengracht index, 1628-1973", *Real Estate Economics*, Vol. 25 No. 2, pp. 175-92.

Hill, R.J. and D. Melser (2008), "Hedonic Imputation the Price Index Problem: An Application to Housing", *Economic Inquiry,* Vol. 46 No.4, 593-609.

Hill, R.J. (2011), "Hedonic Price Indexes for Housing", OECD Statistics Working Paper 2011/1, OECD Publishing, Paris.

IMF dan FSB (2015), Suplementary Annex G-20 Economies' Status vis-a-vis Implementation Targets of the DGRI Recommendations.

Malpezzi, S. Chun, G.H. and Green, R.K. (1998), New place-to-place housing price indexes for US metropolitan areas, and their determinants", *Real Estate Economics*, Vol. 26 No. 2, pp. 235-74.

Meese, R. and Wallace, N. (1991), "Nonparametric estimation of dymanic hedonic price models and the construction of residential housing price indices", *Journal of the American Real Estate and Urban Economics Association*, Vol. 19 No. 3, pp. 308-32.

Statistics Estonia (2008), *Frequently Asked Questions about Chain-linking Method,* Statistics Estonia.

This page intentionally left blank

CPS05:
Probability Theory & Statistical Modelling (1)

A Naïve Method for Variable Reduction in Multidimensional Data

Noppachai Wongsai[1],*
[1]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Muang, Pattani, 94000 Thailand – naong69@gmail.com

Sangdao Wongsai[1]
[1]Faculty of Technology and Environment, Prince of Songkla University, Phuket Campus, Kathu, Phuket, 83120 Thailand – sangdao.w@phuket.psu.ac.th

Apiradee Lim[1]
[1]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Muang, Pattani, 94000 Thailand – api_45@hotmail.com

Abstract

In correlated multidimensional data modelling, a variable reduction technique such as Exploratory Factor Analysis (EFA) is common used. EFA requires users to specify the number of common factors in advance, and also to specify a rotation method. This preferred statistical method for factor analysis involves maximizing the likelihood of the correlation matrix of the original data structure (Venables & Ripley, 2002). It assumes that these variables are normally distributed and is not yet fully understood or accepted by many scientists. We propose a "naïve" method that creates common factors simply by grouping variables having average correlation coefficient greater than specified values. This proposed method does not only provide the simple and straightforward procedure to decrease the dimension of data but also can be applied to data where the statistically-approved method fails because the correlation matrix is singular, such as data where the number of variables exceeds the number of records in a data table.

Keywords: Multidimensional data, Common factor analysis, Correlation coefficients
JEL Classification: C38

1. INTRODUCTION

Information overload in most sciences continuously occurs during the past decades due to the advances technologies in data collection and storage capabilities. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation (Fodor, 2002). There are many dimensionality reduction techniques. The most widely used in Statistics is Principal Component Analysis (PCA). PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations of the original variables with the largest variance (Fodor, 2002). Another linear method is Factor Analysis (FA) which entirely dependent on correlations between variables. FA is a multivariate analysis procedure that operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are not directly measured, which is known as reducing dimensionality (Bartholomew *et al.*, 2011). The main idea is to look for the sets of variables in which all the variables in one group correlate strongly with one another and weakly with all of the variables in the other sets (Norman & Streiner, 2003).

There is two kinds of FA that focus on different applications. Confirmatory factor analysis (CFA) that based on *path analysis* and *structural equation modeling* technique is used to confirm the hypotheses. And exploratory factor analysis (EFA) is used to indicate patterns within the data for data exploration. EFA is used when a researcher wants to discover the number of common factors influencing variables and to analyze which variables go together (Yong & Pearce, 2013). To find the common factor, *principle factor analysis*, which is obtained the same way as PCA, is the simplest and computationally most efficient method. *Maximum-likelihood factor analysis* is another one that most statisticians prefer because its produces better estimates than principle factor analysis in the large samples (Hox & Bechger, 1998). Targeting to reduce dimensions of the large and complex dataset using FA involves many methods and procedures. As well as, it requires mathematical and statistical understanding of the underlying theoretical background to select the right methods and the number of factors to retain in *factor extraction* and *rotation method* steps. And also the problematic issues such as *bipolar factor loading*, *factorial complexity*, or *factor indeterminacy* (Grice, 2001) are needed to be concerned during *interpretations of factor loadings* step. Without the comprehend knowledge of those matters; it is difficult to generate the reliable and responsible results, even encounter with unable to produce one at all. The cause that makes this statically factor analysis to be nearly impossible to produce the result is singular correlation matrix, which requires in the factor extraction step, as well as the violation of multivariate normality distributed and random error assumptions.

In this paper, we present the simple method based on only the correlation matrix of all variables, and one criterion value that be able to achieve the similar result as the preferred statistical method for factor analysis can produce. The method and its application on ecology data are detailed in Section 2 and 3 respectively. Section 4 and 5 present discussion and concussion of this work.

2. THE NAÏVE METHOD

The proposed "naïve" method totally depends on the correlation coefficient of a pair of variables. It will be used to form a group of variables as a new factor that has the same strange of association among them. The algorithm of this method involves creating factors simply by find the most highly correlated variable pair as the basis for the factor forming and sequentially attaching variables that are highly correlated with these variables using a criterion (or cut-off) based on the sum of these correlation coefficients. Once this criterion fails to find any new variable to attach, a new factor is created from the remaining variables. The process recursively carried on until all variables have been dealt with.

The cut-off algorithm requires the matrix contain correlation coefficient of the relationships between variables or observed data values which called *correlation matrix*. Suppose P is a number of variables that measured for N subjects (or observations). The correlation matrix is defined as

$$R_{(p\times p)} = \frac{V_{(p\times p)}}{(N-1)}$$

where and $V_{(p\times p)}$ is the variance-covariance matrix that shows the variance within each column and the covariance between columns of deviation scores matrix. Thus

$$V_{(p\times p)} = D'_{(p\times N)} D_{(N\times p)}$$

in whish $D_{(N\times p)}$ is the data matrix of deviation scores with N rows and P columns that transformed from the raw data. In this proposed method, the squared correlation matrix R will satisfy one of the algorithm requirements. The matrix R is in the form:

$$R = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,j} \\ r_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & r_{i-1,j} \\ r_{i,1} & \cdots & r_{i,j-1} & 1 \end{bmatrix} \quad \text{where} \qquad i,j \in \{1,\dots,p\}$$

Note that this is the mirror matrix of the correlation coefficient elements with 1 on the diagonal. Another requirement is the *cut-off criterion* (σ) or *cut-off value*, which is the only parameter needed to be provided. Since the triangular matrix of R, named R_t, can be defined in the form as following:

$$Rt := R \begin{cases} r_{i,j} = 0\ for\ i < j \\ \qquad or \\ r_{i,j} = 0\ for\ i > j \end{cases} \quad \text{where} \quad i,j \in \{1,\dots,p\}$$

Thus, in the case of upper triangular matrix is used and given a list of all variables is a vector $V = [v_1 \dots v_p]$, the variables that belong to the factor (f_n) can be denoted as

$$V_{f_n} \subseteq V \text{ in which } V_{f_n} = V[v_k] \text{ where } \{k \in \forall i \mid Rt\left[\, r_{*,j} \,\right] \ni \max(Rt) \wedge R\left[\, r_{i,j} \,\right] \geq \sigma\}$$

Therefore the variables in the vector V at the position that match with the position of rows (i) in a single matrix at column (j) of matrix R_t which contains the highest correlation and have value of its elements greater than or equal the cut-off value are chosen to be assigned in the factor. Finally, the matrix R is reduced by removing rows ($Rt_{i,*}$) and column ($Rt_{*,j}$) at the selected correlation are found, which can be denoted as

$$R = R_{-m,-n} \quad \text{where} \quad \{m, n \in \forall i \mid Rt\left[\, r_{*,j} \,\right] \ni \max(Rt) \wedge R\left[\, r_{i,j} \,\right] \geq \sigma\}$$

As well as subtracting already allocated variables from the list of all variables V as

$$V = V - V_f \quad \text{where} \quad \{n \in \mathbb{Z} \mid n > 0\}$$

The process is repeated afterward, from determining the triangular matrix Rt from the remaining matrix R and the list of all variables V from the remaining matrix R row or column name, finding the maximum value of correlation coefficient in the remaining matrix R. Then allocating the remained variables into new factor until the maximum value in the matrix is less than the cut-off value. Subsequently all the variables that have been left over are the individual one that cannot be assigned to any factor, this could be called as the *unique variables*.

3. APPLICATION

In this section, we present an example using ecology data to illustrate the application of the method described in section 2. The dataset consists of the plankton abundance in the Na Thap River, Songkhla province, southern of Thailand. To demonstrate the potential of the proposed method, we compare the result of EFA method with the naïve method under R programming environment (R Development Core Team, 2011), which mainly aim to understand and reduce the structure of the data.

3.1. Plankton Abundance in The Na Thap River

The micro-organism densities data were collected from ten sites on the Na Thap River. It consists of 602 observations with bimonthly sampling intervals from June 2005 to October 2015. The Na Thap study has found 139 different plankton taxa, which comprises of 87 phytoplankton and 52 zooplankton. This kind of taxonomic data from studies in ecology is highly multidimensional data since there are, in principle, the relationships among those planktons. The interaction between them could be very complex and be specific in one ecosystem, such as the food web association in the rivers system that covers the freshwater and estuarine ecosystem.

Based on a previous inspection of the distribution of the residuals using the histogram and the Q-Q plot, we found that this organism count distributions are very *highly skewed* and contain a *high proportion of zeros* (more than one-third in this case). This kind of properties is usually found in most ecological microorganism counts. Thus, we consider appropriate to use a log transformation, such as log(1+*density*), for this space-time data set. Moreover, because of the distinct properties of the data, it is useful to model prevalence and abundance (or density) separately, even though these ecological metrics are often related. Thus, the subset of plankton density data now contains 106 commonly occurring plankton taxa, which exclude 33 taxa with average values of log-transformed densities less than 0.19. In this example, we focus on the abundance and more concern on the correlation between plankton taxa densities.

The statistical method, when considering using the 'factanal' function, in R's native package, with "promax" rotation and specified the number of factors as two, allocates variables to factors when loadings exceed 0.2 as shown in Figure 1. Note that the variables in each factor are descending ordered by factor loading and the variables that are not assigned to either factor are ascending ordered by uniquenesses. These factor groups have 45 and 40 plankton taxa, respectively, with 21 in a unique group omitted from the model. There are 29 variables give negative factor loading, majority assigned to Factor 1.

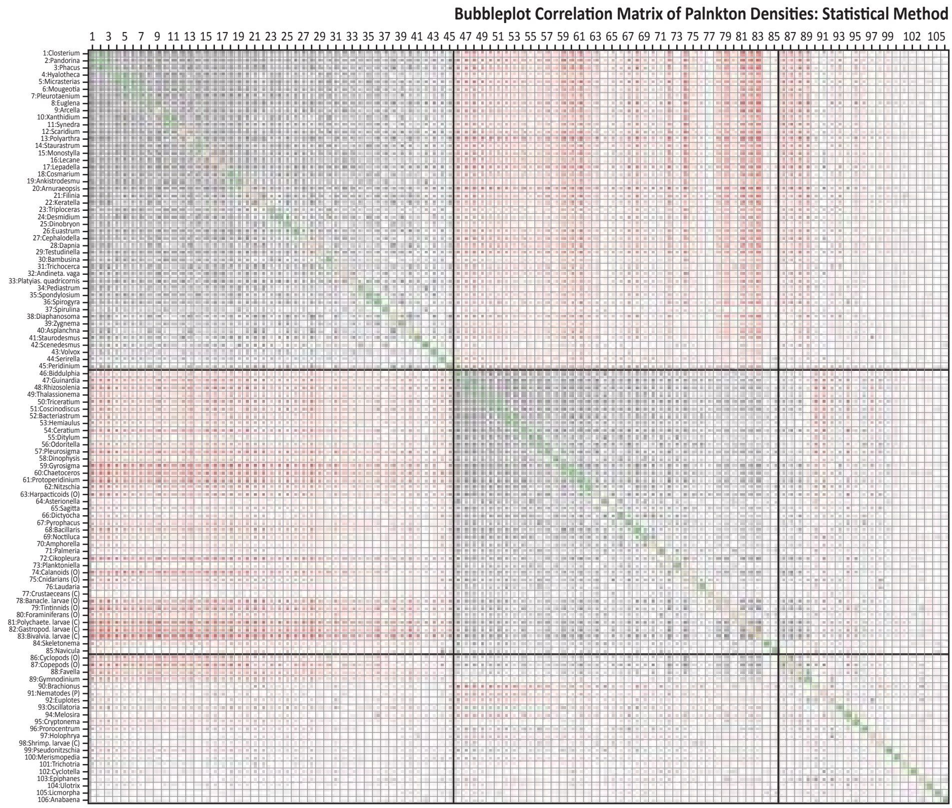


Figure 1: The bubble-plot correlation matrix of 106 variables allocated into 2 Factors using the Factor Analysis method (red color denote negative correlation)

There are several taxa that considering have high negative factor loading which are *Bivalvia Larvae*, *Polychaete Larvae*, *Tintinnids*, and *Gastropod Larvae* (–0.593, -0.501, -0.469, and -0.447, respectively). Those taxa densities are highly negatively correlated with other variables in Factor 1 and have positive correlations with variables in Factor 2. However, they have a high negative loading with Factor 1. Moreover, *Oikopleura* has loading value -0.324 for Factor 1 and 0.356 for Factor 2, same as *Calanoids* that has loading value -0.368 and 0.325 for both factors, respectively. In general, high loadings suggest stronger factor contributions to those variables. Also, it is possible to have both positive and negative factor loadings, however having both positive and negative loadings that are large is rare. The signs of the loadings identify the direction of the correlation and do not affect the interpretation of the magnitude of the factor loading or the number of factors to retain (Kline, 1994).

Figure 2 shows the result from the naïve method. This method requires the average correlation between all variables in a group to exceed 0.2. It initially creates six groups having 44, 42, 3, 4, 2 plankton taxa for the first five groups, respectively, and the last group includes 13 taxa which their correlations less that the cut-off value. The average of correlations and the number of taxa assigned to group 3 to 6 are relatively small. Thus we combined all 22 taxa to become one group. The two methods produce quite a similar result with allocations of plankton taxa to three main groups. There are some differences as follows. The first group contains mostly the same taxa for each method. The naïve method allocates 44 taxa in this group whereas the statistical method allocates only one additional taxa, which is *Serirella*. The second group contains 40 taxa for each method. These groups have 34 taxa assigned by both methods. *Shrimp Larvae*, *Cyclopods,* and *Copepods* are allocated to this group by the naïve method, but not by the statistical method, whereas *Navicula*, *Planktoniella* and *Skeletonema* are allocated to this group by the statistical method but not by the naïve method. Moreover, the third group comprises remaining taxa that have smaller correlations with other taxa.

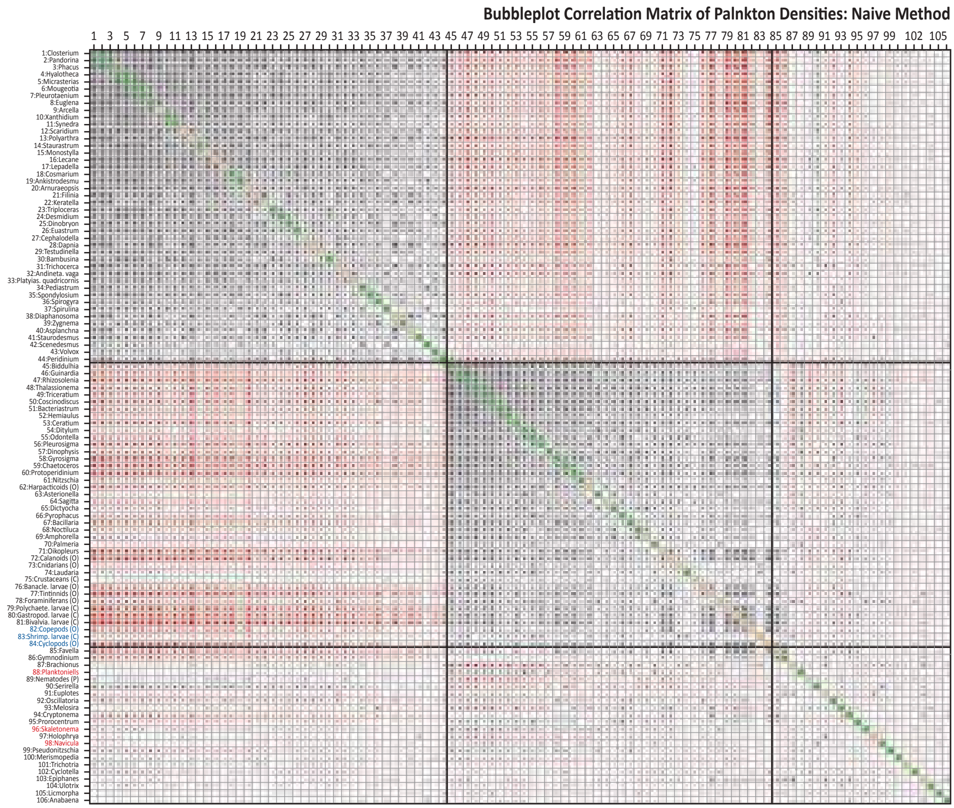


Figure 2: The bubble-plot correlation matrix of 106 variables allocated into 2 Factors using the naïve method (red color denotes negative correlation)

4. DISCUSSIONS

The proposed method uses triangular matrix (upper or lower) as the matrix to find the maximum value of correlation. It may produce some degree of different in the result. The method demonstrated in this work uses upper triangular matrix and concerns the column of the correlation matrix to identify a single matrix to find the variables that have a strong correlation with the highest correlation pair of variables. To remain consistent result, if the lower triangular matrix is used, the condition of allocating variables into the group must be as the following:

$$V_{f_n} \subseteq V \text{ in which } V_{f_n} = V[v_k] \text{ where } \{k \in \forall j \mid Rt[\,r_{i,*}\,] \ni \max(Rt) \wedge R[\,r_{i,j}\,] \geq \sigma\}$$

and this condition must be use to reduce the matrix *R* as well.

4.1. Correlation Type and Missing Values

As this proposed method requires an appropriate correlation matrix, the process of creating correlation matrix need to be sincerely concerned. The correlation coefficient type needs to be carefully chosen. If the data are not normally distributed data, *spearman* correlation type may be required, or log transformation of data may be considered before creating correlation matrix using *pearson* correlation type. In the presence of missing values in the data, *casewise* or *pairwise* deletion for computing covariance matrix needs to be cautiously selected. The *casewise* deletion of missing value removal exclude all cases that have missing data in at least one of the selected variables, whereas *pairwise* deletion calculates a covariance or correlation separately between each pair of variables from all case that has valid data on those two variables. In the case of using R program, it computes correlations for each pair of columns using vectors formed by omitting rows with missing values on a *pairwise* basis. Thus each column vector may vary depending on its pairing, resulting in correlation values that are not even comparable (Lewis, 2015). It may sometimes lead to serious problems in a situation of large matrices with many of missing values, where it may be impossible to use the *pairwise* option in a meaningful way. It is desirable that the pattern of missing data be random. Otherwise, the factor structure that computed will be influenced systematically by the pattern of how values are missing.

4.2. Factor Score Indeterminacy

Although factor analysis methods from different software packages can produce the desirable result, they require several parameters and options to be specified. Sometime it gives hard-to-explain results, such as in the examples. The issue of signs of loading within factors (or components) is arbitrary and indeterminate. Indeterminacy arises from the fact that, under the common factor model, the parameters are not uniquely defined, due to the researcher's choice of the communality estimate. As noted by EFA researchers, the problem of indeterminacy arises with most factor extraction techniques found under the common factor model (DiStefano *et al.*, 2009). In general practice before creating a factor score, if a factor has more negative than positive loadings, changing minus signs to plus and vice versa may be the solution in EFA. It would be appropriate to reverse-score the negative loading variables after variables standardization. If a variable has a negative factor loading, the raw score of the variable is subtracted rather than added in the computations because the variable is negatively related to the factor. However, by doing this and then using sum/average method for factor scores, it could lose all the benefits of latent variables or estimated factor scores. In many cases, researchers do not intend to find the factor score. One may only need a method to do the data structure reduction statistically. If the preferred statistical method produces unhelpful results, the proposed method may be the possible alternative one.

4.3. Singular Matrix

The main advantage of the naïve method not only is that producing the reasonable and explainable result, but also it can be applied to data where the statistically approved method fails because the correlation matrix is singular. Most of factor analysis extraction methods require non-singular matrix as the output of the process of analyzing correlation or covariance matrix. The main reason is, at the various stages of the analysis (preliminary, extraction, scoring), factor analysis algorithm addresses true inverse of the matrix or needs its determinant. As commonly known that ML factor analysis cannot be used with a singular correlation matrix, and it is especially subject to Heywood cases (Kolenikov & Bollen, 2012). However, *minimal residuals* method can work with singular matrix at the extraction stage. The singular correlation or covariance matrix can be presented when there are many non-randomly missing values in the dataset or when the number of columns is greater than the number of rows in the data table.

5. CONCLUSIONS

We developed a naïve method to simply create factors using only one criterion based on the sum of these correlation coefficients. Other criteria for including a variable in a factor may be used, such as one based on the *average* correlations instead of the *maximum*. When scientific data with highly complex and too many variables need structural reduction, researchers may find that the statistical method, such as EFA, could be problematic. In the situation that the EFA (1) results in too many or too few factors which happens all the time when extracting factor based on eigenvalues, (2) introduces factor crossloading (Costello & Osborne, 2005) that difficult to define statistically meaningful cut-off value, (3) generates loadings that greater than one in absolute value, (4) presents *mix factor* where some variables load onto more than one factor which known as split loadings (Yong & Pearce, 2013), and most importantly (5) produces little underlying meaning for the factor, the naïve method is worth considering. The proposed method demonstrates its potential to be an alternative technique when the preferred statistical method falls to give the interpretable result or simply cannot produces the result due to singular correlation matrix (Beavers *et al*., 2013). All together, the main advantages of our approach are the few parameter requirements and statistical assumptions, the data-driven choice of variables reduction based on its correlation structure, as well as fast and uncomplicated numerical implementation.

REFERENCES

Bartholomew, D., Knotts, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach. (3rd ed.). West Sussex, UK: John Wiley & Sons.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical Considerations for Using Exploratory Factor Analysis in Educational Research, *Practical Assessment, Research & Evaluation*, **18(6)**.

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, **10(7)**, 1–9.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, **14 (20)**.

Fodor, I. K. (2002). A survey of dimension reduction techniques. Report for U.S. Department of Energy, University of California. Available from https://e-reports-ext.llnl.gov/pdf/240921.pdf.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, **6(4)**, 430-450.

Hox, J.J., &Bechger, T.M., (1998). An introduction to structural equation modeling, *Family Science Review*, **11**, 354–373.

Kline, P. (1994). An easy guide to factor analysis. New York, NY: Routledge.Norman, G. R., & Streiner, D. L. (2003). PDQ statistics Third Edition, Shelton, CT:PMPH.

Kolenikov, S. &Bollen, K. A. (2012). Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?, *Sociological Methods & Research*, **41(1)**, 124–167.

Lewis, B. W. (2015). Pairwise-complete correlation considered dangerous. Access on Jan 10, 2017. http://bwlewis.github.io/covar/missing.html.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Venables, W. N., &Ripley, B. D. (2002). Modern Applied Statistics with S, New York: Springer-Verlag.

Yong, A. G., &Pearce S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis, *Tutorials in Quantitative Methods for Psychology*, **9(2)**, 79–94.

Flexible Functional Clustering of Accelerometer Data Using Transformed Input Variables

Yaeji Lim*
Department of Statistics, Pukyong National University, Busan, Korea – yaeji.lim@gmail.com

Hee-Seok Oh
Department of Statistics, Seoul National University, Seoul, Korea – heeseok.oh@gmail.com

Ying Kuen Cheung
Department of Biostatistics, Columbia University, New York, USA

Abstract

This paper considers a problem of constructing a grouping of accelerometer data that measure physical activity by computerized accelerometer. Classical methods such as K-means and PAM are not efficient in handling the accelerometer data are high-dimensional and have inherent multiscale structures. Some functional clustering approaches have a limitation of identifying the dynamic structures of the accelerometer data that might be necessary to cluster them success- fully. This paper introduces new input variables for clustering based on rank-based transform and thick-pen transform, which efficiently reflect specific structures of the data such as the amount or the pattern of physical activity with holding a functional form. The proposed clustering methods are obtained by coupling the transformed input variables with functional clustering that considers marginal contribution for building clustering criteria. Furthermore, for achieving a better clustering result that integrate both amount and pattern features of physical activities from 365 patients, various clustering schemes are developed. Empirical performance of the proposed methods is evaluated through simulation study, which demonstrates promising results of the proposed clustering.

Keywords: Accelerometer data; Functional data; High-dimensional data; Rank-based transform; Thick-pen transform.

1. INTRODUCTION

The goals of this paper are to provide groupings or clusters of physical activity patterns based on accelerometer data, and to further identify activity features of each group. For illustration purposes, we consider accelerometer data from 365 individuals.

In this paper, we propose new clustering methods that reflect the quantity and pattern of physical activity, via two important techniques, namely, transformation techniques and functional data clustering approach. Transformation is applied to construct input variables for clustering so that the inherent features of the data can be represented. Functional data analysis aims to handle the high dimensionality of the data and preserve the dynamic structure of the functional data.

2. CONSTRUCTION OF INPUT VARIABLES

2.1. Rank-based Transform

We consider a rank-based transform to generate input variables for clustering. Let $X_{(l)}$ be the l th smallest of a real-valued process $\{X(t)\}_{t=1}^{T}$, and $\{X_{(t)}\}_{t=1}^{T}$ be the order statistic of $\{X(t)\}_{t=1}^{T}$. By connecting or interpolating the data $\{t, X_{(t)}\}_{t=1}^{T}$, we obtain a functional form of the data, which can be used as functional input variable for amount of the activity.

In addition, we consider rank statistic of $\{X(t)\}_{t=1}^{T}$ as $\{r(X(t))\}_{t=1}^{T}$, where r is any function such that $r(X(t)) \leq r(X(s))$ whenever $X(t) \leq X(s)$. By transforming the data $\{X(t)\}_{t=1}^{T}$ into $\{r(X(t))\}_{t=1}^{T}$, it is feasible to identify the pattern of the activity. However, it is typically noisy; hence, it is necessary to denoise $\{r(X(t))\}_{t=1}^{T}$. Here we take a wavelet shrinkage procedure for $\{r(X(t))\}_{t=1}^{T}$, with Haar wavelets.

2.2. Thick-Pen Transform

Thick-pen transform (TPT) of Fryzlewicz and Oh (2011) is a multiscale visualization technique by connecting the dots with a range of pens. In this paper, we utilize TPT to construct input variables for clustering the accelerometer data.

3. PROPOSED CLUSTERING

With the transformed input variables in Section 2, we consider the concept of marginal representation of a curve for building a criterion of functional clustering, which has been used in Chiou and Li (2007). Let *X(t)* is the accelerometer record at time point t = 1, . . . , T . Let C be a random cluster variable that is randomly distributed among clusters {1,2,...,K} with probability mass $PC^{(c)}$ for c ∈ {1,2,...,K}. For criterion for functional clustering, we first define $\tilde{X}^{(c)}$ for a curve X as a difference between the curve X and the mean curve of members in cluster c, say $\mu^{(c)}$. Thus, in the case that the curve X belongs to cluster c, the magnitude $||\tilde{X}^{(c)}||$ might be small. We remark that in this study, $\tilde{X}^{(c)}(t)$ is some functional form of input variables introduced in Section 2. The detailed description of the functional form for each input variable will be discussed in the full version of the paper. Then, we consider the following optimization for clustering a random curve *X(t)* as

$$c^* = \arg min \left|\left|\tilde{X}^{(c)}\right|\right|.$$

4. CLUSTERING RESULTS OF ACCELEROMETER DATA

In this section, we discuss clustering results of the accelerometer data performed by the proposed methods based on transformed input variables. Here, we briefly show the results by Rank-based transform.

Figure 1 shows the clustering results based on the order statistic. In the first row of the figure 1, the mean curves of the accelerometer records in each group are displayed, and the mean curves of the functional input variables $X_{(l)}$ in each group are shown in the second row. We observe that the proposed method divides the data into three subgroups that might have different amounts of activity: 177 subjects are classified as a mild activity group, 129 subjects are in a moderate activity group, and 59 subjects belong to a severe activity group. From the second row, we observe that the three exponential functions are quite similar, but these are some difference in terms of maximum values at the right end of the functions. This difference is reflected in the above-proposed criterion and the corresponding results.

Figure 2 shows the clustering results based on the rank statistic. The plots in the first row are the mean curves of the raw data for each cluster, and in the second row, the average values of the *r(X(t))* in each cluster are plotted and the mean reconstructions by Haar wavelet shrinkage are shown by the red line, where the difference of activity patterns between subgroups is clearly identified. The subjects who active near time t = 0 belong to clusters 2 and 4, and individuals who are less active near time t = 0 and t = 1400 are classified in groups 1 and 3.

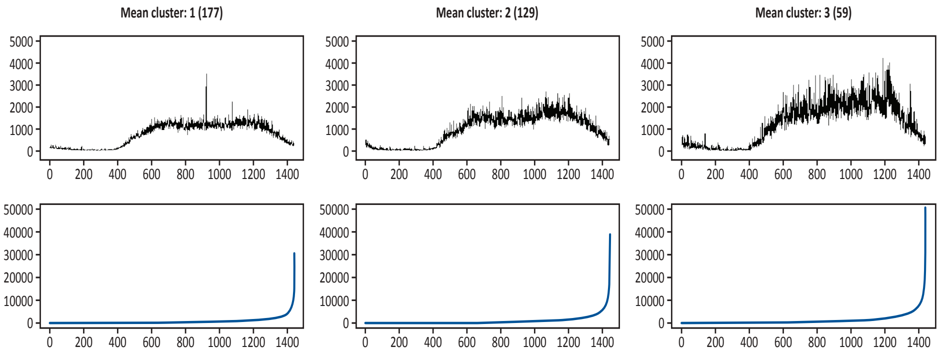


Figure 1: Clustering results based on the amount of the activity using ordered data. Three mean curves of the three groups (first row), and the corresponding three mean curves of the ordered data (second row)

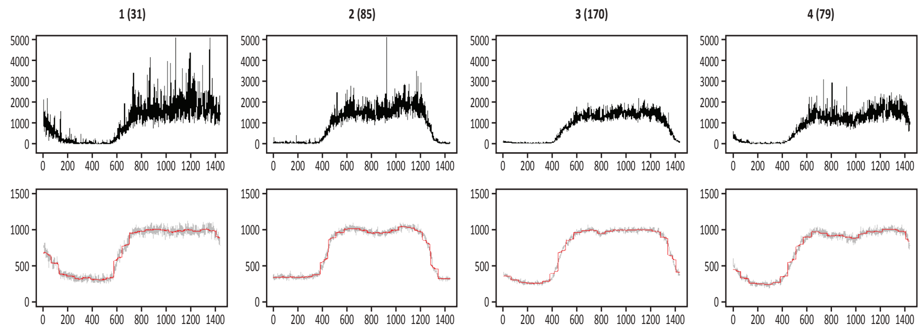


Figure 2: Clustering results based on the pattern of the activity using rank statistic. Four mean curves of the four groups (first row), the corresponding four mean curves of the rank statistic (second row), and the Haar wavelet reconstructions (red line)

5. CONCLUSIONS

In this paper, we examine some new functional clustering algorithms for physical activity data measured by accelerometer. Specifically, we construct input variables for the clustering algorithms based on rank-based and thick-pen transforms, each of which aim to reflect the intrinsic structures of the functional data, such as activity quantity and activity pattern. The proposed functional clustering algorithms for the transformed input variables make use of a marginal representation of a random curve for a particular cluster.

REFERENCES

Chiou, J. M. and Li, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society Series B,* 69(4), 679– 699.

Fryzlewicz, P. and Oh, H. S. (2011). Thick pen transformation for time series. *Journal of the Royal Statistical Society Series B,* 73(4), 499–529.

Technical Efficiency Analysis of SMEs in Nusa Tenggara Timur Province 2015: An Integration of Stochastic Frontier Analysis and Six Sigma Methodology

Erli Wijayanti Prastiwi*
Statistics of West Sumba Regency, West Sumba, Indonesia – erliwijayanti@bps.go.id

Sari Ayutyas
Statistics of Nusa Tenggara Timur Province, Kupang, Indonesia – sariayutyas@bps.go.id

Dewi Kurnia Ayuningtyas
Statistics of Nusa Tenggara Timur Province, Kupang, Indonesia – dewi_kurnia@bps.go.id

Valent Gigih Saputri
Statistics of West Sumba Regency, West Sumba, Indonesia – valent.gs@bps.go.id

Abstract

Nusa Tenggara Timur (NTT) is one of less developed province in Eastern Indonesia which has long coastline and dry landscape in its most regencies. Limited natural resources, high cost transportation, lack of machinery and equipment support, and also undeveloped entrepreneurship are problems and obstacles in NTT especially for Small Medium Enterprises (SMEs) development. Therefore, it is critical for SMEs in NTT to have good technical efficiency to overcome the challenges and increase their production. This paper aims to measure the Technical Efficiency (TE) of SMEs in NTT 2015 and the factors affecting the technical inefficiency effect (TIE) by integrating Stochastic Frontier Analysis (SFA) into Six Sigma Methodology. In this paper, the implementation of SFA is well incorporated into Define, Measure, Analyse, Improve, and Control (DMAIC) process in Six Sigma Methodology. The results show that the best model chosen in the analysis is Translog Production Function and the average of TE of SMEs in NTT 2015 is 78 percent. It means that there is 22 percent chance to achieve optimum production with combination of current inputs. The factors significantly affect the Technical Inefficiency Effect (TIE) of SMEs in NTT 2015 are entrepreneur's education level, average of working hours per week, membership in association, and government's supports. The contributions of this paper are not only for the government but also for entrepreneurs in SMEs and researchers interested in productivity analysis as well as Six Sigma Methodology practitioners. For the government, this paper presents quantitative reference about efficiency to make better policy for SMEs in the future. For entrepreneurs in SMEs, informations regarding how to measure their business performance will make their upcoming planning and evaluation become easier. For researchers and Six Sigma Methodology practitioners, this paper provides other perspective about productivity measurement as well as the implementation to improve the capability of the business process using Six Sigma techniques.

Keywords: Optimum production, Industry, Inefficiency, Entrepreneurship.

1. INTRODUCTION

The development of Small Medium Enterprises (SMEs) in Nusa Tenggara Timur (NTT) Provinces are predicted to be declined in the long term (IRGSC, 2014). The numbers of SMEs are increasing but they do not have capability of enlarging the business scale. Limited natural resources, high cost transportation, lack of machinery and equipment support, and also undeveloped entrepreneurship are problems and obstacles faced by the enterpreneurs. For example is kain tenun (the signature fabric of NTT) industry, a traditional industry which dominates SMEs in NTT. It takes very long time to produce a piece of kain tenun and the price is not high enough to cover the production cost. However, it is critical for NTT to protect the existence of kain tenun. It raises curiousity among researchers in NTT province regarding how to achieved the great success of the SMEs in NTT without loosing the value of the local wisdom attached in the SMEs in NTT. The SMEs need to be efficiently run. If SMEs in NTT are able to prodiuce more efficiently with better quality, employ better workers, use more decent equipments in the process, NTT will have a great opportunity to develop modern SMEs to prosper the human live in the province. Measuring the technical efficiency and controling the quality of SMEs are needed to evaluate the readiness of SME to be developed further. The technical efficiency measurement also can be used as reference to the government to make any development policy. Therefore, this paper aims to measure the Technical Efficiency (TE) of SMEs in the frame of Six Sigma methodology to maximize the analysis.

2. METHODOLOGY

This paper use 1651 SMEs data sample from SMEs annual survey in 2015 in Nusa Tenggara Timur Province held by Directorate of Industry, Statistics Indonesia. The variables used in this paper are output value of the company, the expenditure for raw materials and auxiliary materials, number of workers, the value of fixed capital, enterpreneur's education level, age of the company, average working hours per week, enterpreneur's age, production difficulties experienced by the SME, partnership, membership of koperasi or association, and the aid received by the SME.

Six Sigma

Six Sigma is the application of the scientific method to the design and operation of management systems and business processes which enable employees to deliver the greatest value to customers and owners (Pyzdek, 2003). The benefits of Six Sigma are achieved through the utilization of sistematic approach which are the Define, Measure, Analyse, Improve, and Control (DMAIC) process. Six Sigma project stay on track by establishing deliverables at each phase of the DMAIC process, in which various graphical or analytical tools are implemented. The primary reason for the success of Six Sigma is that it provides an overall approach for quality and process improvement. Six Sigma provides a systematic process for improvement, DMAIC, that clearly shows how to link and sequence individual tools (Feng, Q., & Antony, J., 2009).

Based on accumulated experience in implementing Six Sigma, researchers and practicioners observed that Six Sigma has its inherent limitation and cannot be used as a universal solution for any process in any organization (Tang et al, 2007). One way to enhance Six Sigma is to integrate additional effective tecniques. Recent technical development in the field of econometrics and statistical analysis has provide mor effective methodology for improving the efficiency and producivity of organizations. One of these developments is Stochastic Frontier Analysis (SFA).

Stochastic Frontier Analysis

The first concept of efficiency was introduced by M. J. Farrell 1957 in a research entitled "*The Measurement of Productive Efficiency*" which was applied in agriculture production data of the US. Technical efficiency of a company is also defined as a measurement of a company's success in producing maximum output with given inputs. When output can not be increased without reducing other outputs or add raw materials, then it is said that the company had achieved the technical efficiency. Technical efficiency can also be interpreted as a comparison of observations and the optimal value of the output of a production (Farrell, 1957 & Chiona, 2011).

This research uses Econometrics method named frontier production function to measure the technical efficiency of SMEs. Aigner *et al.* (1977) and Meeusen (1977) introduced the stochastic frontier production function by adding random error (v_i) in the former function deterministic frontier introduced by Aigner and Chu (1968). The frontier production function was writen as follows:

$$\ln(y_i) = x_i'\beta + \varepsilon_i \quad (1)$$

$$\ln(y_i) = x_i'\beta + v_i - u_i \ ; i = 1, 2, \ldots N \quad (2)$$

It can be seen that the random error (ε_i) consists of two components, they were v_i and u_i. The first component of the random error is v_i which is not be able to be controlled by company such as measurement error, bad weather, government policy, any sudden shocks in politics or natural disaster. In other words, v_i can be negative or positive.

Another random component in the stochastic frontier production function beside v_i is u_i which is a deviation caused by factors that could be controlled company, such as defective and damaged products, workforce skills, management ability, the amount of hours of work, or other controlled disorder. Components inefficiency effects u_i is assumed to follow identical and independent truncated normal distribution ($u_i \sim iid\ N^+(\mu, \sigma_u^2)$) or half normal distribution ($u_i \sim iid\ N^+(0, \sigma_u^2)$). The explanation of stochastic frontier production function can be seen in figure 1.

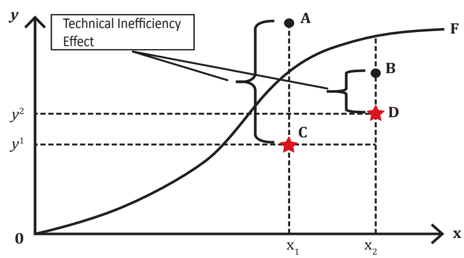


Figure 1: Stochastic frontier production function curve

Based on figure 1, the deterministic component of the model frontier (Y = exp (xβ)) is assumed to have deminishing returns. Production activities of the two companies represented by the indices 1 and 2. Line 0-F indicated the lines deterministic frontier production function which indicated a potential condition that could be achieved by the company. As with the regression analysis, the line of frontier production function was a line drawn from the values of existing observations in which this line was the line that came closest to the existing observations. Points A and B was the frontier output achieved by companies with existing conditions while points C and D were the observations in the data that shows the output produced by the company in fact.

The first company used input for X_1 and produced Y_1. In other words, the production of the first company is at point C. Output frontier of the first company is at point A which was above the line 0-F which is a line deterministic frontier production function. Because of the deviation between the output frontier that should be achieved by the company with the output of the observation, it can be said that there is technical inefficiency in the first company. AC shows the technical inefficiency effects experienced by the first company.

Meanwhile, the second company input is X_2 and produce Y_2 as the output. In other words, the first company production was at point D. The output from the frontier of the second company was at point B, below the 0-F which is a line deterministic frontier production function. This is due to the conditions which inhibit the production process or in other words, the error component v_i was negative. As happened in the first company, the second company also has deviation between the output frontier that should be achieved by the company with the output of the observation, it can be said that there is technical inefficiency in the second company. Line BD indicates that there is technical inefficiency effects experienced in the second company.

Production function that used often are the Cobb-Douglas and translog. Cobb-Douglas production function is more easily assessed and modified to be estimated mathematically. From the general model specification Stochastic Frontier Analysis, translog production function is as follows (Nicholson and Christopher, 2008):

$$\ln Y = \beta_0 + \sum_{i=1}^{n} \beta_i \ln X_i + 0,5 \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_{ij} \ln X_i \ln X_j \quad \text{where } \beta_{ij} = \beta_{ji} \tag{3}$$

Battese and Coelli *et al.* in Coelli (1998) described estimates for the value of technical efficiency or Technical Efficiency (TE) for the cross section data is as follows:

$$TE_i = \frac{y_i}{\exp(\mathbf{x}_i'\boldsymbol{\beta})} \tag{4}$$

$$TE_i = \frac{\exp(x_i'\beta + v_i - u_i)}{\exp(x_i'\beta + v_i)} = \exp(-u_i) \tag{5}$$

Integrating Six Sigma and SFA

This paper explores the feasibility of integrating the SFA method and the Six Sigma to enhance the usefulness and effeciveness of Sig Sigma. The role of SFA is identified in each phase of the DMAIC process. The core of the SFA will be explained further in Measure phase and the results will be further elaborated in Analysis phase. In the other words, integrating Six Sigma and SFA means to strengthen Six Sigma's Measurement and Analysis phases by using SFA and also enrich SFA to become more systematic and holistic package using DMAIC phases in Six Sigma.

4. RESULTS

Define Phase

In Define Phase, the research usually use pareto chart. In this paper, pareto chart is not appropriate to applied. Therefore, the Define Phase will be explained based on the descriptive analysis. The observation of the variables indicates that the participation of SMEs in several government;s program has not been optimized yet. There are only 18 percent of SMEs in NTT become the member of Koperasi (local business organization) supported by the government. There is only 9 percent of SMEs in 2015 received aid from the government. In the perpective of education used to represent the capability of management, there are only 18 percent of the enterpreneurs in NTT graduated from Junior High School or in the other words, there are 82 percent of the SMEs enterpreneur in NTT 2015 that are not well educated. It shows that the obtacles faced by the SMEs in NTT 2015 not only in the term of input process of the industry but also from the side of human development.

Measure Phase

The first step done to do SFA is to measure the Technical Efficiency (TE) of the company in all possible production fuction and statistics distribution. The production function chosen through the Likelihood Ratio (LR) test is Translog Production Function with truncated normal distribution (Table 1 and Table 2). The variables significantly contributed in the production fuction can bee seen in table 3.

Table 1: Likelihood Ratio (LR) test in Stochastic Frontier Production Function Selection

Hypothesis	Log Likelihood	LR	Critical Value	Decision
$H_0 : \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{22} = \cdots = \beta_{44} = 0$ **(Cobb-Douglas)**	-1336.22	80.14	18.31	Reject H_0 **(Translog)**
H_1 : at least one parameter $\neq 0$ **(Translog)**	-1296.15			

Table 2. Likelihood Ratio (LR) test in Technical Inefficiency Effect Distribution Selection

Hypothesis	Log Likelihood	LR	Critical Value	Decision
$H_0 : \mu = 0$ **(Half Normal)**	-1296.15	24.31	3.84	Reject H_0 **(Truncated Normal)**
$H_1 : \mu \neq 0$ **Truncated Normal**	-1283.99			

Table 3. Estimated Parameters of Translog Production Function

Variable	Parameter	Coefficient	Standard-Error	t-ratio
(1)	(2)	(3)	(4)	(5)
Intercept*	β_0	10.92450	1.00424	10.87842
In Raw Material$_i$	β_1	0.00963	0.13672	0.07040
In Worker$_i$	β_2	0.34498	0.29931	1.15261
In Fixed Capital$_i$*	β_3	-0.22478	0.08587	-2.61765
Interaction Raw Material*	β_{11}	0.05377	0.01040	5.17151
Interaction (Raw Material and Worker)	β_{12}	0.10100	0.10422	0.96911
Interaction (Raw Material and Fixed Capital)*	β_{13}	0.01767	0.00607	2.91337
Interaction (Worker)	β_{22}	0.02091	0.02121	0.98562
In teraction (Worker and Fixed Capital)	β_{23}	0.00041	0.00424	0.09724
Interaction (Fixed Capital)*	β_{33}	-0.04016	0.01305	-3.07742

Analysis Phase

The result of the estimated TE shows that in average, the TE of NTT Province 2015 is 78 percent. It means that SMEs in NTT 2015 has reached 78 percent of the potential production which can be achieved. In other words, the SMEs in NTT still has 22 percent chance to achieve the optimum production with the current inputs in the current circumstances. Entrepreneur's education level, average of working hours, membership in "koperasi" or association, and the aid received from the government are all factor that significantly affect the Technical Inefficiency effect of SMEs in NTT Province (Table 4).

Table 4: Estimated Parameters of Factors which affect the Technical Inefficiency Effect

Variable	Parameter	Coefficient	Standard-Error	t-ratio
(1)	(2)	(3)	(4)	(5)
Intercept*	∂_0	0.39198	0.09059	4.32682
Education$_i$*	∂_1	-0.13104	0.04297	-3.04922
Age of Company$_i$	∂_2	-0.00083	0.00159	-0.52463
Average of Working Hours$_i$*	∂_3	-0.00707	0.00094	-7.49861
Age of Enterpreneur$_i$	∂_4	0.00062	0.00134	0.46190
Main Obtacle$_i$	∂_5	0.07799	0.05657	1.37871
Partnership$_i$	∂_6	0.00000	1.00000	0.00000
Membership in Koperasi$_i$*	∂_7	0.16190	0.07419	2.18226
Aid$_i$*	∂_8	-0.18436	0.04521	-4.07744

Improve Phase

After the grass root understanding about the problem and the factors affecting technical inefficiency effect are found, the next step is to take a real action to overcome the problem and help the SMEs to help themselves. It takes further research and government policy to optimize the production of SMEs and to give solutions to the SMEs.

Control Phase

Since this paper only provide analysis form the secondary data, control phase will not be possible in this current stage. Therefore, the result of this paper can be used by the practitioners of Six Sigma and the government to make further cooperation to control the development of SMEs in NTT in the near future.

5. CONCLUSIONS

The average of technical efficiency of SMEs in NTT province 2015 is 78 percent so there is 22 percent chance to achieve optimum production with combination of the current inputs. Entrepreneur's education level, average of working hours, membership in "koperasi" or association, and the aid received from the government are all factor that significantly affect the Technical Inefficiency effect of SMEs in NTT Province. Although SFA can be integrated well in the Six Sigma Methodology, it takes further action from the government, researchers, and Six Sigma Practitioner to make Improve Phase and Control Pase possible.

REFERENCES

Chiona, Susan. (2011). *Technical and Allocative Efficiency of Smallholder Maize Farmers in Zambia.* [Thesis]. Lusaka: The University of Zambia.

Coelli TJ, et al. (1998). *An Introduction of Efficiency and Productivity Analysis.* London: Kluwer Academic Publisher.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A (General)*. Vol. 120, No.3, (1957) pp. 253-290.

Feng, Q & Antony, J (2010). *Integrating DEA into Six Sigma Methodology for Measuring Health Service Efficiency*: Journal of the Operation Research Society, Vol 61, No 7.

Ercoc, Taptuk Emre. (2012). *Estimation Methodology of Economic Efficiency: Stochastic Frontier Analysis vs Data Envelopment Analysis*. International Journal of Academic Research in Economics and Management Science. January 2012, Vol 1. No 1.

Pyzdek, Thomas. (2003). *The Six Sigma Handbook.* The McGraw Hill Companies.

Prediction Intervals of Model Averaging Methods for High – Dimensional Data

Septian Rahardiantoro*
Department of Statistics-Bogor Agricultural University, Bogor, Indonesia – rahardiantoro.stk@gmail.com

Khairil Anwar Notodiputro
Department of Statistics-Bogor Agricultural University, Bogor, Indonesia –khairilnotodiputro@gmail.com

Anang Kurnia
Department of Statistics-Bogor Agricultural University, Bogor, Indonesia – anangk@apps.ipb.ac.id

Abstract

High dimensional data analysis, in recent year, became one of the most popular topics. In case of regression analysis, model averaging is known an alternative method in the purpose of prediction the response variable. The problems occur when the researchers take of specify prediction intervals based on this method. The pattern of prediction distributions are unknown, in other word we can't easily say that the predictions normally distributed or simmetric tailed. Defining the prediction intervals in unknown pattern of distributions based on model averaging predictions would be the focuss of this study. The approach that applied in this study for getting the prediction intervals is specified quartile approach. Firstly, we define the specified pairs of quantile based on the certain probability, and then evauated to get the shortest intervals. The data set of simulation in condition of highly dimension, with number of predictors 1000, and number of observation 100. Then, model averaging applied in these data in order to create the predictions. In a thousand replications, the propose method could be an alternative method to get the shortest prediction intervals in high – dimensional regression.

Keywords: High dimensional data; Quantile; Model averaging; Prediction intervals

1. INTRODUCTION

Regression analysis as we know very well to reveal the relationship among variables, is also developed to deal with high – dimensional data. In recent century, high – dimensional data is known as data with number of variables (p) exceed the number of observations (n), that is $p >> n$. For handling those data condition, classical regression analysis has a lot of limitations. Classical approaches can't to use because multicollinear problems, and also the anova isn't defined well cause restriction in degree of freedom (Myers, 1990).

In 2014, Ando and Li developed the classical approaches therefore it can be use in high – dimensional cases, that called model averaging (MA) (Ando & Li, 2014). This method is designed to get the predictions of response variable accurately. Simply, MA takes some model candidates from k partitions of the predictor variables in the order of each marginal correlation with the response variable. Finally, the predictions take based on the ensamble approach by taking averaging of the predictions of k model candidates.

The problems occur when the researchers want to take some prediction intervals when using MA in high – dimensional data. The normal – based prediction intervals are not always recommended method for that. Basic approach, the researchers can built the prediction intervals based on the data conditions. In this study, we focus on the method to define the prediction intervals in high – dimensional data using MA.

This study is based on the simulation study, by taking high – dimensional data that applied MA, and then the proccess of creating prediction intervals taken. The approach that applied in this study for getting the prediction intervals is specified quartile approach. In practice, this study is using R software in implementing all of process.

2. MODEL AVERAGING (MA)

The main idea of MA is getting the weighted averaging of some model candidate predictions to be the final predictions. Assume we have a matrix $X_{n \times p}$, that indicates the matrix of predictor variables with $p >> n$. The first step is ordering every predictor variables by taking the marginal correlation between every predictor variables with the response variables. The matrix result of this proccess is called $\tilde{X}_{n \times p}$. Then we take k partitions from $\tilde{X}_{n \times p}$, that is in every partition has m predictor variabels. The OLS regression model of the response variable with the predictor variables matrix partition that is called model candidate. Finally, from each model candidates, we can take the predictions, and the last step is taking weighted averaging of that (Ando & Li, 2014).

In this study, we define the weight w_i as AIC weight which is based on the value of AIC in each model candidates. Suppose there are k model candidates, therefore the i – th AIC weight follows

$$w_i = \frac{exp\left(\frac{1}{2} a_i\right)}{\sum_{i=1}^{k} exp\left(\frac{1}{2} a_i\right)} \tag{1}$$

where a_i denotes the value of AIC in the i – th model candidates, and $w_i \geq 0$; $\sum_{i=1}^{k} w_i = 1$ (Claeskens & Hjort, 2008).

The weighted average of MA define by the following formula that take the average of each model candidate predictions. Suppose $\hat{\boldsymbol{y}}_i$ is the prediction vector of i – th candidate model, therefore

$$\hat{y} = \sum_{i=1}^{k} w_i \hat{y}_i \tag{2}$$

become the ensamble prediction vector.

3. PREDICTION INTERVALS METHODOLOGY

In this section, the topics would be described are the methodology of prediction intervals. This study would be applied the approach based on the basic concept to get the shortest interval in specified probability (Casella & Berger, 2002). The method to get the prediction intervals is specified quartile approach. The main idea of this approach is taking some pairs of quartile and then it would be evaluated to get the shortest interval. The following parts would be described the algorithm of the method.

The algorithm of the method:

1. Define a probability of prediction intervals (pb). This probability indicates the accuracy of intervals designed. In this study, the probabilty is 0.95, therefore we have 0.05 probability outside the intervals, that means 5% toleration of errors.
2. Define pairs of number of quartiles, (Q_0, $Q_{0.95}$), ($Q_{0.005}$, $Q_{0.955}$), ($Q_{0.01}$, $Q_{0.96}$), ($Q_{0.015}$, $Q_{0.965}$), ($Q_{0.02}$, $Q_{0.97}$), ($Q_{0.025}$, $Q_{0.975}$), ($Q_{0.03}$, $Q_{0.98}$), ($Q_{0.035}$, $Q_{0.985}$), ($Q_{0.04}$, $Q_{0.99}$), ($Q_{0.045}$, $Q_{0.995}$),($Q_{0.05}$, Q_1). The value of each quartile calculated using the following formula:

$$Q_{pb} = arg_Q \frac{\sum_{h=1}^{z} I_{x_h \leq Q}}{z} = pb$$

3. Count the difference of each pairs of quartiles, and find the shortest to be the solution of intervals.

4. SIMULATION STUDY

High-dimensional data are designed by simulation using R software with the predictors dimensions are p = 1000 & n = 100. It designed by following the multivariate normal distribution with mean vector 0, and variance-covariance matrix $S = Ip$. The number of true regressors s = 50 and the true regressors x_i be spaced evenly, $i = 20(j - 1) + 1, j = 1,2,...50$. Random effects in this data are generated follows *log - Normal (0,4)*, that indicate non-simmetric distribution of random effects.

In practice, the data separated to be two parts, the training set and testing set. Both of it contain 50% observation of the data. The training set is used to take a modelling proccess with MA, and then the model evaluated to predict the value of testing set. This simulation generated on 1000 times replications. Therefore the prediction intervals applied in the 1000 results of replications.

5. RESULT AND DISCUSSION

The figure below describes the pattern of prediction mean from the result of testing set based on the model that built by the training set using MA. As shown in figure 1, the pattern of prediction is non-simmetric, and subject to right-tailed. This is also recall that the normal-base prediction intervals are not recommended. Next, the proposed method applied to take the prediction intervals.

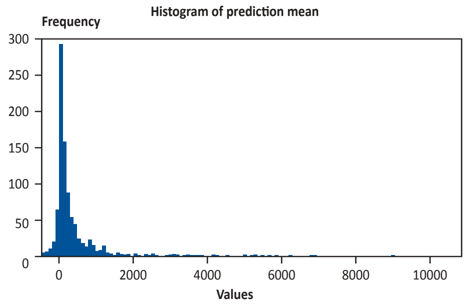


Figure 1. Histogram of prediction mean

The prediction intervals are taken by applying the proposed method in testing set. The individual values predicted of each intervals, therefore the result has 50 intervals. The Figure 2 shows the prediction intervals of each individual value in testing set.

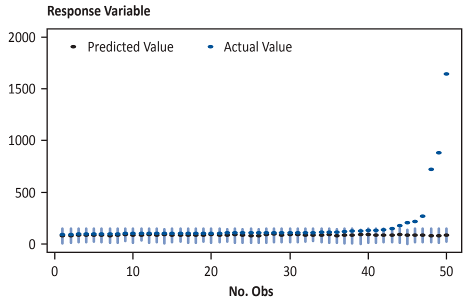


Figure 2. Plot of prediction intervals in testing set

From Figure 2 we can see that almost the prediction intervals contain the actual values of response variable. It is about 86% the true value consist in the prediction intervals. It seems pretty good to say that the proposed method is also a good choise to get the prediction intervals when handling the data in high dimension, specially using MA to analyze.

6. CONCLUSIONS

It can be conclude that in high – dimensional data which applied MA to predict the response variable, the prediction intervals can be taken by using the proposed method, that is using the shortest quantile value in specified probability. From the simulation in this study, it is about 86% accuracy of prediction intervals that contain the true value of response variable.

REFERENCES

Myers, R. H. (1990). Classical and Modern Regression with Applications Second Edition. PWS-KENT Publishing Company.

Ando, T., & Li, K. C. (2014). A Model-Averaging Approach for High-Dimensional Regression, Journal of the American Statistical Association. 194: 254-265.

Claeskens, G., & Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge University Press.

Casella, G., & Berger, R. L. (2002). Statistical Inference Second Edition. Duxbury Thomson Learning.

The Characteristic Function Property of Convoluted Random Variable from a Variational Cauchy Distribution

Dodi Devianto
Department of Mathematics, Andalas University,
Padang 25163, West Sumatra Province, INDONESIA
E-mail address: ddevianto@fmipa.unand.ac.id

Abstract

The variational *Cauchy* distribution is constructed by setting the shape parameter of *Cauchy* distribution multiplied with part of random variable in standard *Cauchy* distribution. The convolution of a variational *Cauchy* distribution is obtained by using properties of inversion theorem of characteristic function. Then it is given some potensial properties of characteristic function of convoluted random variable from a variational *Cauchy* distribution. The properties of uniform continuity and complex conjugate of characteristic function is determined by mathematical analysis methods, and it is confirmed that the characteristic function has never vanish on the complex plane and infinitely divisible.

Keywords: Variational *Cauchy* distribution; Convolution; Characteristic function, Infinitely divisible.

1. INTRODUCTION

The *Cauchy* distribution is due to Augustin Louis Cauchy in 1853 who introduced a continuous distribution as standard *Cauchy* density. The *Cauchy* distribution is a case of continuous stable distribution for which does not have mean, variance and other moments. Furtermore, the infinitely divisible property of *Cauchy* distribution is the closed invariant property under convolution, then it is on class of infinitely divisible because all convolutions belonging to this type again yield the same distribution after convolution stability of this distribution. The specific case of infinite divisibility of this distribution has shown by Bondesson [2] for the half *Cauchy* densities, while Takano [8] has exhibited the infinite divisibility of normed product of *Cauchy* densities. However, Dwass [6] has developed the convolution of *Cauchy* distribution by using induction from definition of convolution of two random variables. However, the convolution theory also very widely use to express properties of some distribution, such as Devianto et. al [4, 5] are also using convolution to determine the sum of exponential distribution with stabilizer constant and its properties. These previous result of convolution is to confirm that convolution of random variables now very interesting to establish its mathematical properties.

The study of convolution and infinitely divisible *Cauchy* distribution has widely developed to the property of its characteristic function, since the characteristic function is always exist as the most general tools to determine convolution of random variables. The characteristic function from a random variable X is defined as Fourier-Stieltjes transform as the following

$$\phi_X(t) = E[\exp(itX]$$

where $\exp(itX) = \cos tX + i \sin tX$ and i as imaginary unit. The characteristic function has inversion theorem as the uniqueness property that is for every probability distribution function $f(x)$ from random variable X has a unique characteristic function $\phi_X(t)$ such that the probability distribution function can be obtained by using inversion of characteristic function as follows

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[-itx]\ \phi_X(t)\, dt\,.$$

The refinement of this characteristic function property also can be used to determine convolution of a distribution. The convolution of *Chauchy* distribution can be obtained from its characteristic function by using linerity property of expectation.

The establishment theory of characteristic function and its uses to determine convolution and infinite divisibility have developed the stability properties of *Cauchy* distribution and its variation such as half *Cauchy* distribution by Bondesson [2] and normed normed product of *Cauchy* densities by Takano [8], while Devianto [3] also defined a variational *Chaucy* distribution and its concolution. The probability density function of a variational *Cauchy* distribution is constructed by setting parameter γ multiplied with part of random variable in standard *Cauchy* distribution, so that it can be defined in the following term

$$f(x) = \frac{\gamma^{1/2}}{\pi(1+\gamma x^2)} \quad \text{for } \gamma > 0 \text{ and } x \in (-\infty, \infty).$$

The cummulative probability distribution function of a variational *Cauchy* distribution is obtained as follows

$$F(x) = \frac{1}{2} + \frac{2\arctan\left(\gamma^{1/2}\, x\right)}{\pi} \quad \text{for } \gamma > 0 \text{ and } x \in (-\infty, \infty).$$

The direction of this paper is to contruct the characteristic function of convoluted of a variational *Chaucy* distribution and its property of characteristic function. The section 2 is devoted to establish the property of characteristic function of convoluted of a variational *Chaucy* distribution, while in Section 3 is discussed its infinite divisibility.

2. THE CHARACTERISTIC FUNCTION PROPERTIES OF CONVOLUTED RANDOM VARIABLE FROM A VARIATIONAL CAUCHY DISTRIBUTION

This section is started by using definition of non-negative function and necessary and sufficient conditions for a function to be a characteristic function in Bochner's Theorem. We use definition of non-negative function as necessary and sufficient conditions to be a characteristic function from Lukacs [7].

Definition 2.1. *A complex-valued function* $\phi(t)$ *with real variable t is said to be non-negative if the following two conditions are satisfied*

(i) $\phi(t)$ is continuous;

(ii) *For any positively defined function with quadratic form*

$$\sum_{1\le j\le n} \sum_{1\le l\le n} c_j \overline{c_l}\, \phi_X(t_j - t_l) \ge 0$$

for any complex number $c_1, c_2, \ldots, c_n$ *and real number* $t_1, t_2, \ldots, t_n$.

Theorem 2.2. (Bochner's Theorem). *A complex-valued function* $\phi(t)$ *of a real variable t is a characteristic function if, and only if,*

(i) $\phi(0) = 1$;

(ii) $\phi(t)$ is non-negative definite.

The characteristic function of a variational *Cauchy* distribution has introduced by Devianto [3] by using concept of complex integration, that is for *X* as random variable from a variational Cauchy distribution has characteristic function

$$\phi_X(t) = \exp\left[-\frac{|t|}{\gamma^{1/2}}\right]$$

for $\gamma > 0$ and $t \in (-\infty, \infty)$. The convolution of random variables from a variational *Cauchy* distribution is determined by using property of characteristic function. Let Random variable X_i has a variational *Cauchy* probability distribution with parameter γ, then random variable

$$S_n = \sum_{i=1}^{n} X_i$$

as convolution of random variables X_i has probability distribution function

$$f(s_n)=\frac{n\,\gamma^{1/2}}{\pi\,(n^2+\gamma\, s_n{}^2)}$$

for $\gamma>0$ and $s_n\in(-\infty,\infty)$. The characteristic function of random variable S_n is obtained by using linearity of expectation for characteristic function as follows

$$\phi_{S_n}(t)=E[e^{itS_n}]=\exp\left[-\frac{n\,|t|}{\gamma^{1/2}}\right]$$

The characteristic function has some refinement properties, this section gives basic properties of characteristic function from convoluted random variable of a variational *Cauchy* distribution in the propositions.

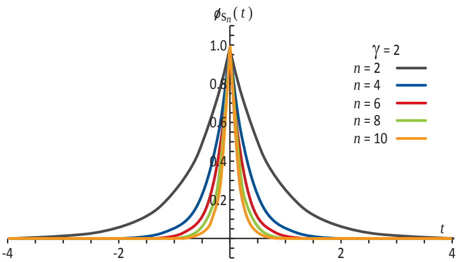


Figure 1. Curve of characteristic function $\varphi_{S_n}(t)$ as simetric distribution from random variable S_n with parameter γ = 2 and various value of *n*-fold convolution.

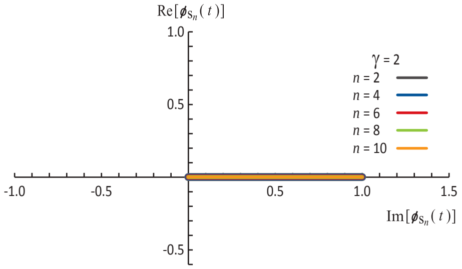


Figure 2. Parametric curves of characteristic function $\varphi_{S_n}(t)$ with parameter γ = 2 and various value of n-fold convolution

The simetric shape of distribution for *n*-fold convolution a variational Cauchy distribution causes the characteristic function only has real part. Figure 1 has plotted the function $\phi_{Sn}(t)$ on the cartesian coordinate with t in x-axis, the curve comes as smooth lines with extreme values on $\phi_{S_n}(0)=1$. Furtehermore, the characterization of characteristic function can be seen from the shape of parametric curves. The parametric curves are governed by using parametric plot at cartesian coordinate system for x-axis as real part of characteristic function and y-axis as imaginary part of characteristic function. Figure 2 has shown that characteristic function $\phi_{Sn}(t)$ lies only in the real part in the interval $0\le \mathrm{Re}[\phi_{S_n}(t)]\le 1$. These properties imply that characteristic function $\phi_{Sn}(t)$ is positively defined function and never vanish on the complex plane.

Proposition 2.3. *Let S_n be a random variable from convolution of a variational Cauchy distribution with characteristic function*

$$\phi_{S_n}(t)=\exp\left[-\frac{n\,|t|}{\gamma^{1/2}}\right].$$

then it is satisfied $\phi_{S_n}(0)=1$;

Proof. (i) It is easily obtained that for t = 0 then $\phi_{S_n}(0)=1$.

Proposition 2.4. *Let S_n be a random variable from convolution of a variational Cauchy distribution with characteristic function*

$$\phi_{S_n}(t)=\exp\left[-\frac{n\,|t|}{\gamma^{1/2}}\right].$$

then characteristic function $\phi_{S_n}(t)$ is uniformly continuous.

Proof. The property of uniformly continuous of characteristic function from convoluted random variable of a variational Cauchy distribution is explained by setting for every ε > 0 there exists δ > 0 such that $\left|\phi_S(t_1) - \phi_S(t_2)\right| < \varepsilon$ for $\left|t_1 - t_2\right| < \delta$ where δ depends only on ε. Let us set the the following equation

$$\left|\phi_{S_n}(t_1) - \phi_{S_n}(t_2)\right| = \left| \exp\left[-\frac{n|t_1|}{\gamma^{1/2}}\right] - \exp\left[-\frac{n|t_2|}{\gamma^{1/2}}\right]\right|$$

Then let us define $h = t_1 - t_2$, so that we have

$$\left|\phi_{S_n}(t_1) - \phi_{S_n}(t_2)\right| = \left| \exp\left[-\frac{n|h-t_2|}{\gamma^{1/2}}\right] - \exp\left[-\frac{n|t_2|}{\gamma^{1/2}}\right]\right|$$

By taking limit for $h \to 0$, then it is obtained

$$\lim_{h\to 0}\left| \exp\left[-\frac{n|h-t_2|}{\gamma^{1/2}}\right] - \exp\left[-\frac{n|t_2|}{\gamma^{1/2}}\right]\right| = 0$$

This result of limiting process hold for every δ < ε where $\left|\phi_{S_n}(h+t_2) - \phi_{S_n}(t_2)\right| < \varepsilon$ for $\left|t_1 - t_2\right| < \delta$. Then the characteristic function of $\phi_{S_n}(t)$ is uniformly continuous.

Proposition 2.5. *Let S_n be a random variable from convolution of a variational Cauchy distribution with characteristic function*

$$\phi_{S_n}(t) = \exp\left[-\frac{n|t|}{\gamma^{1/2}}\right].$$

then characteristic function $\phi_{S_n}(t)$ is positively defined function with quadratic form

$$\sum_{1\le j\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \phi_{S_n}(t_j - t_l) \ge 0$$

for any complex number $c_1, c_2, ..., c_n$ and real number $t_1, t_2, ..., t_n$.

Proof. It will show that characteristic function $\phi_{S_n}(t)$ from convoluted random variable of a variational *Cauchy* distribution satisfies the quadratic form

$$\sum_{1\le j\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \phi_{S_n}(t_j - t_l) \ge 0$$

for any complex numbers $c_1, c_2, ..., c_n$ and real numbers $t_1, t_2, ..., t_n$. Let us use definition of characteristic function from convoluted random variable of a variational *Cauchy* distribution, hence we have

$$\sum_{1\le j,l\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \phi_{S_n}(t_j - t_l) = \sum_{1\le j\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \exp\left[-\frac{n|t_j - t_l|}{\gamma^{1/2}}\right]$$

It is used the modulus property of inequality $|t_j - t_l| \ge |t_j| - |t_l|$, then we have the following form

$$\begin{aligned}\sum_{1\le j,l\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \phi_{S_n}(t_j - t_l) &\le \sum_{1\le j\le n}\ \sum_{1\le l\le n} c_j \overline{c_l}\ \exp\left[-\frac{n|t_j| - n|t_l|}{\gamma^{1/2}}\right] \\ &= \sum_{1\le j\le n} c_j \exp\left[-\frac{n|t_j|}{\gamma^{1/2}}\right] \overline{\sum_{1\le l\le n} c_l \exp\left[-\frac{n|t_l|}{\gamma^{1/2}}\right]} \\ &= \left|\sum_{1\le j\le n} c_j \exp\left[-\frac{n|t_j|}{\gamma^{1/2}}\right]\right|^2\end{aligned}$$

The last part of equation above has been in the quadratic form then we have

$$\sum_{1\le j\le n}\sum_{1\le l\le n} c_j \overline{c_l}\, \phi_{S_n}(t_j - t_l) \ge 0$$

It is proved that $\phi_{S_n}(t)$ as positively defined function where the quadratic form has nonnegative values.

3. THE INFINITELY DIVISIBLE CHARACTERISTIC FUNCTION OF CONVOLUTED RANDOM VARIABLE FROM A VARIATIONAL CAUCHY DISTRIBUTION

This section will explain the infinitely divisible characteristic function of convoluted random variable from a variational Cauchy distribution. The definition of infinitely divisible characteristic function is referred to Artikis [1] that confirmed for distribution function of F with characteristic function $\phi(t)$ is infinitely divisible if for every positive integer m there exists a characteristic function $\phi_m(t)$ such that $\phi(t) = (\phi_m(t))^m$.

Proposition 3.1. *Let us define a function*

$$\phi_S(t) = \exp\left[-\frac{n|t|}{\gamma^{1/2}}\right]^{\frac{1}{m}}$$

for every $n, m \in Z^+$, $\lambda > 0$ *and* $-\infty < t < \infty$ *, then it is satisfied*

(*i*) $\phi_S(0) = 1$;

(*ii*) $\phi_S(t)$ *is uniformly continuous*;

(*iii*) $\phi_S(t)$ *is positively defined function with quadratic form*

$$\sum_{1\le j\le n}\sum_{1\le l\le n} c_j \overline{c_l}\, \phi_S(t_j - t_l) \ge 0$$

for any complex number $c_1, c_2, ..., c_n$ *and real number* $t_1, t_2, ..., t_n$.

Proof. The outline proof of this proposition is similar way with Proposition 2.3, Proposition 2.4 and Proposition 2.5, then Proposition 3.1 is obviously proven.

Theorem 3.2. *The function*

$$\phi_S(t) = \exp\left[-\frac{n|t|}{\gamma^{1/2}}\right]^{\frac{1}{m}}$$

is a characteristic function.

Proof. Proposition 3.1 has established the necessary and sufficient condition for $\phi_S(t)$ to be a characteristic function that required by Theorem 2.2. Then the function $\phi_S(t)$ is a characteristic function.

Theorem 3.3. *The c*onvolution of a variational Cauchy distribution is infinitely divisible.

Proof. The infinite divisibility is shown by using the property of characteristic function $\phi_{S_n}(t)$ such that satisfies necessary and sufficient condition in term of characteristic function, that is $\phi_{S_n}(t) = (\phi_S(t))^m$. Based on Theorem 3.2, there is exist a characteristic function $\phi_S(t)$ such that

$$\phi_S(t)^m = \left(\exp\left[-\frac{n|t|}{\gamma^{1/2}}\right]^{\frac{1}{m}}\right)^n = \exp\left[-\frac{n|t|}{\gamma^{1/2}}\right] = \phi_{S_n}(t)$$

for any positive integer number m. The function $\phi_{S_n}(t)$ is a characteristic function from convolution of a variational Cauchy distribution, so that the characteristic function of convolution of a variational *Cauchy* distribution is an infinitely divisible.

4. CONCLUSION

The characteristic function of convoluted random variable from a variational *Cauchy* distribution is obtained as $\phi_{S_n}(t)$ with some basic properties are uniformly continuous and its complex conjugate of characteristic function is determined by mathematical analysis methods. It is confirmed by graphically that characteristic function of convolution a variational *Cauchy* distribution has lied only at the real part and never vanish on the complex plane. The most important property of convolution a variational *Cauchy* distribution is the infinite divisibility of its characteristic function.

REFERENCES

Artikis T. 1983. Constructing infinitely divisible characteristic functions. Archivum Mathematicum, Vol. 19, No. 2, Page 57-61

Bondesson L. 1987. On the infinite divisibility of the half-Cauchy and other decreasing densities and probability functions on the nonnegative line. Scandinavian Actuarial Journal, Vol. 1987, No. 3-4, Page. 225-247.

Devianto D. 2016. On the convolution of a variational *Cauchy* distribution. Proceeding International Conference on Theoretical and Applied Statistics.

Devianto D, Oktasari L and Maiyastri. 2015. Some properties of hypoexponential distribution with stabilizer constant. Applied Mathematical Sciences, Vol. 9, No. 142, Page 7063-7070.

Devianto D, Maiyastri, Oktasari L and Anas M. 2015. Convolution of generated random variable from exponential distribution with stabilizer constant. Applied Mathematical Sciences, Vol. 9, No. 96, Page 4781-4789.

Dwass M. 1985. On the convolution of Cauchy distributions. The American Mathematical Monthly, Vol. 92, No. 1, Page 55-57.

Lukacs E. 1992. Characteristic Function. Hafner Publishing Company, London.

Takano K. 2003. On Infinite divisibility of normed product of Cauchy densities. Journal of Computational and Applied Mathematics, Vol. 150, No. 2, Page 253

This page intentionally left blank

IPS05: Survey Sampling

Small Area Models for Brazilian Business Skewed Data

Denise Britz do N. Silva*, Fernando A. S. Moura, André Felipe Neves
IBGE, Brazil – denise.silva@ibge.gov.br

Abstract

The Brazilian Institute of Geography and Statistics (IBGE) carries out a Service Sector Annual Survey that focuses on segments of the tertiary sector. Sample estimates for some economic activities in the North, Northeast and Midwest regions of Brazil have low precision due to the sample design. Furthermore, one of the main variables of interest is considerably skewed with potential outliers. To overcome this problem, skew models are proposed to produce model-based estimators. The small domain estimation models relate operating revenue variables with potential auxiliary variables (number of employed persons and wages) obtained from a business register. The proposed models have been compared with the usual Fay-Herriot model under the assumption of unknown sampling variances. The evaluation studies with real business survey data show that the proposed models seem to be more efficient for small area predictions under skewed data than is the customarily employed normal area model.

Keywords: Small area models, survey sampling.

IPS06:
Financial Inclusion

Measures of Financial Inclusion – A Central Bank Perspective

Blaise Gadanecz and Bruno Tissot*

Bank for International Settlements, Basel, Switzerland – Bruno.Tissot@bis.org

Abstract

Central banks' experience shows that better statistics can be instrumental to promote financial inclusion. Well-founded data frameworks are essential when developing financial services for the poor, in both formal and informal markets, and adequate indicators are a precondition for good policies. They ensure that financial inclusion is properly assessed and that measures aimed at developing it are adequately implemented, monitored and adjusted as required. Good statistics can also help to strike a proper balance between encouraging innovation and the growth of financial services on the one hand, and ensuring that financial stability is preserved on the other.

Keywords: Financial services, Financial stability, Poverty, Data measurement, Public policy.

JEL classification: E58, G18, G21, I22, I30, O16

1. INTRODUCTION

Financial inclusion, broadly defined as access to financial services, is expanding globally but remains a key issue for policymakers worldwide. In particular, it is an important public policy goal that directly relates to central banks' key objectives and activities (Mehrotra and Yetman (2015)). Financial inclusion can contribute to sustaining economic welfare and to reducing poverty. It also supports economic, monetary and financial stability, by making saving and investment decisions more efficient, enhancing the transmission of monetary policy and facilitating the functioning of the economy. The international standard-setting bodies (SSBs), especially those hosted by the BIS, have been actively engaged with financial inclusion policies for more than a decade. From a payments perspective, the current focus is on facilitating financial inclusion and access through payment systems (CPMI (2016)). From a supervisory perspective, initial attention was devoted to the microfinance activities of banks and other deposit-taking institutions. Since then, the focus has also shifted to the full range of financial products and services that low-income households should be able access (BCBS (2015)).

A key issue for central banks is the need for monitoring the impact of financial inclusion policies. Adequate indicators are a prerequisite for properly assessing the access to financial services, and for formulating, implementing and monitoring public policy designed to enhance it. Good statistics can also help to strike a balance between encouraging innovation and the growth of financial services on the one hand, and ensuring that financial stability is preserved on the other. A number of international initiatives, undertaken by various public and private sector organisations and SSBs, have helped to set up common frameworks for developing financial inclusion indicators and informing policymakers. But despite many and encouraging improvements, measurement of financial inclusion remains work in progress.

For its part, the Irving Fisher Committee on Central Bank Statistics (IFC) surveyed its member central banks on financial inclusion end-2015 (IFC (2016)). The survey covered 47 countries, of which 30 emerging economies. Specifically, the objective of the survey was to compare financial inclusion policies and practices along four dimensions: definitions; central bank mandates, policies and governance; data types and sources; contributions to international initiatives and global forums.

2. DEFINITIONS OF FINANCIAL INCLUSION

Official definitions of financial inclusion are neither widespread nor harmonised across countries. A large majority of reporting central banks do not rely on an official definition. And when there is actually such a definition, in most cases it refers to ***access to financial services***. Access mainly relates to the ability of firms and households to use financial products and services, given in particular the constraints of time and distance. Relevant measures include the proximity of access points, the variety of access channels, as well as socio-economic barriers limiting such use. In a broader sense, the pricing and other terms and conditions of financial products and services can also be relevant factors limiting the scope for access to financial services for segregated groups. Another widely used dimension is the effective use of financial products and services, eg whether deposit accounts, payment services, microcredit schemes and insurance products are actually used by the population. Measuring this concept is done by looking at the observed consumption of financial products, their usage patterns and customer behaviour.

Most approaches consider both ***supply and quality indicators***. Supply (or availability) of financial products and services relates to the various types of such products and services offered to potential customers. For instance, one will focus on the limited number and type of savings products, credit, payment and insurance services offered by various providers to financially excluded groups. Pricing and other terms and conditions can also be captured to complete this assessment, in particular to ascertain whether the targeted populations can afford the products on offer. Also monitored are the various constraints on the supply of financial products and services, which may include administrative/regulatory prescriptions, a lack of interest on the part of providers in serving certain customer segments, business models, unaffordable costs and inadequate product design. Appropriateness and suitability to users' needs are also an important aspect of the quality dimension of financial services. Various concepts can be mobilised in this perspective, such as the pricing of products (is it commensurate with risk?), their convenience and security, the quality of customer relationship management and the degree of consumer protection. Other dimensions also viewed as important relate to financial literacy – which encompasses people's knowledge about financial concepts, inflation and investment risk, their financial numeracy and behaviour regarding money management and savings, as well as their awareness of financial products and services – and the SME financing dimension of financial inclusion related to small and medium-sized enterprises (SMEs).

In contrast, the dimension of ***financial infrastructure*** is more rarely taken into consideration. This dimension relates to the various elements that support the functioning of the financial system. For instance, robust, safe, efficient and widely accessible information and communication technology infrastructure is a key factor underpinning the provision of transaction account services and broader financial products (CPMI (2016)). The quality of infrastructure can be determined by various logistical, geographical, political and environmental, as well as legal factors. Logistical factors would cover technical reliability, such as the error rate on executing payment orders, and how failed orders are handled or corrected. Geographical, environmental and political factors can also play an important role: for instance, national policies regarding regional autonomy may determine the extent to which far-flung regions of a country are within easy reach of telecommunications or other infrastructure networks. Last, the legal aspects of financial infrastructure relate to the ease with which financial claims can be enforced in court. In the area of payment services, a sound legal infrastructure should include a user-friendly and effective recourse and dispute resolution mechanism to address consumer claims and complaints.

3. CENTRAL BANK MANDATES, GOVERNANCE STRUCTURES, CONTRIBUTIONS AND OBJECTIVES IN THE AREA OF FINANCIAL INCLUSION

Accountability is a key element for consideration in the area of financial inclusion, especially as regards the conduct of related public policies (Gadanecz and Tissot (2015)). Indeed, most central banks perform some direct or indirect reporting of their financial inclusion activities. Another important governance aspect is the organisation of the central bank financial inclusion function itself. Activities related to financial inclusion are often decentralised, i.e. several departments and units deal with different aspects of it. Nevertheless, authorities try to ensure a certain degree of policy consistency even when various units are involved. Some countries have set up some sort of collegial structure, eg a formal committee in charge of financial inclusion issues located within the central bank, or a committee comprising representatives of the central bank and other institutions interested in financial inclusion. Not only can such a collegial structure serve as a useful coordination vehicle among the different internal entities working on financial inclusion; it can also help to provide the necessary impetus and buy-in from various stakeholders, including central banks' senior management.

More than half of the central banks surveyed have some sort of ***mandate to focus on financial inclusion***. For those countries where there is no explicit national financial inclusion strategy, it is general felt that there is a need to have one. Nevertheless, in practice central banks can intervene at various levels of the financial inclusion policy agenda. They can in particular directly contribute to it in three major ways:

(i) by ***promoting financial education***: monetary authorities have often a mandate to promote financial education and literacy as well as consumer protection. They can, for instance, achieve this by publishing financial literacy standards, together with clear information that serves to protect consumers of financial services. That, in itself, is a key element for supporting financial inclusion. As individuals increase their financial literacy, they gain knowledge of the benefits of adopting transaction accounts, using those accounts effectively for payments and storing value, and for accessing other financial services;

(ii) by ***acting as financial supervisors*** and overseers: central banks are often involved in the supervision and oversight of financial services, products, institutions and/or payment systems. This supervisory role can contribute to financial inclusion to a great extent, notably by fostering a clear framework for delivering financial services and ensuring sound market practices. On example is the mobile payments system in Kenya, which can serve as a gateway to other financial services and enhance financial inclusion. Yet another illustration is the promotion by a number of central banks of innovative payment and remittance mechanisms; this is often seen as instrumental in facilitating access to, and reducing the costs of, payment and settlement services; and

(iii) by supporting ad hoc ***initiatives targeted at financially excluded population segments***: some central banks are themselves actively involved in facilitating the delivery of financial services to the population. For instance, they promote microfinance programmes and/or help provide subsidised funding to commercial banks to support their lending to priority borrower groups. Such activities are usually conducted with a view to fostering economic growth and reducing poverty more generally; they also contribute to reducing inequalities in accessing financial services as well.

In a more indirect way, a large part central banks see their ***financial stability role*** as their most important contribution to financial inclusion. This role often requires central banks to work on promoting sound and efficient payment systems, improving the functioning of the financial system and protecting consumers and users of financial services. Indeed, central bank efforts to safeguard financial stability can potentially contribute significantly to financial inclusion. A smoothly functioning, efficient and stable financial system is likelier to engage with financially excluded households or firms than a system where financial instability or stress prevails. Likewise, financially excluded parties are likelier to access formal financial services when a minimum level of consumer protection is offered. In turn, a higher level of financial inclusion is beneficial not only for those directly affected, but also for the national payments infrastructure, financial system and ultimately, the economy (CPMI (2016)). Thus, a virtuous circle can be created.

Certainly, a dilemma may exist between financial inclusion and financial stability (Khan (2011). One view is that greater financial inclusion enhances financial stability, for instance, by offering banks a means of diversifying their loan portfolios, or providing them with a more stable source of funding. Another, opposite view, is that improving financial inclusion can be detrimental for financial stability if, say, expanding the pool of borrowers lowers lending standards. This dilemma figures prominently on the current work agenda of international SSBs. Interestingly, the 2015 IFC survey suggested that central banks see little or no conflict at all in reconciling their "traditional" policy objectives on the one hand, and the promotion of financial inclusion on the other.

One important factor facilitating financial inclusion is the ***traditional monetary stability objective*** of central banks. Price stability is seen as instrumental in anchoring inflation expectations, in turn allowing individuals to make better-informed saving and investment decisions. As such, this is likely to make households and SMEs avail themselves of financial services to a greater extent than they would otherwise do, alleviating financial exclusion. In turn, monetary policy tools can become more effective. For instance, when the saving, spending and investment decisions of households and firms are influenced by banks' interest rates, policy rates can be transmitted more broadly to the economy.

In working towards improving financial inclusion, what are central banks' operational objectives?

Interestingly, their primary focus is on improving financial literacy, together with the broad demand- and supply-side aspects of financial inclusion. Another important area for action is promoting the quality of financial services (appropriateness to customers' needs). They can do so by ensuring the effective use of available financial services, and implementing proportionate risk-based regulation (Muhammad (2015)). This objective aims to calibrate the intensity of regulation to the risk profile and systemic importance of the products, services, channels and/or institutions being regulated and supervised (see BCBS (2015)). The idea is to refine the regulatory approach so as to address the wide spectrum of institutions being supervised – say, by treating large internationally active banks differently from the small, non-complex deposit-taking institutions that are key to supporting financial inclusion.

4. HOW WELL DO CURRENTLY AVAILABLE DATA ALLOW FINANCIAL INCLUSION TO BE MEASURED?

Data on financial inclusion are already widely collected (Tissot (2015)). In the vast majority of countries that collect such statistics, the central bank or the monetary authority is primarily responsible for this task. Data can also come from other sources, such as the supervisory authority, the statistical office, various ministries, or private organisations such as the Bill & Melinda Gates Foundation. Not surprisingly, the most widely available data are supply side indicators (eg availability and access to financial services), which are relatively well covered. Most of these data come from administrative, regulatory and supervisory sources. Typically, as part of their oversight mandates, central banks and supervisory authorities have access to information on financial institutions' supply of services to specific segments of the population. As overseers or operators of national payment systems, they often also have payment data at their disposal. Turning to the demand and use of financial services, data are significantly less available, with surveys being the main data source. In this context, the main shortcoming is that the quality of micro data to assess the number of account holders is often compromised by double-counting. Indeed, it is common practice to count the number of accounts, rather than the number of (different) account holders. Moreover, the cost of demand-side surveys is often high, and response rates and/or response incentives can be low.

As regards the main data gaps, information is often insufficiently available to assess the implementation of financial inclusion policies, in particular to measure the related welfare benefits. Existing data collection frameworks are rarely designed to directly assess whether policy targets in the area of financial inclusion are being met. Data are also scarce regarding the quality of financial services and of financial infrastructure. In any case, the availability of financial inclusion data differs for various segments of financial services and products. In general, the activities of commercial banks as suppliers of financial services are relatively well covered. Perhaps more surprisingly, supply-side indicators related to state-owned banks, specialised financial institutions and even more so post offices are significantly less available. Turning to alternative financial providers such as non-bank financial intermediaries, cooperative and charitable organisations, lending and savings associations and bureaux de change, information is even scarcer. Last, little indicators appear available on the supply of financial services by telecom firms. Existing data collection systems may have difficulties in capturing new non-bank financial intermediaries or to cope with alternative new data sources (eg big data) that are becoming available to measure financial inclusion.

A large number of central banks produce regional or sectoral aggregated indicators of financial inclusion. There has been a public debate about whether countries should produce specific indices synthesising the various dimensions of financial inclusion, such as access and usage. Such indices do have analytical merits, for instance by providing a useful way of synthesising various aspects of the topic for analytical purposes. A single indicator capable of charting general trends and facilitating comparison between geographic units can prove very useful for mapping policy progress in the area of financial inclusion for identifying advances and/or barriers. In some countries, a synthetic index may be used in the context of the national strategy for financial inclusion.

Despite the advantages of financial inclusion indices, the views are mixed. While several aspects of financial inclusion (availability, access, usage) are usually incorporated into these indices, important dimensions are often missing: e.g data on the quality of services, SME financing, financial literacy and the financial infrastructure are often lacking. Moreover, indices can appear to be less meaningful than a dashboard of individual indicators, which may in addition be easier to communicate to the public than a composite index. Last, indices may also suffer from poor cross-country and temporal comparability.

A last important point is that financial inclusion data collection process is generally not sufficiently integrated with national statistical frameworks. Perhaps more worryingly, the various indicators are typically not collected by a single independent authority. Not having data collated in one place may be a corollary of the decentralised fashion in which most countries conduct financial inclusion work. But this may result in inefficient or incomplete data access and could limit the scope of any analytical work. This puts a premium on adequate data-sharing arrangements within countries. Other data collection challenges pertain to the cost and non-response aspects of demand-side surveys. Often, the survey population has insufficient incentives to respond, and data may suffer from double-counting. Conducting a cost-benefit analysis before launching such surveys was thus seen as important. Turning to supply side data, they fail to adequately cover alternative providers.

Concerning the data, there are insufficient data available to assess the implementation of financial inclusion policies. This is compounded by the fact that the data collection framework is very rarely designed to assess whether policy targets in the area of financial inclusion are being met. In other words, more attention could be paid to the need to adequately measure the effects of measures or policies implemented to develop financial inclusion, and their impact on general welfare.

The lack of granularity, low frequency and – to a lesser extent – confidentiality restrictions are additional difficulties. As regards granularity, having more breakdowns of the data is an important objective, as it would allow central banks to conduct more detailed regional analysis and to better measure correlations between the various dimensions of financial inclusion. Meanwhile, financial inclusion data are often collected too infrequently to allow policymakers to make an adequate and timely assessment of the impact of their actions. Turning to the issue of confidentiality, it is particularly important to be able to guarantee to households and firms that respond to demand-side surveys that their answers will be protected and that surveying authorities will strictly adhere to their commitments. Confidentiality issues may also arise for the suppliers of services to the financially excluded, when their (restricted) supervisory data are mobilised for financial inclusion policies.

5. INTERNATIONAL COLLABORATION

The general view is that important benefits can be derived from collaboration with international groupings in the area of financial inclusion. A significant part of central banks already contribute to international data collection frameworks and initiatives, such as those of the World Bank (Global Findex Initiative), the OECD (International Network on Financial Education), the IMF (Financial Access Survey), the Alliance for Financial Inclusion, the Global Partnership for Financial Inclusion and the Consultative Group to Assist the Poor. In addition, they often participate in the financial inclusion-related activities of regional organisations.

The benefits of engaging with these international forums can be manifold. They include access to other countries' or cross-country data; sharing experience in developing related methodologies, concepts and survey questionnaires; forming partnerships; as well as benefiting from capacity-building and technical assistance in the area of financial inclusion. International collaboration also provides exposure to benchmarks of best practices. Another is that sharing data on financial inclusion brings clear benefits. When data is shared between different authorities of the same country, it can be leveraged to conduct financial inclusion policies, especially when those policies are a coordinated effort between different administrations. International data-sharing also allows the full benefits of international collaboration on financial inclusion to be reaped. In addition, internationally harmonised data on financial inclusion can be a key input for national policymakers when they benchmark themselves on assessing financial inclusion as well as on developing, designing, implementing and reviewing policies to improve it. Last, harmonised cross-county data are essential to conduct meaningful analyses and to support global public and private organisations working on financial inclusion to better target their resources geographically.

However, a general view is that definitions and measures of financial inclusion are currently not well enough harmonised to allow such effective cross-country comparisons. Coming up with a uniform set of indicators across countries for the various dimensions of financial inclusion is thus a key challenge.

Yet, each country's experience with financial inclusion is determined by its own domestic circumstances. These include the level of economic development, the specificities of its financial system, the relative importance of the agricultural, manufacturing and service sectors that are the mainstays of the economy, economic inequalities, as well as social, demographic and cultural factors. In this sense, there can be no "one size fits all" measure to gauge financial inclusion universally. In particular, the nature of financial exclusion issues faced by advanced economies may significantly differ from those experienced by emerging markets.

To be sure, there is a clear trade-off between international data harmonisation and the need to adequately capture local characteristics. The implication is that further harmonisation of definitions and measures of financial inclusion is warranted but should not happen at the expense of accounting for national specificities. Dealing with this trade-off is certainly no easy task. One way forward is to leverage existing international collaboration efforts as a good way to further coordinate the development of financial inclusion measures that are both comparable internationally and meaningful domestically. Of particular relevance is the work of the Alliance for Financial Inclusion, the Global Partnership for Financial Inclusion, the OECD/INFE project and the World Bank's Findex initiative. Central banks also see a need to deepen international outreach, technical assistance, knowledge-sharing, financial support and capacity-building; this puts a premium on the work of international forums in general, and of the IFC in particular. In any case, and as acknowledged by the G20, international collaboration can serve as a useful impetus for the advancement of policy for financial inclusion.

6. CONCLUSIONS

The IFC 2015 survey highlighted six key messages related to financial inclusion policies and practices:

1) Definitions. There is no standard, universally accepted definition of financial inclusion. Central banks that do not currently use an official definition of financial inclusion should consider the merits and drawbacks of having one.
2) Central bank contributions. Most central banks have some form of direct or indirect remit to promote financial inclusion. But first and foremost, they make indirect contributions to financial inclusion by pursuing their traditional objectives of price and financial stability. The central bank community could thus clarify and communicate more on their contributions to improving financial inclusion by pursuing their traditional policy objectives and should consider the pros and cons of having an explicit financial inclusion mandate from this perspective.
3) Internal coordination. Although most central banks report formally or informally on their activities in the area of financial inclusion, operations in this domain are often decentralised. Central banks should address logistical and organisational challenges to enhance the internal coordination of financial inclusion work. This effort should cover the various functional areas in charge of monitoring and implementing financial inclusion policies, and in particular also data-gathering and analysis. Adequate governance structures may have to be put in place, especially when units are involved both within and outside the central bank.
4) Data collection. Although data on financial inclusion are widely collected, significant gaps still exist. Work to improve data availability in the following areas would be desirable: (i) data on access, usage and quality of financial services and of financial infrastructure, on SME financing and on non-bank financial service providers; and (ii) data that allow policy implementation to be directly assessed.
5) International cooperation. Collaboration between central banks and interaction with international groupings should be further enhanced, especially in order to favour an effective exchange of views and best practices when defining, measuring and analysing financial inclusion.
6) International data-sharing. Ongoing efforts to internationally harmonise definitions and measurements relating to financial inclusion should be encouraged, not least to facilitate data-sharing. However, greater harmonisation should leave room for capturing country specificities and should primarily leverage existing international collaboration initiatives.

REFERENCES

Basel Committee on Banking Supervision (BCBS) (2015): "Range of practice in the regulation and supervision of institutions relevant to financial inclusion", consultative paper, January.

Committee on Payments and Market Infrastructures (CPMI) (2016): "Payment aspects of financial inclusion", April.

Gadanecz, B and B Tissot (2015): "Key messages of the Sasana Workshop on Financial Inclusion Indicators – Promoting financial inclusion through better data", IFC Bulletin, no 38, February.

Irving Fisher Committee (2016): "Measures of financial inclusion – a central bank perspective", *IFC Report*, June.

Khan, H (2011): "Financial inclusion and financial stability: are they two sides of the same coin?", Address by Shri H R Khan, Deputy Governor of the Reserve Bank of India, at BANCON 2011, organised by the Indian Bankers Association and Indian Overseas Bank, Chennai, 4 November.

Mehrortra, A and J Yetman (2015): "Financial inclusion – issues for central banks", BIS Quarterly Review, March.

Muhammad, I (2015): "Towards proportionality in practice – financial inclusion and implementation of global standards", Opening remarks by Mr Muhammad bin Ibrahim, Deputy Governor of the Central Bank of Malaysia (Central Bank of Malaysia), at the Global Symposium on Proportionality in Practice, Kuala Lumpur, 28 May 2015, www.bis.org/review/r150603e.htm.

Tissot, B (2015): "Statistical challenges related to financial inclusion", paper presented at the 60th congress of the International Statistical Institute in Rio, http://unstats.un.org/unsd/accsub-public/Session-ISI/Documents%202015/Paper-BIS-Tissot.pdf.

Financial Inclusion and the G20 Agenda

Beatrice Timmermann*
Deutsche Bundesbank, Frankfurt am Main, Germany – beatrice.timmermann@bundesbank.de

Dr. Philipp Gmehling
Deutsche Bundesbank, Frankfurt am Main, Germany – philipp.gmehling@bundesbank.de

Disclaimer
This article represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Abstract

The following article provides a short overview on the general development of financial inclusion as a major part of the G20 agenda. Furthermore the agenda of the German G20 Presidency 2017 in advancing financial inclusion is described in more detail, highlighting how the specific workstreams may contribute to progress to the overarching priorities of the German Presidency.

Keywords: Financial inclusion; Digitisation of finance; G20.

1. INTRODUCTION

The G20 is a group comprised of the 20 most important industrial and emerging economies, serving as a central forum for international cooperation in economic and financial policies. It aims at strengthening and broader sharing the benefits of globalisation. With regard to these goals the term "financial inclusion" describes the pursuit of providing useful and affordable access to financial services to all individuals and businesses worldwide. As such, financial inclusion may contribute to economic growth and employment, may reduce inequality and could potentially increase financial stability.

Since the 2009 Pittsburgh Summit, financial inclusion plays a major role on the G20 agenda. The G20 Leaders recognised financial inclusion as one of the main pillars of the global development agenda at the G20 Summit in Seoul (2010) and endorsed a Financial Inclusion Action Plan (FIAP) and principles for innovative financial inclusion. Subsequent to the Seoul Summit, the G20 launched the Global Partnership for Financial Inclusion (GPFI) in December 2010. The GPFI is the main coordinating and implementing mechanism for the FIAP and serves as an inclusive platform for G20 countries, non-G20 countries and relevant stakeholders for peer learning, knowledge sharing, policy advocacy and coordination. In particular, the GPFI helps countries implementing the G20 principles for innovative financial inclusion and aims at strengthening data for measuring financial inclusion.

During previous G20 presidencies additional focus was put on digital financial inclusion, acknowledging that digitisation can be one of the most important factors for increasing inclusive finance by leveraging the opportunities provided by new technologies with regard to financial services. During the Chinese Presidency in 2016 the G20 approved the High Level Principles for Digital Financial Inclusion (HLP). The G20 member countries committed themselves to taking concrete actions to promote digital financial inclusion at their own country level, based on the specific circumstance of each country.

2. FINANCIAL INCLUSION PRIORITIES OF THE GERMAN G20 PRESIDENCY IN 2017

The three pillars of the German G20 Presidency in 2017 are: building resilience; improving sustainability; and assuming responsibility. For the G20 finance track the following priorities have been defined: enhancing economic resilience; improving investment conditions: partnership with Africa; shaping digitalisation. Guided by these overriding goals, the focal areas on financial inclusion for the GPFI during the German Presidency are:

a. Review and Update of the Financial Inclusion Action Plan
b. Implementation of the High Level Principles for Digital Financial Inclusion
c. Leverage relevance of and increasing Financial Literacy
d. Implementation of the G20 Action Plan on SME Financing
e. Financial Inclusion of Vulnerable Groups: Forcibly Displaced Persons
f. Further work on improving conditions for remittances

Review and Update of the Financial Inclusion Action Plan

In order to maintain relevance and to reflect the proceeding rapid digitalisation of financial services a review and update of the FIAP will be conducted in 2017. Therefore, amongst other goals, the review aims at addressing the opportunities and challenges for financial inclusion through advances in digitisation. Additionally, the objective of the review is to align the FIAP with other development goals such as the UN 2030 Agenda for Sustainable Development and the Addis Ababa Action Agenda.

Implementation of the High Level Principles for Digital Financial Inclusion

Especially in the financial sector, digitisation is advancing rapidly and thus enabling new innovative business models. Hence, under the priority "shaping digitalisation" for the German G20 Presidency the GPFI will take key first steps in implementing the High-Level Principles for Digital Financial Inclusion adopted during the G20 Hangzhou Summit in 2016. This includes a peer exchange on emerging approaches to digital financial inclusion. Peer exchange may help to identify practices of different groups of users, types of services and technologies that may again vary according to specific countries' circumstances.

A review requested from the World Bank Group on emerging practices at the country level that bring the HLPs to life, by illustrating the HLP actions through real life examples, could be used for other countries as a basis for better shaping their national implementation strategies. The report will show country examples as role models how to deal with legal and regulatory challenges, coordination among public and private sector entities and infrastructure issues. This approach takes into account that adjustment to country specific needs is key and the promotion of a global inclusion strategy would not serve the need of all countries.

Further member-driven activities, such as for example installing a digital platform for a continuous peer-exchange among countries' authorities involved in enhancing digital financial inclusion (e.g., central banks, supervisory authorities, regulators), could also be conducive to promote implementation of the HLPs at the country level.

Leverage relevance of and increasing Financial Literacy

A prerequisite for successful and sustainable financial inclusion is that individuals have the ability to use financial services and are aware of the associated financial risks. In order to make sound financial decisions and ultimately achieve individual financial wellbeing a certain degree of financial literacy is therefore required. Increased sophistication of financial markets and expansion in access to financial services – among else through digitisation – requires consumers and investors to keep pace and acquire additional skills and competencies.

Knowing what works in financial literacy helps to optimise implementation of effective financial education strategies. A survey on the level of financial literacy and financial inclusion on the basis of the OECD/International Network for Financial Education (INFE) toolkit is being conducted by G20 countries and will provide comparable information on financial knowledge, attitude and behaviour for participating countries. The survey results will be summarised in a report of the OECD/INFE which serves as a starting point to foster research on the impact and effectiveness of financial literacy programmes, underlining the significance of policy evaluation in general. To further intertwine policy guidance and research the OECD/INFE is working on establishing a global database of evaluated financial literacy initiatives.

A further OECD/INFE report on "ensuring financial education and financial consumer protection for all in the digital age" examines the demand-side opportunities and challenges resulting from the digitisation of finance. The report supports implementation of the HLPs for digital financial inclusion, in particular principles 5 and 6 on consumer protection and on financial education, by exploring how financial consumer protection and policy measures for financial education can support the development of inclusive digital finance. It also highlights concrete examples of the use of digital tools to deliver financial education and enhance overall financial decision-making.

Implementation of the G20 Action Plan on SME Financing

The G20 Action Plan on SME Financing was adopted during the Turkish Presidency in 2015. Under the German Presidency there will be a self-assessment by all G20 countries regarding the status of frame conditions for SME financing. This will be consolidated into a baseline report by mid 2017 by the Presidency, which then is to be updated every two years. Furthermore, the action plan highlights the need to further analyse the impact of innovative financing instruments for SMEs, which is done by a GPFI stock-tacking report on "Alternative Data Transforming SME Finance". The report will highlight the digital opportunities of closing the financing gap for SMEs which is prevailing in many countries and thus limits their potential contribution to driving growth and employment. Moreover the action plan aims to close data gaps on SMEs.

In light of the limitations caused by the tremendous financing gap for SMEs, the GPFI also aims at conducting further research on how to foster participation of SMEs in sustainable global value chains through better access to financial services.

In addition to that, as part of a GPFI working topic on access to financing for rural SMEs, the GPFI SME Finance subgroup will look at the question on how resilience against Climate-Related Risks, in particular for rural SMEs, can be increased.

Financial Inclusion of Vulnerable Groups: Forcibly Displaced Persons

For forcibly displaced persons, access to affordable, secure and quickly usable financial services (financial inclusion) is often unavailable or inadequate in host countries; this is especially true for women. Therefore, the German Presidency aims to define fields of action on the basis of available surveys and to draft policy recommendations in consultation with non-G20 countries. Furthermore, a report by the Alliance for Financial Inclusion (AFI) will examine current regulatory practice including the role of the central bank or the financial regulator and other sectorial actors, and will consider related challenges and how these are being/could best be addressed.

Further work on improving conditions for remittances

Remittances are a key source of income for millions of households and businesses. The G20 commits to continue its efforts to reduce the costs for money transfers, thereby supporting the goals of the UN 2030 Agenda. Under the German Presidency the GPFI will continue its various efforts in regard to remittances, including efforts to strengthen competition, making use of new digital technologies, improving transaction infrastructure or strengthening customer protection.

3. CONCLUSIONS

According to the Global Financial Inclusion Database by the World Bank there are still more than 2 billion people and about 300 million companies worldwide without any access to formal financial services. The priorities of the German G20 Presidency with regard to financial inclusion outlined here address concrete actions in order to further improve the degree of financial inclusion and foster inclusive growth.

Measuring Financial Inclusion in Malaysia: *Unlocking Shared Benefits for All through Inclusive Finance*

Zarina Abd Rahman

Bank Negara Malaysia, Kuala Lumpur, Malaysia – zar@bnm.gov.my

1. INTRODUCTION

Promoting inclusive finance, where all segments of society have access to suitable and affordable formal financial services, is a key focus and specific mandate of Bank Negara Malaysia (the Bank) in contributing towards equitable and sustainable growth. Essential financial services provide equal opportunities for all Malaysians, including the lower income segments of society, to safely save and invest, borrow for productive activities and buffer themselves against unforeseen shocks. In advancing the financial inclusion agenda, the Bank in 2011 introduced the Financial Inclusion Framework, a comprehensive plan outlining the strategies for an inclusive financial system over the coming decade. Subsequently, the Financial Inclusion Index was developed to track the progress and impact of the Bank's financial inclusion policies. This article reports on the progress of financial inclusion in Malaysia since 2011 when the first Financial Inclusion Demand-Side Survey was conducted.

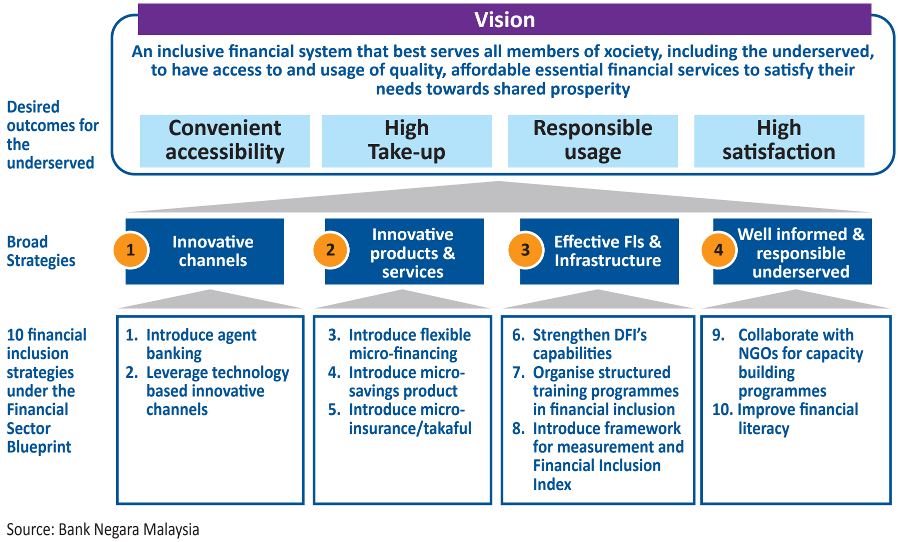


Diagram 1. Financial Inclusion Framework

2. SIGNIFICANT ACHIEVEMENTS IN FINANCIAL INCLUSION

The Financial Inclusion Index score for the general population in Malaysia improved significantly to 0.90 in 2015 from 0.77 in 2011 (where 1.00 reflects full inclusion). These improvements were driven by increased accessibility to financial access points across the country, more responsible usage of products and higher levels of satisfaction among financial consumers. Meanwhile, gaps continue to be observed in the take-up level of financial products and services, particularly among low-income households.

(i) Convenient Accessibility

Convenient accessibility, which measures the availability of financial access points at the district (daerah) and sub-district (mukim) levels, recorded a marked improvement. All 144 districts and 97% (2011: 46%) of the 886 sub-districts with a population of at least 2,000 now have access to essential financial services. This expansion in the number of access points nationwide provides 99% (2011: 82%) of Malaysians with convenient access to safe, reliable and affordable financial services.

This achievement was in large part due to the establishment of agent banks, which had an important impact in increasing access to financial services particularly in the rural areas. Following the operation of agent banks, the volume of financial transactions conducted through agent banks has increased from three million transactions in 2012 to RM63 milion in 2015 (amounting to RM5.7 billion in value)

Financial Inclusion Index

The Financial Inclusion Index measures the level of financial inclusion in Malaysia and the effectiveness of the Bank's policies in achieving four desired outcomes of financial inclusion: (i) convenient accessibility; (ii) high take-up; (iii) responsible usage; and (iv) high satisfaction.

The Financial Inclusion Index is constructed from both supply-side data from financial institutions and demand-side data collected through the Financial Inclusion Demand-Side Survey (Survey), conducted periodically by the Bank.

Advancement of Agent Banking

Agent banking enables consumers to obtain banking services by licensed financial institutions through third-party agents such as retail outlets and post offices. First introduced in 2012, the agent banking regulatory framework was further enhanced in April 2015 to allow agents to facilitate the opening of saving accounts[1] on behalf of financial institutions via online real-time systems and biometric identity verification.

As at end-2015, 6,902 agent banks have been established nationwide, with over 13,600 new accounts opened and 63 million transactions amounting to RM5.7 billion facilitated by agent banks. Most of these transactions involved bill payments (59.2%; RM3.4 billion) and cash deposits (28.1%; RM1.6 billion).

Another important development that has intensified since 2011 has been the expansion of Internet banking and mobile banking. As at end-2015, the number of Internet banking subscribers increased to 19.8 million (2011: 11.9 million) representing 63.7% of the total population, while the number of mobile banking subscribers increased to 7.3 million (2011: 1.6 million) representing 23.5% of the total population. These digital channels have had an important impact in increasing access to banking services, with greater convenience and flexibility for consumers to keep track of their personal finances.

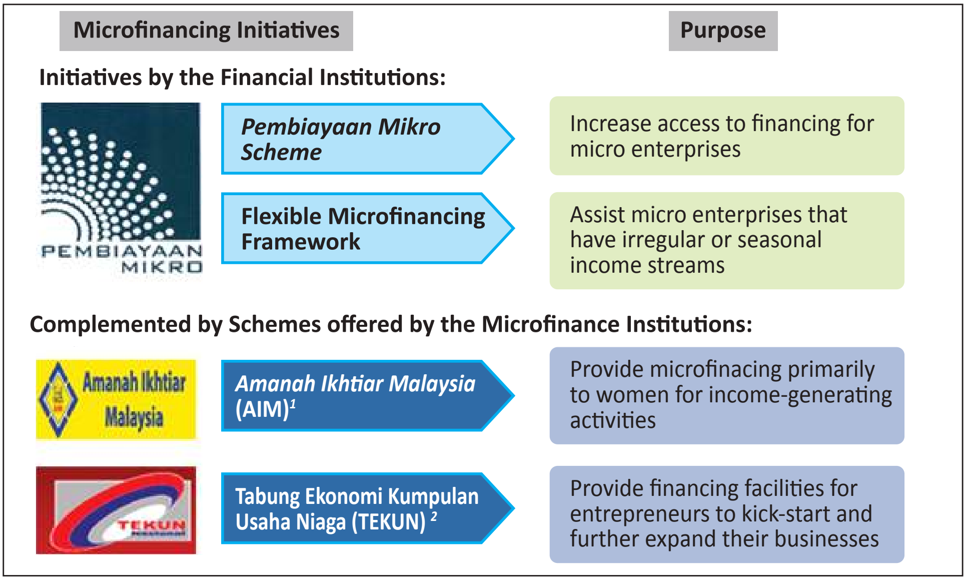
(ii) Take-up of Financial Products and Services

The take-up rate, which represents the population's usage of specific financial services namely deposit accounts, financing accounts and insurance policies, recorded a slight decline. However, while the percentage of adults with deposit accounts remained high at 91%, the percentage of adults with financing

1 In addition to allowable services under the guidelines on Agent Banking (2012) namely accepting deposits, Facilitating withdr awals, fund transfers, bill payments and financing repayments.

accounts had declined from 36% to 25%, contributed mainly by the decrease in credit card ownership. This, in part, reflects improved debt management practices and affordability assessments following the introduction of a number of measures by the Government and the Bank, such as the credit card service tax, the Credit Card Guidelines and the Guidelines on Responsible Financing. Meanwhile, the percentage of adults surveyed who indicated that they purchased a life insurance or takaful policy moderated from 18% to 16%. These findings show that while financial inclusion has increased significantly since 2011, certain gaps remain, particularly among the low-income segment, where affordability remains a challenge.

In recognition of these gaps and opportunities, the Bank has undertaken various initiatives to encourage the development of products and services that are targeted towards this segment. This has included encouraging financial institutions to offer microsavings products with low committed periodical savings to encourage regular savings among the low-income population; facilitating the offering of microinsurance/microtakaful products by insurance companies to provide financial protection against unexpected adverse events; and supporting the provision of tailored microfinancing solutions for micro, small and medium enterprises (SMEs).

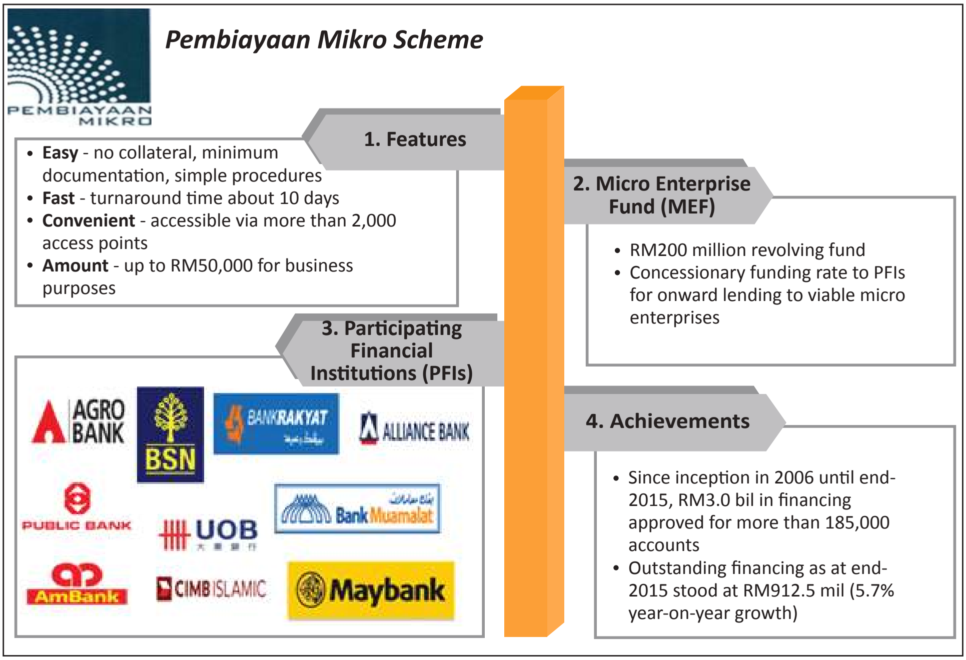


1 AIM was established in 1987 with the objective of assisting the hard-core poor to rise out of the poverty trap. This is done through the provision of microfinancing through more 130 branches nationwide.

2 TEKUN Nasional was established in 1998 and since a rebranding exercise in 2008, it has played a role that is more than just being a financial provider agency but also provides entrepreneurship development and support services. TEKUN has more than 190 branches to-date.

Diagram 2. Targeted Microfinancing Solutions

The take-up of financial services is also supported by a comprehensive financing ecosystem comprising: (i) an enabling financial infrastructure; (ii) financing and guarantee schemes; (iii) avenues to seek information and redress; (iv) debt resolution and management arrangements; and (v) outreach and awareness programmes. These arrangements have been important to promote confidence and reduce anxiety in using financial services, in addition to supporting individuals and businesses to understand how the financial system can help them manage risks and improve their financial well-being.



Source: Bank Negara Malaysia

Diagram 3. Micro Financing Scheme

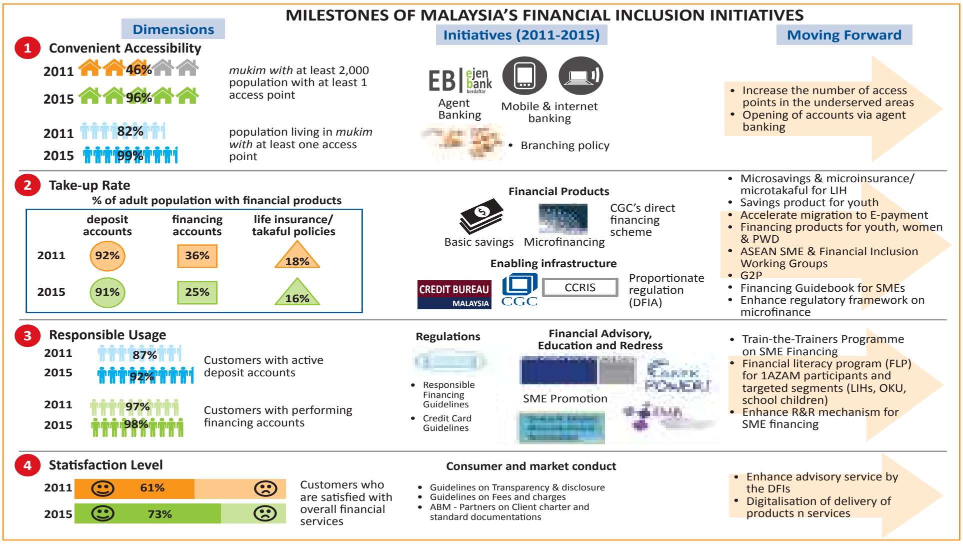
(iii) Responsible Usage

Responsible usage, which measures whether financial products are utilised appropriately, remained high. The percentage of banking customers with active deposit accounts, an indication that they are saving regularly, has increased to 92% from 87% in 2011 while the percentage of banking customers with performing financing accounts increased to 98% from 97% over the same period.

Financial education initiatives have had a key role in encouraging responsible usage by helping consumers make better financial decisions, thus promoting a positive experience from their participation in the financial system. The Bank continues to collaborate with the public and private sectors to organise and implement these financial capability programmes. Meanwhile, agencies such as the Financial Mediation Bureau, the Credit Counselling and Debt Management Agency and the Small Debt Resolution Scheme serve to ensure that financial consumers are able to get the help that they need to effectively manage their financial affairs.

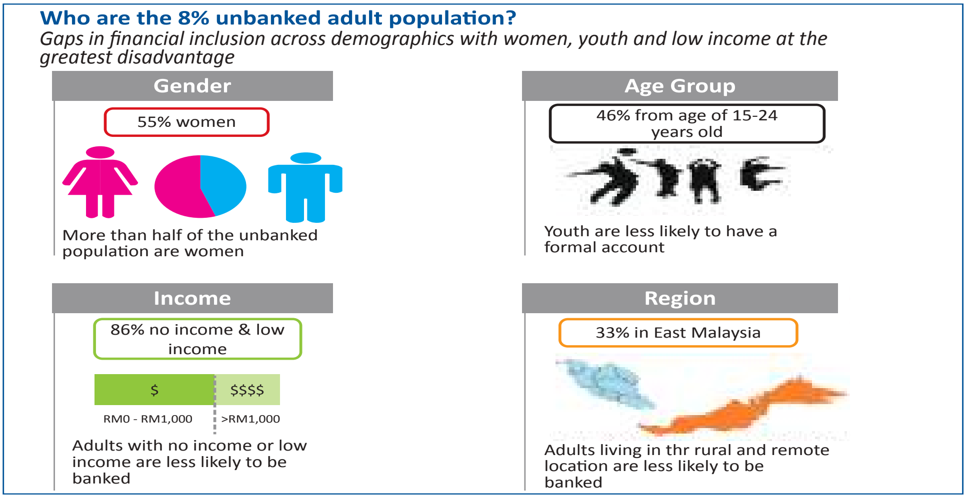
(iv) Satisfaction Level

The index revealed a significant improvement in the level of satisfaction with financial services in Malaysia. The percentage of customers of financial institutions who are satisfied with overall financial services increased to 73% (2011: 61%), with higher satisfaction levels observed across all segments of the population, including low-income households who also recorded an increase in satisfaction to 67% (2011: 60%). This in part reflects improvements in the conduct and services of financial institutions arising from various initiatives to promote a positive experience for all financial customers. Of note were substantially strengthened standards issued by the Bank to regulate product transparency and disclosures, and the imposition of fees and charges by banks. Also notable have been key private sector initiatives such as the PARTNER programme by the banking industry which simplifies documentation and improves the turnaround time for processing SME financing and housing loans. A similar initiative to introduce plain language in insurance contracts has also been pursued by the insurance industry.



Source: Bank Negara Malaysia

Diagram 4. Milestones of Malaysia's Financial Inclusion Initiatives



Source: Bank Negara Malaysia

Diagram 5. Who are the 8% unbanked adult population?

3. CONCLUSION: ELEVATING THE LEVEL OF FINANCIAL INCLUSION MOVING FORWARD

The Bank remains committed to further increase the level of financial inclusion. This will be supported by a sustained focus on encouraging the development of innovative delivery systems, products that are responsive to the needs of the underserved, and effective education, support and protection for financial consumers. With this focus, the Bank expects that the remaining 8% that constitutes the unbanked population in Malaysia will be further reduced to 5% by 2020.

Assessing Financial Inclusion in Portugal from The Central Bank's Perspective[1]

João Cadete de Matos *

Banco de Portugal, Lisbon, Portugal – jcmatos@bportugal.pt

Luís D'Aguiar

Banco de Portugal, Lisbon, Portugal – laguiar@bportugal.pt

Abstract

Access to financial services is quintessential in fostering the development and competitiveness of an economy and in mitigating the asymmetry in the distribution of income across and within countries. The active promotion of financial inclusion entails several direct and indirect improvements on the conditions for value creation of economic agents, particularly small and medium businesses and low income households. These span from the enablement of risk pooling to the possibility of financing ambitious development projects that would be otherwise impossible. To fully rip these benefits, it is crucial that countries develop infrastructures and continuously monitor and reform their institutions, in order to keep up with the latest innovations, hence capturing their induced development effects. At this point, a question arises: how can financial inclusion be measured and accounted for? The paper discusses the evolution of financial services in Portugal and, in particular, the role and contribution of *Banco de Portugal* as a financial services provider – with emphasis in the services offered by its Central Credit Register and within the context of its Banking Conduct Supervision activities, and how they have been used to pre-empt over-indebtedness and monitor the access to financial services, while shedding light into the usage of innovations in the payment systems. Finally, an attempt at documenting and measuring the evolution financial inclusion in Portugal is made on the basis of the results of the *Banco de Portugal*'s Survey on the Financial Literacy of the Portuguese Population.

Keywords: Financial inclusion; Financial services; Payment systems; Central credit register

JEL classification: G20; E42

1. INTRODUCTION

Financial inclusion – loosely defined here as the ability of individuals or enterprises to obtain formal financial services that are appropriate to their needs, including access to credit and bank accounts, payments facilities, insurance, and other risk management services – is essential in fostering the development and competitiveness of an economy and in mitigating the asymmetries in the distribution of income across and within countries.

However, while there has been some progress worldwide toward wider access to formal financial services, significant challenges remain – *e.g.*, more than two billion adults do not have access to formal or semi-formal financial services; they are the financially excluded in a world where access to financial services can mean the difference between surviving or thriving (ATISG, 2010). What is more, financial exclusion seems to be no longer only a phenomenon in developing and emerging countries, especially in the aftermath of the 2007-08 financial crisis.

---

1 The analysis, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the *Banco de Portugal* or of the Eurosystem

Data currently available show that, even in developed countries, many people remain excluded from the financial system, with no access to basic banking products, such as is clearly the case of a current account. According to an European Commission study on financial services provision and prevention of financial exclusion, at the end of 2003, 10 per cent of adults aged 18 and over in the EU15 countries (European Union Members States before May 2004) and 47 per cent of adults in EU10 (Member States which joined the EU in May 2004) had no bank account at all. A further 8 per cent in the EU15 and 6 per cent in the other Member States only had a deposit account with no payment card or cheque book.

Seven per cent of all adults in the EU15 countries and 34 per cent in the other Member States had neither a transaction nor a deposit account, revolving credit or savings (European Commission, 2008a).

Meanwhile, increased awareness of this problem led policymakers, regulators and development agencies globally to consider the promotion of financial inclusion as a priority. The G20, for instance, has identified financial inclusion (for individual households as well as enterprises) as a key driver of economic growth, reduced economic vulnerability for individual household, poverty alleviation, and improved quality of life for people around the world. On the occasion of its Seoul Summit in 2010, a number of initiatives directed to improving access to financial services and expanding opportunities for poor households and micro-, small- and medium-sized enterprises was approved (the Financial Inclusion Action Plan, the Global Partnership for Financial Inclusion and a flexible SME Finance Framework).

This paper describes the provision of formal financial services in Portugal and, in particular, the role and contribution of *Banco de Portugal* as a financial services provider – with emphasis in the services offered by its Central Credit Register and within the context of its Banking Conduct Supervision activities, and how they have been used to pre-empt over-indebtedness and monitor the access to financial services, while shedding light into the usage of innovations in the payment systems.

The focus of the paper will be more on the provision of financial services than in measuring the access to them, in view of the methodological difficulties inherent to the measurement of access to finance as well as the lack of comprehensive and reliable data on, *inter alia*, the people using financial services, the types and quality of services they receive and the price they pay, and the barriers to broader access. That said, the absence of significant non-price barriers for firms and households in the use of financial services in Portugal, on the one hand, and the fact that the provision of financial services may be seen as an indication of the potential access to financial services, on the other hand, give justifying argument for the approach that has been followed.

Notwithstanding, an attempt at documenting and measuring the evolution financial inclusion in Portugal will be made, on the basis of the results of the *Banco de Portugal*'s Survey on the Financial Literacy of the Portuguese Population.

2. THE EVOLUTION FINANCIAL SERVICES PROVISION IN PORTUGAL

The provision of financial services in Portugal has been growing at a fast pace since 1986, the year that marked the Portuguese accession to the European Economic Community (EEC). This trend was fuelled by a substantial increase in the use of technological innovations (mainly in telecommunications), which made it possible to implement teleprocessing networks, either within the larger banks or through interbank links, with visible benefits to the supply of financial services, particularly in the field of retail payments. This increase was backed by the creation of SIBS ("Sociedade Interbancária de Serviços"), a company that was founded in 1983 by a number of resident banks (as of today, the company's shareholders stand for practically the whole retail banking sector in the country). Its aim was to introduce a single payment platform that met the banking establishments' needs while developing their facilities and technology, and extending their international scope. In 1985, this network became operational under the Multibanco designation.

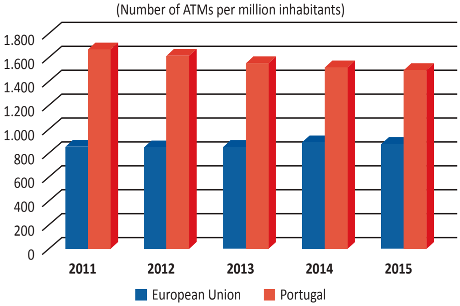
Multibanco, a sophisticated network shared by every bank operating in the economy that fully integrates automated teller machines (ATMs) and electronic funds transfer at point of sale (EFTPOS) terminals, profoundly transformed the way retail payment operations were carried out in Portugal. Since its implementation new features have been constantly added to the system (involving no extra costs to its users): in addition to cash deposits/withdrawals and balance/transactions inquires, customers can undertake a wide range of other operations, *inter alia*: money transfers (both to other customers in the same bank and to other banks), payments for utilities bills, payments to the State and the Social Security, mobile phone top-ups, transport ticketing and event booking and ticketing.

The activity of SIBS was instrumental in generating economies of scale deriving from a more rational and effective use of the financial, technical and human resources needed to develop more advanced payment instruments and systems. Also, SIBS made the Multibanco network open to all those taking part in payment systems operations, which allowed for network economies and the safety inherent in a single system.

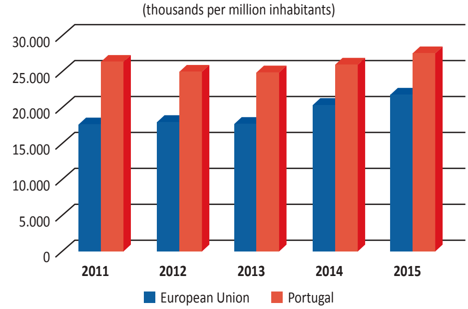
In addition, this scheme has returned to its users a substantial part of the productivity gains generated from on-going technological and organisational developments, both directly, through the supply of a service that is ever wider, of better quality and at a better price, and indirectly, through increasing efficiency in the banking system as a whole.

Quantitative relevance

The Payments and Settlement Systems Statistics dataset of the European Central Bank Statistical Data Warehouse offers ample coverage about the characteristics and the dimension of the Portuguese payments system as well as of the other European Union Member States. Looking into the dataset, one sees, inter alia, that the value of transactions carried through card payments with cards issued by resident Payment Service Providers (PSPs) in Portugal has been growing steadily for the last five years, totaling 65.4 EUR billions in 2015 (55.7 EUR billions in 2011); also, the demographic and geographic coverage of the ATM and POS networks are shown to rank amongst the highest in the European Union and the Euro area.



Graph 2.1. ATM Demographic Coverage



Graph 2.2. ATM Demographic Coverage

The number of ATMs provided by resident PSPs in Portugal reached 15.6 thousands in 2015 – about 1500 per million inhabitants (the 2nd highest demographic coverage in both the European Union and the Euro area) and 169 per thousand Km^2. In addition, the number of POS terminals provided by PSPs in Portugal amounted to 286.4 thousands in 2015 – around 27640 per million people (the 6th highest demographic penetration in the European Union and the 5th in the Euro area) and 3106 per 1000 Km^2. The number of cards with a cash function issued by resident PSPs totalled almost 20 million in 2015 (19 million in 2011). The number of payments with cards issued by PSPs totalled about 1375 million in this year (1237 million in 2011) and the related value of transactions reached approximately Euro 65.4 billion (55.7 EUR billions in 2011).

A *Banco de Portugal*'s study of July 2007 estimates that, during 2005, cash withdrawals from ATMs, rather than at bank counters, saved consumers 11.2 million hours in processing time and banks participating in this study cut costs by 46 per cent. For cash deposits at ATMs, estimated gains were around 443 thousand hours for customers in terms of processing time and cost cuts for the banks to the order of 4.3 EUR millions. With credit transfers through ATMs, gains were estimated to be around 491 thousand hours in processing and 5.1 EUR millions in transaction costs. Consumers also gained around

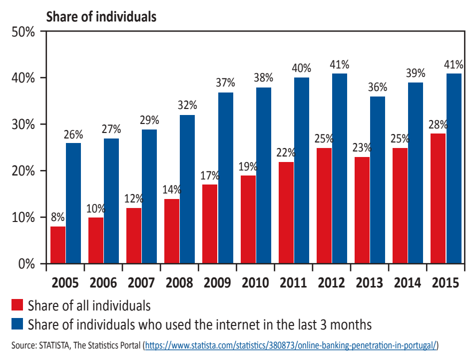
1.5 million hours in checking balances and account entries through ATMs in place of over the counter. Taking gross average salaries as a point of reference (as detailed in statistics for OECD countries in 2005), the benefit for the consumer comes in at around 86 EUR millions, corresponding to 13.6 million hours of processing time.

Home banking and mobile banking in Portugal

The availability of information and communication technologies has also led the Portuguese banks to make considerable efforts to modernise customer access to financial services, in new, cheaper ways, and to more people. As a complement to in-person services, online connections with corporate customers, home banking and mobile banking (m-banking) are now extensively available in Portugal and increasingly used by firms and households.

Home banking has become the third preferred channel to get in touch with a bank – next to ATMs and to face-to-face contact with the bank teller. In 2007, around 19 per cent of those people in Portugal that have access to a wide range of transaction banking services that are appropriate to their needs and socio-economic status (*i.e.*, the "fully banked", following the terminology in the 2008 Commission's report) favoured that channel – more than seven times the percentage in 2000 (2.7 per cent). The rising number of Portuguese households that have at least one computer at home (in the period 2004 – 2008 the average annual growth rate was 6 per cent) and access to a broadband Internet connection (35 per cent) will likely strengthen the use of home-banking as an alternative to traditional outreach.

Information about on-line bank penetration in Portugal (from the *Statista* website) show that, in 2015, 28 per cent of all individuals used the internet for on-line banking, and that usage was higher for those who had used the internet within the last three months, at 41 per cent (see picture below)



Graph 2.3. Online Banking Penetration in Portugal from 2005 to 2015

With m-banking, banks get information to their customers no matter where they are and at reduced costs. This is vital for a number of banking services (*e.g.*, alerts), as well as for traditional marketing campaigns. From the demand-side viewpoint, people often do not have time to get to a bank branch, and the Internet may not be an option in some cases – circumstances that favour the use of a more straightforward channel. For a sizeable part of the Portuguese population, using mobile phones to have access to certain financial services, like checking bank balances or conduct other basic operations, should be like a second nature. The Portuguese are clearly a phone-savvy nation. For many, the mobile phone is the organizational hub of their lives.

In 2008, the number of mobile phone subscribers amounted to almost 15 million, corresponding to a market penetration rate of about 152 per cent[2], one of the highest among the EU-25 Member States and well above the EU-25 average (119 per cent). Also, the total number of possible users of Universal Mobile Telecommunication System services was 4.3 million. (European Commission 2008b).

The continuous expansion of Internet banking and m-banking in Portugal should contribute to improve noticeably the access to financial services, by offering services that are, concurrently, more affordable and more suited to the prospective customers, particularly to that segment of the so-called "marginally banked" (*i.e.*, people with a deposit account that has no electronic payment facilities and no payment card or cheque book) that have a bank account but rarely use the related electronic payment facilities and cards.

In view of the above, it is safe to say that the Portuguese retail payment system is widely recognised today as a highly developed system, in terms of technology, accessibility, time-saving features and nationwide coverage. The system processes millions of operations on a daily basis, both counter-based at thousands of bank branches and electronically through the ATM/POS system. Last but not least, its overall quality can be recognised by what is in relative terms a very small number of complaints on the part of the banks' clientele.

3. THE ROLE OF THE BANCO DE PORTUGAL AS PROVIDER OF FINANCIAL SERVICES

Banco de Portugal is the Portuguese central bank and is an integral part of the European System of Central Banks (ESCB). As such, it pursues the objectives and participates in the performance of the tasks entrusted to the ESCB, particularly the maintenance of price stability, viewed as a pre-condition for increasing economic welfare and the growth potential of an economy. The *Banco de Portugal* is also accountable for the efficient and safe functioning of the country's payment systems – including the issue of banknotes and clearing services –, an essential condition for the sound operation of the economy. In addition, the *Banco de Portugal* provides a wide range of services to the banks – *e.g.*, the running of the Central Credit Register (including the centralization of information on protested bills and on cheque defaulters) – and to the non-financial companies – *e.g.*, the maintenance of the Central Balance-Sheet Data Office. Moreover, the *Banco de Portugal* supervises the resident credit institutions and other financial companies, thus providing for the stability and the soundness of the financial system and ensuring the efficiency of its operation, the safety of deposits and of depositors and the protection of consumers of financial services. Also, Banco *de Portugal* regulates, oversees and sanctions the conduct of credit institutions, financial companies, payment institutions and electronic money institutions offering retail banking products and services. It also promotes the financial information and education of bank customers. Last but not least, another service provided by the *Banco de Portugal* to the community includes the compilation, analysis and dissemination of monetary, financial, exchange and balance-of-payments statistics, which are instrumental to decision-making and, in this way, influence the financial activity in the economy.

Services related with the Central Credit Register

The Central Credit Register (CCR) is a database managed by the *Banco de Portugal* on the basis of credit-related information (including potential liabilities, such as, for instance, unused amounts on credit cards and open credit lines) that is supplied by the participants (all resident credit-granting institutions). The main aim of the CCR is to provide information to back-up the participants in their appraisal of

---

2 According to the Portuguese Authority for Communications (ANACOM), at the end of the third quarter of 2016, the penetration of mobile services was reported at 165.8 per 100 inhabitants. The penetration rate of mobile stations with actual use was 123 per 100 inhabitants. There were 17.1 million mobile stations eligible to use the service. Of these, 12.7 million (74.2 per cent of the total) saw actual use. There were around 6.3 million actual users of services which are typical of mobile broadband (i.e. video telephony, broadband data transmission, mobile TV, etc.), representing 49.2 per cent of mobile stations in actual use.

the risks attached to extending credit. To this end, the participants can assess aggregate information on the credit liabilities of each client *vis-à-vis* the financial system. Any individual person has the right to be informed about the data recorded in his/her name in CCR and, where necessary, ask the participant responsible for the reporting to *Banco de Portugal* to correct and update such information. The CCR database is also used for: (i) the supervision of credit institutions and other financial companies; (ii) the analysis of the stability of the financial system; (iii) monetary policy operations and intra-daily credit; and (ii) the compilation of statistics on, *e.g.*, the distribution of credit by branch of activity. The CCR database fulfils all the requirements for data protection, as laid down by the National Commission for Data Protection. The *Banco de Portugal* is legally authorised to use the CCR information for purposes such as the supervision of financial institutions, the analysis of financial system stability, the conduct of monetary policy and intraday credit operations, and the compilation of statistics.

Banco de Portugal ensures the centralisation and the subsequent dissemination throughout the banking system (generally on a daily basis) of credit incidents (protested bills) submitted to the Notary Public Offices by financial institutions, thus providing the financial institutions with additional means to better evaluate the risks of their active operations. Legally, any citizen or company is entitled to have access to the information on their own situation in the protested bills data base existing at the CCR.

Banco de Portugal is also responsible for checking compliance with the duties assigned to the credit institutions as regards the use of cheques. In particular, *Banco de Portugal* centralises the information reported by credit institutions and discloses through the banking system the list of cheque defaulters. This database is authorised by the National Commission for the Protection of Data, insofar as the legal and constitutional provisions protecting the citizens are observed. All entities concerned have access to the information on their own situations.

Services related with the market conduct supervision of credit institutions

In a market characterized by contractual freedom and financial innovation, it is incumbent upon the *Banco de Portugal* to check for compliance with the minimum requirements of information to customers on the financial conditions applied to the different operations and services, as well as on the respective risks, a mission that will be referred here as banking conduct supervision, as opposed to prudential supervision, which is more focused on guaranteeing the soundness of financial undertakings and contributing to the stability of the financial system.

Informed decision-making by the banks' clientele especially aware of the risks inherent to financial products and services is a key requirement to the efficient operation of the retail financial markets and to mitigate the level of risk in the financial system. The disclosure by credit institutions of relevant information concerning their products and services, in a transparent, intelligible and standardised way, promotes such decision-making. However, the dissemination of information along those lines may not be enough, given that clients' decisions are also determined by their level of financial literacy. Therefore, is also necessary to foster financial education among the public at large.

Banco de Portugal's banking conduct supervision is structured on the basis of a number of reciprocally complementing guiding rules, ranging from the requirement for credit institutions to observe the principle of transparency and rigour when informing their clients along the various stages of the marketing of banking products and services, to the development of the normative framework that governs the conduct of credit institutions in the retail financial markets. Concurrently, they include monitoring compliance with regulations – *e.g.*, via surveillance activity related to the commercialization or promotion of financial products and services, by responding to clients' complaints and through comprehensive on-site inspections – as well as fixing cases of non-compliance and, in the most serious situations, applying administrative sanctions.

Another guiding principle consists of promoting the quality of the demand for financial products and services, by fostering initiatives that contribute to raising the clients' competences in assessing costs, expected income and risks related to those products and services.

4. MEASURING FINANCIAL INCLUSION IN PORTUGAL

The international financial crisis has highlighted the importance of financial literacy and informed decision making by bank customers as a form of fostering financial inclusion and improving the efficiency and stability of the financial system. In this context, central banks and financial supervisors have attributed increasing importance to initiatives promoting financial literacy and becoming more involved in the definition and implementation of national strategies on financial education.

The promotion of financial literacy contributes to foster the benefits of the instruments regulating transparency and duties of information of credit institutions and, therefore, to the more efficient functioning of financial markets. Citizens who are better informed have greater capacity to understand the information that is conveyed to them by the credit institutions, helping, thereby, to monitor the markets. By choosing financial products that are suited to their risk profile and financial needs, bank customers allocate their funds in the most efficient manner and contribute to the stability of the financial system.

In recognition of the importance for the citizens of taking informed and careful decisions in the management of their personal finance, the *Banco de Portugal* decided to carry out a Survey on the Financial Literacy of the Portuguese Population, taking into account the principles and best practices adopted internationally. The Survey was structured so as to enable obtaining information about the financial attitudes, behaviour and level of understanding of financial matters by the population. Through the assessment of the various dimensions of the concept of financial literacy, the Survey contributes towards identifying the population groups and financial topics with the most significant gaps in terms of literacy. This represents an important means of diagnosis of the degree of financial literacy of the population and, as such, is an indispensable step towards the definition of financial education priorities.

Encouraging Financial Inclusion in A New Nation – The Experience of The *Central Bank of Timor-Leste*

Gastão de Sousa

Banco Central de Timor-Leste, Dili, Timor-Leste – gastao.desousa@bancocentral.tl

Abstract

Financial inclusion is continuously underlined, in international *fora* and the academia, has essential for the promotion of social and economic progress of a country. Policymakers and regulators are well placed to be key players in shaping and developing an inclusive financial sector. In this context, having recognised the role that the continuous promotion of financial inclusion has in supporting the sustainability of Timor-Leste's growth, *Banco Central de Timor-Leste* ("the Bank") has been playing an instrumental role in measuring the level of financial inclusiveness of the Timorese economy and in fostering several initiatives to encourage and support evidence-based policies in this domain. In this presentation we explore the findings of the Bank's first ever report on financial inclusion – the *Financial Inclusion Report 2016* – and highlight how a new developing country is tackling this reality through the implementation of a set of comprehensive measures, which span from modernizing the payments systems to the promotion of financial literacy.

Keywords: Financial services; Financial literacy; Payment systems; Timor-Leste

JEL classification: G20; O17

1. INTRODUCTION

We are all aware that in the recent years, more and more countries are introducing extensive measures to enhance access to financial services and improve usage of financial services. These comprehensive efforts illustrate each country's commitment to foster financial inclusion as a domestic agenda. Timor- Leste is a fairly new developing nation, and financial inclusion is especially vital to empower its citizens with the tools that allow them to meet their financial needs. The principle of inclusiveness is regarded with the utmost importance, and Banco Central de Timor-Leste (the Bank) is committed to ensure that no Timorese will be left behind as the financial sector develops and grows. The Bank has carried out various action plans to encourage inclusiveness, including ensuring facilitative regulation, modernizing payments systems, developing various financial products, building capacities and financial literacy. Moving forward, the Bank aspires to strengthen the financial inclusion strategies under a holistic national framework, to fill existing gaps and nurture innovation, towards achieving the vision of universal access to financial services in the country.

To promote financial inclusion in the Timor-Leste, the Bank started in 18 July 2016 the report on financial inclusion in Timor-Leste, which was officially released on 13 September 2016.

Financial Service Providers in Timor-Leste and the access point coverage

Despite being a new country, with a financial system under development, Timor-Leste has already 4 commercial banks: one local – BNCTL – and three branches of foreign banks –ANZ Bank (Australia and New Zealand), Mandiri Bank (Indonesia) and BNU/CGD (Portugal). There are also available microfinance institutions, credit unions, saving groups, insurance companies, and money transfer operators.

Access to financial services is a prerequisite for fostering inclusive finance. The Bank has placed significant emphasis on enhancing the outreach of financial services access points through various channels by participating financial institutions. In fact, access point indicators, have been widely regarded as benchmarks to measure financial inclusion. A facilitative financial sector is pivotal to ensure the access to financial services, which has led to a notable progress in the growth of channels such as bank and nonbank access points, mobile banking vans, e-wallet and internet banking.

The abovementioned banking institutions in Timor-Leste play an important role in widening the outreach of financial services access points in the country to enable communities to transact, save and obtain credit.

Table 1. Financial Sector Access Points Coverage

Financial Service Providers	2016: Access Points					
	Branch/ Field Office	ATM	EFTPOS	Meating Center	Other	Total
Banks	33	52	164	235	-	484
MFIs	37	-	-	634	-	671
Credit Unions	-	-	-	-	32	32
Insurance	2	-	-	-	-	2
Others (MTO, CEB)	10	-	-	-	-	16
Total	88	52	164	869	32	1,205

Source: Banco Central de Timor-Leste Financial Inclussion Report 2016

As at June 2016, a total of 484 access points was reached, including bank branches, automated teller machines (ATMs), electronic funds transfer point of sales (EFTPOS) and meeting centres. Of this total, 48% were related to meeting centres, 34% to EFTPOS, while ATMs and bank branches represented 11% and 7% respectively. This clearly illustrates the significance of the deployment of meeting centres and EFTPOS for outreach of financial services in the country.

Banks have shown a remarkable progress in widening outreach of financial services through the years: the total number of access points expanded from 198 in 2009 to 226 in 2013, an increase of 14%. Following that, the number of access points grew significantly from 2013 to June 2016, where the total access points recorded an increase of 114%. Of all the types of access points, EFTPOS displayed the highest percentage of increase from 2009 to 2016, recording a 254% growth, a leap from just 45 access points in 2009 to 164 by June 2016. ATM access points also recorded a substantial increase of 225%, from 16 access points in 2009 to 52 by June 2016.

Non-bank financial service providers have also been instrumental in shaping a more inclusive financial sector in Timor-Leste by complementing the role played by banks. In fact, non-banks promoted 711 access points, which corresponds to 59% of the 1,205 total access points in the country, while banks access points represents the remaining 41%. Non-banks bridge gaps in areas where banks are not able to serve, making it possible to reach more communities with financial products developed to satisfy their financial needs.

Microfinance Institutions (MFIs)[1] in Timor-Leste are made up of 2 key institutions which enhance the financing access to microenterprises. In 2009, these 2 MFIs recorded 702 access points while in 2013, the number increased to 1,005 access points. As at June 2016, the number of access points for MFIs decreased to 671, as one of the MFIs, KIF, changed its strategy from group loans and meeting centers to individual clients.

The landscape of the non-bank sector in Timor-Leste is complemented by other financial service providers as Credit Unions, Insurance providers, Currency Exchange Bureaus (CEB) and Money Transfer Operators (MTO). As at June 2016, the number of Credit Union stood at 32 while Insurance providers had 2 branches operating in the country. CEBs and MTOs provide their services via branches and, in June 2016, had a combined of 16 branches between them across the country.

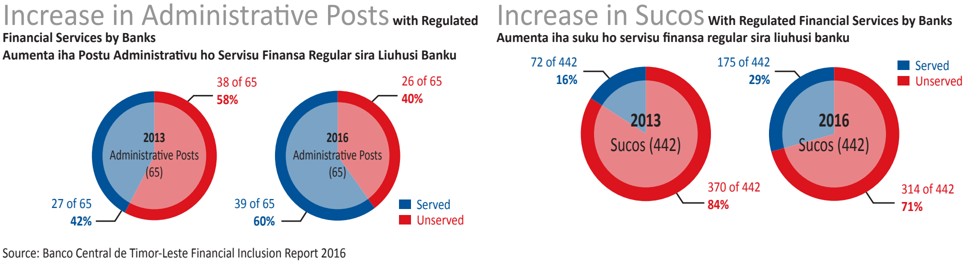
1 The abbreviation MFIs used throughout the text refers exclusively to "Microefinance Institutions" and should not be confused with the generalized used of this abbreviation for "Monetary and Financial Institutions"

Coverage at the administrative posts and sucos levels.

The level of access to financial services in the country is fairly high, whereby all municipalities are served with financial services access points. The majority of administrative posts also benefit from the presence of financial services access points, and as at 2016, 88% of administrative posts were covered.

However, there is a slight drop comparing with 2013 figures where 92% of administrative posts were covered. At the sucos level, 58% of them have the presence of financial services access points as at 2016, a number lower than the one recorded in 2013 where 63% of sucos were covered. The drop in financial access points from 2013 to 2016 was attributed to the drop in MFI access points due to the reduction of the usage of meeting centers after 2013.

At the administrative posts level, regulated financial services by banks were available in 60% (or 39) administrative posts in 2016, while 29% (or 127) sucos were covered.



Graph 1. Financial Services by Banks in Administrative post and Sucos

It is evident that gaps still exist, as there are administrative posts that are not yet covered by financial access. The numbers are more concerning at the sucos level. A large number of sucos continue to be unserved by access points and financial services in the country. 187 among a total of 442 sucos still do not have financial services access points, while 314 sucos remain unserved by banking regulated financial services.

The population density, based on the Census 2010, also highlights the varying densities of administrative posts: administrative posts with higher population densities, especially those which are still underserved, present higher potential for increasing access points in the future.

It was found that 81% of the population, or 858,731 of 1,066,409 people, are living in administrative posts covered by bank access points. From the adult population point a view, the same percentage was recorded, where 81% of adults are living in administrative posts covered by bank access points.

Limited coverage by branches.

Due to the distance, there is some limited coverage by banks' branches. Within a radius of 5km, around one hour of walking, the bank branches cover 6.5% of the national territory, including most urban areas in the municipal cities. 10km radius from the bank branches covers 22.8% of the national territory, while 25km radius covers 80% of the national territory. Hence, people living in remote areas, (such as south Manatuto and west Lautem), may need to travel up to 25km or more to reach the nearest bank branch. They may also face adverse road conditions and transportation amenities to reach a bank branch.

2. USAGE INDICATORS

The number of people who uses financial services, either provided by banks or non-bank, is central to usage indicators. This measurement of financial services acceptance gives an overview of the penetration of financial products and services in the country.

Table 2. Usage Indicators

Financial Institution	Total Clients
Banks	388,018
MFIs	26,873
Credit Unions	8,059
Total	**422,941**

In Timor-Leste, the data collected in June 2016 found that there were 422,941 clients with accounts in banks, MFIs and credit unions. Given this figure, it is estimated that about 42% of the population (or 68% of adult population) use financial products and services.

The largest share of clients in the financial system is held by banking institutions, with a total of 388,018 customers. 96.5% of this total clients are individuals, while the remaining 3.5% are corporate clients. A look into the breakdown by gender of the individual clients shows that a slightly higher percentage of males are using financial services by banks (52% of total banking clients) compared to 48% of females clients.

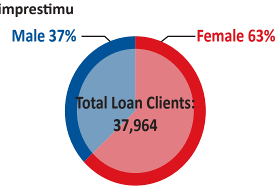
3. CREDIT MARKET

Using Timor-Leste 2010 population census information we can conclude that only 6% of adult population has an outstanding loan, of which 55% borrowed from the banking system and 45% borrowed from MFIs. Loan products offered by banks are granted in a higher proportion to male clients (13,354 men, corresponding to 63% of the total costumers compared to 7,642 female clients, which corresponds to the remaining 37%). However, MFIs loan products are more demanded by female clients, with a total of 16,257 clients (96%) against just 678 male clients (4%).

Gender Distribution of Loan Account Ownership

Distribuisaun Jeneru tuir Propriedade Konta imprestimu

Gender	No. of Clients		
	Banks	MFIs	Total
Female	7,642	16,257	23,899
Male	13,354	678	14,032
Not Informed	33	-	33
Total			**37,964**



Source: Banco Central de Timor-Leste Financial Inclusion Report 2016

Graph 2. Credit Markets

The higher number of female borrowers in MFIs, has being contributed to increase the number of female clientele for loan products, which corresponds to 63% of total loan clients in the overall financial sector.

Just to share some similar observations with regard to the penetration of loan accounts by municipalities in the country, Dili (the capital of the country) records the highest usage in loan accounts offered by both banks and MFIs. The number of bank loan account holders in Dili reaches more than 35% of the total bank loan account holders, while for MFIs, the number lower to 18.6% of the total MFI loan account holders.

Other Financial Products Usage Still Limited

Usage indicators also examine other important features, particularly on how financial accounts are being used. The Bank has evaluated the usage of deposit accounts in banking institutions, including the penetration of the usage of debit cards and the collection of outstanding deposits.

It was found that only 7% (28,177) of the clients are using debit cards, out of the total 387,063 deposit accounts registered in banks. It was also observed that there are 297,630 clients with deposit balances below USD100.00, including a sizable number of 244,332 clients with deposit balances below USD10.00. Clients with deposits below USD10.00 are made up of both corporates (2,860 accounts) and individuals (241,471 accounts). The low account balances indicates that some of these accounts may not be actively used, resulting in a lower savings rate. Another reason behind the low balances is due to the primary usage of a portion of these individual accounts: to receive financial aid from the government.

E-wallet is another new approach adopted to increase outreach of financial products and services for consumers in Timor-Leste. This product is currently offered by CGD/BNU which was officially released in March 2015, after the start of pilot run in November 2014. This product offers a set of services such as transfers between CGD/BNU accounts, including e-wallet to e-wallet or current to e- wallet, purchase of Pulsa prepaid top-up for phone credit, check account balances, transactions status and International Bank Account Numbers (IBAN) for current accounts. There are currently 1,573 clients registered for this product and, on average, 2,000 of monthly transactions were performed using this product. The e-wallet product is a promising strategy to enhance financial inclusion as there are about 1.5 million registered mobile account subscriptions in the country, a number which is higher than the total population, signifying a wide penetration of mobile phone accounts.

In line with the development in the financial sector, internet banking services were also introduced in Timor-Leste, with the first initiative rolled out in January 2013 by ANZ Bank. ANZ Bank has 2,500 registered users and the volume of transactions ranges from 600 to 800 per month. CGD/BNU launched, in February 2016, their internet banking initiative called BNUdireto, through a platform that is serving 700 individuals and 33 enterprises with an estimated volume of transactions up to 1,500-2,000 per month.

In the recent years, insurance providers in the country have been striving to expand their product range to help all segments of society manage their risks. To date, insurance providers have expanded their product range to more than 7 types of insurance policy types, and currently, there are a total of 1,052 individual insurance policy holders in the country and 212 collective insurance policy holders. Timorese mostly use MTOs to conduct international transfers, being the majority of transactions outbound. International outbound transfers make up 70.5% of total transactions valued at over USD48,900.00, in the first half of 2016. However, international inbound transfers account for 29.45%, of total transactions valued at over USD19,400.00.

The usage of financial products by regions show that Dili has the largest share of deposit account holders, which is consistent with the population of Dili, and with the fact that it is the municipality with highest population density in the country. Totalising 119,119 clients, Dili has 233% more clients than the next highest municipality, Baucau, which has 35,802 clients. The municipalities with the lowest numbers of individual deposit clients are Ailieu (13,753), Manatuto (16,327), Manufahi (16,568) and Ainaro (16,772).

Identified Barriers for Financial Inclusion

The procedure of opening a bank account appears to be a straightforward process. The steps involve going to one of the four commercial banks' branches, presenting an identification document with a photo (electoral card, which is the most common document among adult Timorese, national identity card or passport), filling bank forms and, within the same day, the account application is approved for accepting deposits.

Timor-Leste adopted Tetum and Portuguese as official languages, and English and Bahasa Indonesia as working languages. In addition to Tetum, there are other 15 dialects used in specific areas across the country, making it challenging for financial services providers to select languages for an effective communication. According to the Census 2010, Tetum is the most widely used language, but only 56.1% of the adult population are able to speak, read and write it. Therefore, there is still a need to use language alternatives to cater for the remaining population.

The network of bank branches is still limited to the main city of each municipality. While it covers 29.6% of the population (people living in urban areas), the remaining 70.4% (population living in rural areas) are required to travel a significant distance to reach the nearest bank branch. Public transport connecting main cities to other villages, when existing, is precarious and unreliable. For many, the only alternative is to rely on informal arrangements or walk long distances, at times up to 25 km or more, to the nearest branch.

Identified Gaps & Challenges for Inclusive Finance

The challenges for access to financial services are mainly related with the deployment of access points to areas outside the urban centres, especially to administrative posts and sucos with large land areas and lower population density. Currently, there are 8 administrative posts unserved by any financial service providers, regulated or not, and 26 administrative posts unserved by banks. The numbers are even more concerning at the sucos levels. The lack of access points in these unserved and underserved areas give the residing communities no choice, but to travel a long distance to reach a financial access point. The features of financial products and services available in the market also play a key role in either driving or hindering financial inclusion. If financial services are complicated and expensive to use, it would be natural for people to reject these products. In fact, all four banks in Timor-Leste are charging a range of different fees for monthly maintenance, debit cards, and a range of different minimum initial deposit and bank account balance. The logical next step would be to improve the product availability, the pricing and processes in the financial sector to serve the requirements and needs of those without access to banks so they can also benefit from the financial system. The actual state of financial sector also suggests that there is a need for supporting the infrastructures to improve information asymmetry and manage emerging risks.

Financial literacy, Consumer protection & Capacity building are important challenges for inclusive finance. The lack of awareness and knowledge about financial products and services, especially amongst the poor and rural communities, also pose as a hindrance towards fostering financial inclusion. There are still many people within the unbanked and underserved segments who simply do not comprehend what financial services are available, why they are important, how to use and what are the benefits of using them.

BCTL Key Initiatives to Promote Financial Inclusion

- E-Wallet: on September 2014, the Bank authorised the BNU/CGD to start a pilot program that allows a set of banking transactions to be carried out by mobile phone.
- Banking agent: During 2015, the Bank worked with the ADB and the University of New South Wales (UNSW) on a branchless banking agents network project. As a result, a Circular Letter 14/2015: Guidelines on the Use of Agents in Branchless Banking was introduced. In addition, the Bank, ADB and the UNSW team produced a policy paper on supervision and protection of float in branchless banking, and also produced a policy paper on consumer protection in branchless banking.
- Small and Medium Enterprises' (SME) access to credit: the Bank initiated in 2013 a feasibility study on options for agricultural finance and a loan guarantee schemes granted by Government to support the private sector in Timor-Leste.
- Financial literacy and consumer protection: the Bank spearheaded its first awareness program of Financial Education on 29 November 2014, which was launched officially by the President of RDTL (Democratic Republic of Timor-Leste) H.E. Mr Taur Matan Ruak. This was a nationwide awareness program aimed at improving financial literacy. The theme for the program, "Save now for a better future" was developed in consultation with community groups, and its main objective is to raise awareness about personal and small business finances and thereby change spending behaviour to inculcate savings habits.

Financial Literacy Program

The program of "Save now for a better future" has the purpose of improve basic financial literacy levels of the Timor-Leste people through a nationally coordinated network of financial education programs and awareness campaigns. Its mission is to convey the Timorese the idea that they can use products and services provided by the finance sector to enhance their wellbeing thus contributing to the economic growth of the nation. The expected outcomes include a recognition of target groups with improved financial knowledge, skills and attitudes that will foster economic growth in the country.

4. CONCLUSION

The Bank has carried out a lot of initiatives in the last few years, in collaboration with the government, financial service providers and other partners, as an important effort to build the foundations for inclusiveness in the financial sector in Timor-Leste. However, the Bank still needs to strive to make this real. To enhance financial inclusion, the Bank is concentrating its efforts on identifying opportunities to through coordinated strategies in the coming future.

REFERENCES

BCTL (2016); BCTL Financial Inclusion Report 2016

IPS07:
Modelling And Analysis With Application To Finance And Insurance

Real-Time Valuation of Large Variable Annuity Portfolios: A Green Mesh Approach

Kai Liu
Ph.D. Candidate
Department of Statistics and Actuarial Science
University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
k26liu@uwaterloo.ca

Ken Seng Tan*
University Research Chair Professor
Department of Statistics and Actuarial Science
University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
kstan@uwaterloo.ca

Abstract

The valuation of large variable annuities (VAs) portfolio is an important problem of interest, not only because of its practical relevance but also of its theoretical significance. This is prompted by the phenomenon that many sophisticated algorithms are typically efficient at valuing a single VA policy but they are not scalable to valuing large VAs portfolio consisting of hundreds of thousands of policies. As a result, this sparks a new line of research direction exploiting machine leaning methods (such as data clustering, Kriging, neutral network) on providing more efficient algorithms to estimate the market values and sensitivities of the large VAs portfolio. The idea underlies these approximations is to first determine a set of VA policies that is ``representative" of the entire large VAs portfolio. Then the values from these representative VA policies are used to estimate the respective values of the entire large VAs portfolio. A substantial reduction in computational time is possible since we only need to value the representative set of VA policies, which typically is a much smaller subset of the entire large VAs portfolio. Ideally the large VAs portfolio valuation method should adequately address issues such as (1) the complexity of the proposed algorithm, (2) the cost of finding representative VA policies, (3) the cost of initial training set, if any, (4) the cost of estimating the entire large VAs from the representative VA policies, (5) the computer memory constraint, (6) the portability to other large VAs portfolio valuation. Most of the existing large VAs portfolio valuation methods do not necessary reflect all of these issues, particularly the property of portability which ensures that we only need to incur the start-up time once and the same representative VA policies can be recycled to valuing other large portfolio of VA policies. Motivated by their limitations and by exploiting the greater uniformity of the randomized low discrepancy sequence and the Taylor expansion, we show that our proposed method - green mesh method - addresses all of the above issues. The numerical experiment further highlights its simplicity, efficiency, portability, and more importantly, its real-time valuation application.

Keywords: Variable annuities; Large portfolio valuation; Data mining; Quasi-Monte Carlo methods.

Arbitrage Model with Factor-Augmented Predictors and Applications to China's Stock Market

Xiaorong Yang

Zhejiang Gongshang University, Hangzhou, China – xiaorongyang@hotmail.com

Abstract

Quantitative equity portfolio management has become a fundamental building block of the investment management. The development of general equilibrium asset pricing models enables statistical arbitrage strategies to capture the effect factors of the market returns. In empirical analysis, a crucial step in the model-building process is the selection of essential factors, which may contribute to the positive excess returns. However, it could be challenging due to the thousands of candidate factors obtained. In this study, we employed a factor-augmented model to identify the effective factors for excess returns, and rank the portfolios according to the selected factors. We perform a trading strategy of the combination of buying stock portfolio and stock index futures hedging for Alpha arbitrage. Moreover, we derive a prediction of value at risk of the portfolio by using quantile regression techniques.

Keywords: Arbitrage model, Portfolio management, Stock market.

A Bayesian Quantile Regression Model for Insurance Company Costs Data

Karthik Sriram

Indian Institute of Management, Ahmedabad, India – karthiks@iima.ac.in

Abstract

We examine the average cost function for property and casualty insurers. The cost function describes the relationship between a firm's minimum production cost and outputs. A comparison of cost functions could shed light on the relative cost efficiency of individual firms, which is of interest to many market participants including the insurance industry. To identify and to compare the cost function, current practice is to rank insurers according to the centre of the cost distribution. However, the assumption of a common cost–output relationship could be misleading because insurers tend to adopt different technologies that are reflected by the cost function in their production process. To address these issues, we model the average production cost of insurers by using a Bayesian quantile regression approach. Quantile regression enables the modelling of different quantiles of the cost distribution as opposed to just the center. The Bayesian approach helps to estimate the cost-to-output functional relationship at a firm level by borrowing information across firms. In the analysis of US property–casualty insurers, we show that better insights into efficiency are gained by comparing different quantiles of the cost distribution.

Keywords: Bayesian joint, Insurance industry.

IPS08:
Recent Advances in The Analysis of Biomic Data

Network Analysis of Next-Generation Sequencing Count Data

Somnath Datta

University of Florida, Gainesville, USA – somnath.datta@ufl.edu

Abstract

We propose a combination of partial least squares and negative binomial regression with a model offset in order to conduct a network analysis of genomic data obtained from next generation sequencing experiments. The methodology is evaluated in a number of simulation experiments. Its utility is demonstrated via an application to a large scale genomic data set of palate development in mice.

Keywords: Gene expression; Genetic association; Interaction network; Partial least squares.

Statistical Association Based Analysis for Genomic Data

Susmita Datta
University of Florida, Gainesville, USA – susmita.datta@ufl.edu

Abstract

We introduce various measures of association scores to predict the functional interaction of molecular products such as gene-gene and between lipid interactions. These association measures are computable from the array and next generation sequencing genomic data or mass spectrometry based lipidomics data. We have used these pairwise association scores to construct an association based network via an empirical Bayes approach leading to a better understanding of the collective working of genes or lipids during a biological process. Additionally, we introduce permutation based statistical tests for detecting various changes in the network topology of the molecular products from one biological condition to another. In the later part of the talk, we introduce a statistical method based on association scores to identify key transcription factors such as the master regulator transcription factor which controls the other transcription factors and their target genes differentially under two different biological systems or the samples coming from two different tissue types.

Keywords: Microarray; Association; Interaction; Transcription factor.

This page intentionally left blank

CPS06: Statistical Theory & Methodology (1)

Bayesian Approach to Errors-in-Variables in Poisson Regression Model

Adriana Irawati Nur Ibrahim*
University of Malaya, Kuala Lumpur, Malaysia – adrianaibrahim@um.edu.my

Nur Aainaa Rozliman
University of Malaya, Kuala Lumpur, Malaysia – aainaaroz@gmail.com

Rossita Mohammad Yunus
University of Malaya, Kuala Lumpur, Malaysia – rossita@um.edu.my

Abstract

Poisson regression is widely used to model count data. When count data show evidence of overdispersion and bias, negative binomial regression is often used instead to treat this occurrence. The treatment however, would lead to the wrong conclusion when measurement error or errors-in-variables (EIVs) is the actual cause. When at least one of the explanatory variables in the regression model is measured with error, EIVs models can be used to examine the relationship between the outcome variable and the unobserved explanatory variables (also known as covariates or exposures) given the observed mismeasured explanatory variables. Even though Poisson EIVs regression model has been extensively researched using the frequentist approach, very few have utilized Bayesian adjustment for this model. Therefore, our study focused on applying the flexible parametric approach, which uses flexible distributions, to the Bayesian Poisson EIVs regression model. For computational purposes, Markov chain Monte Carlo techniques were applied. Simulation studies were then conducted to investigate the performance of the proposed approach. The results showed that the approach worked well and was able to estimate the true regression parameters consistently and accurately.

Keywords: Markov chain Monte Carlo; Flexible parametric approach.

Journal of Economic Literature (JEL) Classification: C6; C8.

1. INTRODUCTION

Count data are data that has non-negative integers as their response values. Poisson regression is vitally used to model count data and has witnessed a rich set of applications such as econometrics, epidemiology and agriculture. When count data show evidence of overdispersion and bias, negative binomial regression is often used instead to treat this occurrence. The treatment however, would lead to the wrong conclusion when measurement error or errors-in-variables (EIVs) is the actual cause. EIVs can contaminate data when instead of observing the real value of the explanatory variables, their surrogate estimates are observed. It is important to address the EIV problem as ignoring EIV without treating it to model the association between the outcome and the true explanatory variables often leads to biased and inconsistent estimate. This in return would produce the wrong conclusion on the underlying relationship between these variables.

Over the years, most studies correcting EIVs in Poisson regression has focused on the use of frequentist methods. Corrected score estimator was proposed by Stefanski (1989) and Nakamura (1990) for parameters contaminated with error in the mean function. Kukush et al. (2004) and Shklyar and Schneeweiß (2005) later investigated the performance of three estimators, that is the naive, corrected score and structural quasi score estimator specifically for Poisson regression.

On the other hand, to the best of our knowledge, there is relatively few studies on the usage of Bayesian approach to correct bias in Poisson EIV model. This is due to the computational difficulties that Poisson regression presents when applying Bayesian method. Mallick and Gelfand (1996) proposed Bayesian method for EIV in GLMs and provided Poisson regression as an example in their approach. Fu et al.

(2015) constructed a Bayesian approach for zero-inflated Poisson model where the data has skewness and measurement error. The advantages of the Bayesian method compared to various frequentist methods are that we can assume non-normality of the exposure data and also the variance of the measurement error is estimated instead of assumed known.

2. EIVS MODEL

When at least one of the explanatory variables in the regression model is measured with error, EIVs models can be used to examine the relationship between the outcome variable and the unobserved explanatory variables (also known as covariates or exposures) given the observed mismeasured explanatory variables. Consider Poisson regression with outcome variable Y and its accurately measured but unobserved exposure X. Typically, in real life application of count data, X is contaminated with error and its surrogate X^* instead is observed such that,

$$X^* = X + \varepsilon \tag{1}$$

$$\text{or } X = X^* + \varepsilon, \tag{2}$$

where ε is assumed to be independent and identically distributed (iid) and it follows *that* $\varepsilon \sim N(0, \sigma^2_\varepsilon)$. Equation (1) is referred to as classical measurement error where ε is independent of the true underlying exposure variable and equation (2) is referred to as Berkson measurement error model. Note that, in this paper we consider the measurement error model (1). We follow the general Bayesian formulation for the joint density of EIVM based on the framework provided by Richardson and Gilks (1993),

$$f(Y, X, X^*|\theta_R, \theta_M, \theta_E) = f(Y|X, \theta_R) f(X^*|X, \theta_M) f(X|\theta_E). \tag{3}$$

The first component of the joint density, known as the outcome model describes the relationship between response Y and true exposure variable X. Our main inferential focus and the parameters that we want to estimate are their vector of parameters, θ_R . The second component describes the relationship between surrogate value X^* and unobserved true value X and it is known as the measurement model. The last component, a model that is subjected to misspecification in this paper, is called the exposure model and the model describes the distribution of the true exposure variable X.

Structural approach to correct bias in EIV model requires the specification of exposure model. Unfortunately, in real life applications, the exposure model is unknown as the independent covariate X is unobserved. X might take a different distribution than what is specified as the exposure model. There are extensive research done on how to take on these problems, especially for data that follows logistic distribution. Richardson and Leblond (1997) concluded from their investigation that the misspecification of an exposure model will cause biased estimates. Some used functional approach to avoid any distributional assumptions, however efficiency is sacrificed. Therefore, in structural approach, to reduce sensitivity to model misspecification, the usage of flexible parametric distribution is often considered. For logistic EIV model, Richardson et. al (2002) specified their exposure model as normal mixture model; Huang et. al (2006) and Hossain and Gustafson (2009) utilized flexible distributions as their exposure model. Bolfarine and Lachos (2007) made use of skew-normal as the exposure model for probit regression with its covariate contaminated with error. To our knowledge, there are none that utilize flexible exposure model in the structural EIV model for Poisson regression. Therefore, our study focuses on applying the flexible parametric approach, which uses flexible distributions, to the Bayesian Poisson EIVs regression model, which will be discussed further in the next section. In this paper, we will test the robustness of a flexible distribution introduced by Nekoukhou et al. (2013) known as flexible skew-generalized normal (FSGN) as the misspecified exposure model.

3. BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN POISSON REGRESSION MODEL

We consider Poisson regression with mean function $\exp(\beta_0 + \beta_1 x_i)$, for $i = 1, \dots, n$, where n is the sample size. To simplify computation, we introduce latent variable η_i such that for the observed count data $Y_1, Y_2, \dots, Y_n$ we have

$Y_i|\eta_i \sim Poisson\ (\exp(\eta_i))$ where $\eta_i = \beta_0 + \beta_1 X_i$.

The three components of the structural measurement error (Richardson and Gilks, 1993) for Poisson outcomes which will be used to form the likelihood are

Outcome Model $\quad f(Y_i|X_i, \theta_R) = Poisson\ (\exp(\eta_i))$ (4)

Measurement Model $\quad f\left(X_{ij}^*\middle|X_i, \theta_M\right) = N(X_i, \tau^2)$ (5)

Exposure Model $\quad f(X_i|\theta_E) = FSGN(\alpha, \lambda_1^2, \lambda_2, \omega_1, \omega_2),$ (6)

where $\theta_R = (\beta_0, \beta_1, \sigma^2)$, $\theta_M = \tau^2$, $\theta_E = (\alpha, \lambda_1^2, \lambda_2, \omega_1, \omega_2)$ and X_{ij}^* signifies the j^{th} replicated surrogate for i^{th} observation of X, $j = 1, \dots, m$. Note that to help estimate the parameters in the EIVs model, error assessment data or validation data can be used. In this paper, the error assessment data are chosen to contain m values of observed X^*.

FSGN is introduced by Nekoukhou et al. (2013) where the distribution shows flexibility in modeling unobserved, skewed and bimodal data. The FSGN distribution with location parameter α, shape and scale parameter ω_i and λ_i respectively, for i = 1, 2 is given by

$$f(x) = \frac{2}{\lambda_1}\,\phi\left(\frac{x-\alpha}{\lambda_1}\right)\Phi\left[\frac{\omega_1(x-\alpha)+\frac{\omega_2}{\lambda_1^2}(x-\alpha)^3}{\sqrt{\lambda_1^2+\lambda_2(x-\alpha)^2}}\right], \tag{7}$$

where $\omega_1, \omega_2 \in \mathbb{R}$, $\lambda_1 > 0$ and $\lambda_2 \geq 0$..

4. JOINT POSTERIOR DISTRIBUTION

The joint posterior distribution, $f(X_i, \theta|X_{ij}^*, Y_i)$ for EIV model can be obtained as the product of likelihood stated in (4), (5) and (6) and also the prior distributions of all the parameters involved $f(\theta)$, where

$$f(\theta) = f(\eta_i|\beta, \sigma^2)f(\beta)f(\sigma^2)f(\tau^2)f(\alpha)f(\lambda_1^2)f(\lambda_2)f(\omega_1)f(\omega_2)$$

where $\theta = (\eta_i, \beta, \sigma^2, \tau^2, \alpha, \lambda_1^2, \lambda_2, \omega_1, \omega_2)$ and $\beta = (\beta_0, \beta_1)$.

We adopt the prior used in Asfaw Dagne (1999) for the latent variables η_i, where they are normally distribution with mean $X'\beta = \beta_0 + \beta_1 X_i$ and variance $\sigma^2 I$. Meanwhile for the hyperparameter σ^2 and scale parameter λ_1^2, they follow Inverse-Gamma (Gelman et al., 2014). The same prior is used for τ^2. β and α follow flat priors, meanwhile both ω_1 and ω_2 assume normal distribution with mean zero and large variance Similarly, to assume the noninformative prior property we let the prior for λ_2 be a halfnormal distribution with variance 1 (Gelman, 2006).

5. SAMPLING METHODS

For computational purposes, Markov chain Monte Carlo techniques are applied. The posterior conditional for latent variable η_i is not in closed form (i.e., it does not follow any known distribution). Thus, random-walk Metropolis-Hastings (RWMH) algorithm is used to update the latent variable by block-wise sampling. Similarly, ω_k is also updated using RWMH. The posterior conditional for parameters β, σ^2 and τ^2 is in closed form, therefore, Gibbs sampler is used to update these parameters. Parameter λ_1^2, λ_2, α and X are updated using Metropolis-Hastings.

6. SIMULATION DESIGNS AND SETTINGS

This section is to illustrate the performance of Bayesian approach in reducing bias for Poisson EIV regression model when the exposure model is misspecified. To misspecify the model, unobserved true exposure X is generated from a skewed mixture of normal. In this simulation design, we let $n = 50$. The following is the distribution of simulated data for Y_i, X_i and X_{ij}^* :

$$Y_i \sim Poisson\,(\exp(\beta_0 + \beta_1 X_i))$$
$$X_i \sim 0.5N(0.19, 0.08^2) + 0.2N(1.05, 0.2^2) + 0.3N(2, 0.48^2) \quad (8)$$
$$X_{ij}^* = X_i + \epsilon_j\,, \quad \text{where} \quad \epsilon_j \sim N(0, \tau^2)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. For the sake of simulating data that is similar to real life research situation, the number of replicated surrogates is limited to $m = 2$. True values for the parameters in our model are shown in Table 1 while Figure 1 shows the histogram for the simulated X. Note that $\tau^2 = 1.11$ corresponds to a high measurement error.

Table 1. True Parameter Values

Parameter	True values
$\beta = (\beta_0, \beta_1)$	$(0.5, 1)$
τ^2	1.11

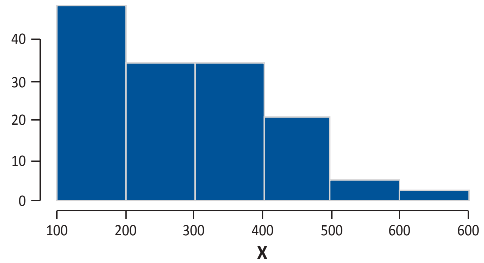


Figure 1. Histogram for Simulated X

7. RESULTS

To investigate the performance of our model, we compare our estimation for the regression parameters, β_0 and β_1, with the benchmark estimate as well as the naive estimate. The benchmark estimate is obtained using direct regression of true parameter values. Meanwhile, by assuming the mean of the surrogate values, $\bar{X}_i^* = \frac{\sum_{j=1}^{m} X_{ij}^*}{m}$, as the true underlying exposure we implement direct regression on it to obtain the value of naive estimates. These naive estimates are also known as biased estimates since they are produced from covariates contaminated with error without bias correction. The estimate for our model is the average of β_0 and β_1 over a decided number of iterations after burn-in. We investigate the ability of our model to accurately estimate the regression parameters by conducting the simulation N times and calculating its bias and mean squared error (MSE). The results are shown in Tables 2 and 3.

Table 2. Estimated value of parameters of the EIV Poisson regression model

Parameter	Naive	Our Model	Benchmark
β_0	1.2998	0.5912	0.4980
β_1	0.3696	0.8501	0.9990

Table 3. Accuracy of estimation of parameters of the EIV Poisson regression model using our approach

N	Parameter	Bias	MSE
30	β_0	0.0267	0.0008
	β_1	0.0044	0.0001
50	β_0	0.0266	0.0008
	β_1	0.0045	0.0001
100	β_0	0.0247	0.0007
	β_1	0.0058	0.0001

After 100,000 burn-ins, we find the posterior means of β_0 and β_1 over 200,000 iterations which is shown in the third column of Table 2. The convergence for all the parameters are analyzed using trace plots; example of the trace plots for β_0 and β_1 from one of the simulation study is presented in Figure 2. These plots show that MCMC chains have good mixing and convergence is achieved. Results shown in Table 2 show that our approach provides good estimation even though the exposure model is misspecified, as the estimates obtained from our model are closer to the benchmark estimates than the naive estimates. From Table 3 we can see that our approach gives accurate and consistent estimates of the regression parameters.

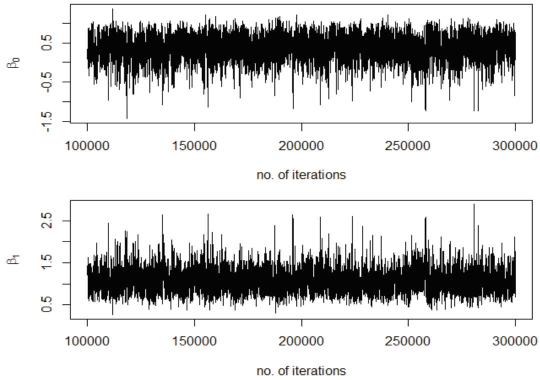


Figure 2. Trace Plots for β_0 and β_1 in one of the simulation studies.

8. CONCLUSION AND DISCUSSION

In this paper, the flexible parametric approach is applied to the Bayesian Poisson EIVs regression model. Here, we test the robustness of the model misspecification by using a flexible distribution, that is the FSGN distribution, as the exposure distribution in order to relax normality assumption. Our study shows that the approach works well and is able to estimate the true regression parameters consistently and accurately. For future research, we may consider X to be generated from other distributions and also using other types of flexible distributions as the exposure model.

9. REFERENCES

Fuller W.A. (1987). *Measurement Error Models*. Wiley, New York.

Stefanski, L. A., & Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**(4), 703-716.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement error models. *Communications in Statistics-Theory and Methods*, **18**(12), 4335-4358.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, **77**(1), 127-137.

Richardson, S., & Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**(6), 430-442.

Dellaportas, P., & Stephens, D. A. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics*, 1085-1095.

Mallick, B. K., & Gelfand, A. E. (1996). Semiparametric errors-in-variables models A Bayesian approach. *Journal of Statistical Planning and Inference*, **52**(3), 307-321.

Asfaw Dagne, G. (1999). Bayesian analysis of hierarchical Poisson models with latent variables. *Communications in Statistics-Theory and Methods*, **28**(1), 119-136.

Richardson, S. et al (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **165**(3), 549-566.

Genton, M. G., & Loperfido, N. M. (2002). Generalized skew-elliptical distributions and their quadratic forms. *Annals of the Institute of Statistical Mathematics*, **57**(2), 389-401.

Ma, Y., & Genton, M. G. (2004). Flexible Class of Skew-Symmetric Distributions. *Scandinavian Journal of Statistics*, **31**(3), 459-468.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, **1**(3), 515-534.

Bolfarine, H., & Lachos, V. H. (2007). Skew-probit measurement error models. *Statistical Methodology*, **4**(1), 1-12.

Hossain, S., & Gustafson, P. (2009). Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in medicine*, **28**(11), 1580-1600.

Nekoukhou, V., M. H. Alamatsaz, and A. H. Aghajani. "A flexible skew-generalized normal distribution." *Communications in Statistics-Theory and Methods* **42(**13) (2013), 2324-2334.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Bayesian Accelerated Life Testing

Lizanne Raubenheimer
Rhodes University, Grahamstown, South Africa – L.Raubenheimer@ru.ac.za

Abstract

In this paper a Bayesian approach for accelerated life testing will be considered. It will be assumed that the failure times at the different stress levels follow a Weibull distribution and exponential distribution, respectively. In accelerated life tests, the components are exposed to an environment that is more severe that the usual environment, such that the components will fail in a shorter than usual period of time. The failure information is then transformed through an accelerated model to predict the reliability under normal operating conditions. The power law model will be used. This accelerated life testing model is typically used when the accelerated stress is non-thermal. Inference for the model will be discussed, and results can be obtained by using Markov chain Monte Carlo (MCMC) methods.

Keywords: Accelerated life testing; Bayesian inference; Exponential distribution; Weibull distribution.

1. INTRODUCTION

In this paper we will consider a Bayesian approach to accelerated life testing for the power law model using the Weibull distribution and the exponential distribution, respectively, as the life distributions. Mazzuchi *et al.* (1997) considered a Bayesian approach for inference from accelerated life tests when the underlying life model is Weibull, using the power law. Their approach is based on the linear model framework by West *et al.* (1985). Singpurwalla *et al.* (1975) obtained least squares estimators of the parameters of the generalized Eyring model, where the life model is exponential. Chaloner & Larntz (1992) studied an experimental design for accelerated life tests where the life times were either lognormal distributions or Weibull distributions, using a Bayesian approach. Van Dorp & Mazzuchi (2005) developed a general Bayes inference model for accelerated life testing, where the failure times at a constant stress level were assumed to belong to a Weibull distribution, but the specification to a parametric time-transformation function is not required. They used prior information to indirectly define a multivariate prior distribution for the scale parameters at the different stress levels. Most modern products are designed to operate without failure for years, thus few units will fail in a test of practical length at normal use conditions. For such applications, ALTs are used in manufacturing industries to assess reliability, see Escobar & Meeker (2006) for further discussion. Erkanli & Soyer (2000) introduced a simulation-based design for accelerated life tests, where they considered an exponential life model with the power law as the time transformation function, using Monte Carlo approaches. This paper will extend on the work done by Erkanli & Soyer (2000), and focus on an objective Bayesian approach for the exponential model and a simulation based approach for the Weibull model.

2. THE MODEL

2.1. Exponential Model

We assume that life length X_i is exponential with failure rate λ_i, denoted by $X_i|\lambda_i \sim Exp(\lambda_i)$ with probability density function given by

$$f(x_i|\lambda_i) = \lambda_i exp\{-\lambda_i x_i\}$$

where $x_i \geq 0$ and λ_i is the failure rate at the accelerated stress environment, S_i. In this paper we will make use of the power law. The relationship between the failure rate and the stress level in the i^{th} testing environment will then be $\lambda_i = \alpha S_i^{\beta}$, where $\alpha > 0$ and $\beta > 0$ are the model parameters to be determined. This implies $X_i|\alpha, \beta, S_i \sim Exp\left(\alpha S_i^{\beta}\right)$ and the density is given by

$$f\left(x_i|\alpha S_i^{\beta}\right) = \alpha S_i^{\beta} exp\left\{-\alpha S_i^{\beta} x_i\right\} \tag{1}$$

The parameter β is a measure of the effect of the stress on life. We are interested in making inference about $\lambda_u = \alpha S_u^{\beta}$, the failure rate at the use stress environment. We assume that there is no censoring in the ALT. From equation (1), the likelihood function is given by

$$L(\alpha,\beta|data) = \alpha^n \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) exp\left\{-\alpha \sum_{i=1}^{m} S_i^{\beta} T_i\right\}.$$

where $T_i = \sum_{j=1}^{n_i} x_j$ and represents the total time on test at stress environment S_i.

The prior information can be obtained from historical data or expert opinion. When there is little or no prior information available, objective Bayes can be used to derive priors for the unknown parameters. The well-known Jeffrey prior will be considered. The Jeffreys' prior is obtained from the Fisher information matrix. The Fisher information matrix is given by

$$I(\alpha,\beta) = \begin{bmatrix} \frac{n}{\alpha^2} & \frac{1}{\alpha}\sum_{i=1}^{m} n_i log S_i \\ \frac{1}{\alpha}\sum_{i=1}^{m} n_i log S_i & \sum_{i=1}^{m} n_i (log S_i)^2 \end{bmatrix}.$$

The Jeffreys prior is obtained by taking the square root of the determinant of the Fisher information matrix,

$$\pi_{E-J}(\alpha,\beta) \propto \sqrt{|I(\alpha,\beta)|} \propto \alpha^{-1}.$$

When using the Jeffreys' prior, the joint posterior will be

$$\pi_{E-J}(\alpha,\beta|data) \propto \alpha^{n-1} \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) exp\left\{-\alpha \sum_{i=1}^{m} S_i^{\beta} T_i\right\}.$$

The marginal posterior of β using the Jeffreys prior is given by

$$\begin{aligned} \pi_{E-J}(\beta|\text{data}) &= \int \pi_{E-J}(\alpha,\beta|\text{data}) d\alpha \\ &= \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) \int_0^{\infty} \alpha^{n-1} exp\left\{-\alpha \sum_{i=1}^{m} S_i^{\beta} T_i\right\} d\alpha \\ &= \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) \left(\sum_{i=1}^{m} S_i^{\beta} T_i\right)^{-n} \Gamma(n), \end{aligned}$$

and the conditional posterior of α given β is

$$\pi_{E-J}(\alpha|\beta,data) = \frac{\left(\sum_{i=1}^{m} S_i^{\beta} T_i\right)^n}{\Gamma(n)} \alpha^{n-1} exp\left\{-\alpha \sum_{i=1}^{m} S_i^{\beta} T_i\right\}.$$

Alternatively, one can assume independent gamma priors on α and β, $\alpha \sim Gamma(a_1,b_1)$ and $\beta \sim Gamma(a_2,b_2)$. When using independent gamma priors, the joint posterior will be

$$\pi_{E-G}(\alpha,\beta|data) \propto \alpha^{a_1+n-1} \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) \beta^{a_2-1} exp\{-\beta b_2\} exp\left\{-\alpha \left(\sum_{i=1}^{m} S_i^{\beta} T_i + b_1\right)\right\}.$$

The marginal posterior of β using the independent gamma priors is given by

$$\begin{aligned} \pi_{E-G}(\beta|\text{data}) &= \int \pi_{E-G}(\alpha,\beta|\text{data}) d\alpha \\ &= \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) \beta^{a_2-1} exp\{-\beta b_2\} \int_0^{\infty} \alpha^{a_1+n-1} exp\left\{-\alpha \left(\sum_{i=1}^{m} S_i^{\beta} T_i + b_1\right)\right\} d\alpha \\ &= \left(\prod_{i=1}^{m} S_i^{\beta n_i}\right) \beta^{a_2-1} exp\{-\beta b_2\} \left(\sum_{i=1}^{m} S_i^{\beta} T_i + b_1\right)^{-(a_1+n)} \Gamma(a_1+n) \end{aligned}$$

and the conditional posterior of α given β is

$$\pi_{E-G}(\alpha|\beta,data) = \frac{\left(\sum_{i=1}^{m} S_i^{\beta} T_i + b_1\right)^{a_1+n}}{\Gamma(a_1+n)} \alpha^{a_1+n-1} exp\left\{-\alpha \left(\sum_{i=1}^{m} S_i^{\beta} T_i + b_1\right)\right\}.$$

It can be shown that both the above joint posterior distributions are proper, it is clear that the conditional posteriors follow a gamma distribution. It can also be shown the marginal posteriors are log-concave, a condition for the Gibbs sampler.

2.2. Weibull Model

We assume that the life length X_i follows a Weibull distribution with scale parameter λ_i and shape parameter β. Assume that under the i^{th} accelerated test environment, the failure behaviour of the items can be described by a Weibull model with density given by

$$f(x_i|\lambda_i,\beta) = \beta\lambda_i x_i^{\beta-1} exp\left\{-\lambda_i x_i^{\beta}\right\}$$

with scale parameter, $\lambda_i > 0$ and shape parameter $\beta > 0$. We will denote the above model by $X_i|\lambda_i,\beta \sim Wei(\lambda_i,\beta)$. The model above implies that the scale parameter λ_i, depends on the stress environment, but the shape parameter, β, does not. This is a common assumption in the literature, see Soyer *et al.* (2008). Under the power law, the relationship between the failure rate and the stress level in the i^{th} testing environment is given by $\lambda_i = \theta_1 S_i^{\theta_2}$, where S_i denotes the i^{th} accelerated stress environment, $\theta_1 > 0$ and $\theta_2 \in \mathbb{R}$ are the unknown model parameters to be determined. This implies that $X_i|\theta_1,\theta_2,\beta \sim Wei\left(\theta_1 S_i^{\theta_2},\beta\right)$, and the density is given by

$$f(x_i|\theta_1,\theta_2,\beta) = \beta\theta_1 S_i^{\theta_2} x_i^{\beta-1} exp\left\{-\theta_1 S_i^{\theta_2} x_i^{\beta}\right\}.$$

Let D_i denote the test data from the i^{th} accelerated stress environment, that is, $D_i = \{n_i, r_i, x_i 1,\ldots,x_{iri}\}$, where n_i is the number of items tested, r_i is the number of failures observed during the observation period, x_{ij} is the time to failure of the j^{th} item under the i^{th} environment, $j = 1,2,\ldots,r_i < n_i$. Our main aim is to make inferences about the failure behaviour of the items at the use-stress environment, S_u. We assume that there is no censoring in the ALT, thus $r_i = n_i$. Assuming conditional independence of the failure times x_{ij} given the stress levels S_i, and the parameters θ_1, θ_2 and β, the likelihood function when k stress levels are used is given by

$$L(\theta_1,\theta_2,\beta|data) = (\beta\theta_1)^{\sum_{i=1}^{k} n_i}\left(\prod_{i=1}^{k} S_i^{n_i\theta_2}\right)\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i} x_{ij}^{\beta-1}\right) exp\left\{-\theta_1\sum_{i=1}^{k}\left(S_i^{\theta_2}\sum_{j=1}^{n_i} x_{ij}^{\beta}\right)\right\}.$$

Soyer *et al.* (2008) assumed that $\theta_1 \sim Gamma(0.01,0.01)$, $\theta_2 \sim Uni(0,100)$ and $\beta \sim Uni(0,10)$. It was further assumed that θ_1, θ_2 and β are independent. Under the power law, the joint prior distribution for $(\theta_1,\theta_2,\beta)$ is given by

$$\pi_{W-G}(\theta_1,\theta_2,\beta) \propto \theta_1^{0.01-1}\exp\{-0.01\times\theta_1\}\times\frac{1}{100}\times\frac{1}{10}.$$

The joint posterior distribution is therefore given by

$$\pi_{W-G}(\theta_1,\theta_2,\beta|data) \propto \beta^{\sum_{i=1}^{k} n_i}\theta_1^{\sum_{i=1}^{k} n_i+0.01-1}\left(\prod_{i=1}^{k} S_i^{n_i\theta_2}\right)\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i} x_{ij}^{\beta-1}\right)$$
$$\times exp\left\{-\theta_1\left(0.01+\sum_{i=1}^{k}\left(S_i^{\theta_2}\sum_{j=1}^{n_i} x_{ij}^{\beta}\right)\right)\right\}.$$

A vague prior can also be assumed, here the uniform prior will be considered.

$$\pi_{W-U}(\theta_1,\theta_2,\beta) \propto \text{constant}$$

The joint posterior distribution of θ_1, θ_2 and β using the uniform prior

$$\pi_{W-U}(\theta_1,\theta_2,\beta|data) \propto (\beta\theta_1)^{\sum_{i=1}^{k} n_i}\left(\prod_{i=1}^{k} S_i^{n_i\theta_2}\right)\left(\prod_{i=1}^{k}\prod_{j=1}^{n_i} x_{ij}^{\beta-1}\right) exp\left\{-\theta_1\sum_{i=1}^{k}\left(S_i^{\theta_2}\sum_{j=1}^{n_i} x_{ij}^{\beta}\right)\right\}.$$

It can be shown that both these posteriors are proper.

2.3. Expected Reliability

The steps involved for evaluating $R(x_u|D)$ is given by :

1. Sample θ_1, θ_2 and β from the posterior densities.
2. The expected reliability can then be computed using the Monte Carlo average of the posterior sample $\left\{\theta_1^{(j)}, \theta_2^{(j)}, \beta^{(j)}\right\}_{j=1}^{J}$ as $R(x_u|D) \approx \frac{1}{J}\sum_{j=1}^{J} R\left(x_u|\theta_1^{(j)}, \theta_2^{(j)}, \beta^{(j)}\right)$ which is the expected reliability at mission time x_u.

3. APPLICATION

We will make use of an example used in Soyer *et al.* (2008) and use the ALT data published in Nelson (1972). The data is given in Table 1 and represents the breakdown of an insulating fluid subjected to various voltage levels. The accelerated stress levels are given by 26, 28, 30, 32, 34, 36, and 38 Kv, where Kv represents kilovolts. We are interested in making inference at the use-stress of 22 Kv.

Table 1. Times to breakdown of an insulating fluid (in minutes) under various values of the stress

38 Kv	36 Kv	34 Kv	32Kv	30Kv	28Kv	26Kv
0.09	0.35	0.19	0.27	7.74	68.85	5.79
0.39	0.59	0.78	0.40	17.05	108.29	1579.52
0.47	0.96	0.96	0.69	20.46	110.59	2323.70
0.73	0.99	1.31	0.79	21.02	426.07	
0.74	1.69	2.78	2.75	22.66	1067.6	
1.13	1.97	3.16	3.91	43.40		
1.40	2.07	4.15	9.88	47.30		
2.38	2.59	4.67	13.95	139.07		
	2.71	4.85	15.93	141.12		
	2.90	6.50	27.80	175.88		
	3.67	7.35	53.24	194.90		
	3.99	8.01	82.85			
	5.35	8.27	89.29			
	13.77	12.06	100.58			
	25.50	31.75	215.10			
		32.52				
		33.91				
		36.71				
		72.89				

A comparison of the four models using the DIC is given in Table 2.

Model	DIC
Exponential gamma priors	614.7
Exponential vague prior	614.6
Weibull gamma priors	607.6
Weibull vague prior	607.5

The model using the Weibull distribution as the lifetime model and a vague prior gave the smallest DIC, and based on the DIC is the preferred model.

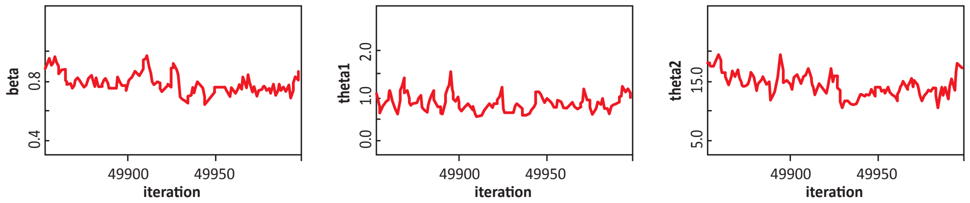


Figure 1. Trace plots of posteriors using uniform priors

From Figure 1 we see that for θ_1, θ_2 and β the chain is moving well around the parameter space and establishes convergence after 50000 iterations.

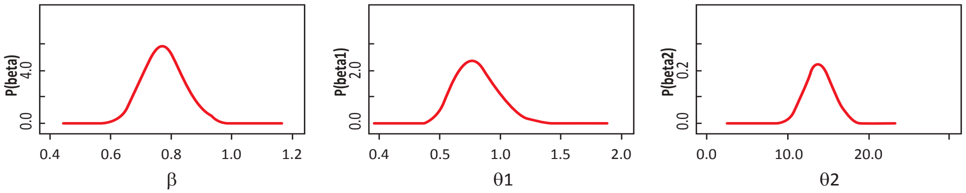


Figure 2. Posterior distributions of θ_1, θ_2 and $\boldsymbol{\beta}$, using the uniform prior.

The posterior distributions of θ_1, θ_2 and β when the uniform prior is used are given in Figure 2.

Table 3 shows the predictive reliability for various values of mission times. Having a closer look at the actual values, it is clear that the predictive reliability is higher when the vague prior is used.

Table 3. Predictive reliability function at use stress S_u = 22 Kv, when using the Weibull distribution

x	Weibull	
	Gamma priors	Vague prior
1	0.9994484	0.9994527
2	0.9990814	0.9990929
3	0.998761	0.9987798
.	.	.
51	0.9898242	0.9901594
52	0.9896757	0.990017
.	.	.
105	0.9825584	0.9832063
106	0.9824347	0.9830881

4. CONCLUSION

In this paper the power law was considered, using the Weibull distribution and the exponential distribution, respectively, as the lifetime models. Various priors were considered. These models were compared using the DIC for model assessment. The vague prior using the Weibull distribution as the lifetime distribution had the smallest DIC, and is the preferred model for this limited study. The predictive reliability was computed at the use-stress level for the Weibull distribution. The results were very similar for the two priors and it was found that the items were quite reliable at the use-stress level. The predictive reliability was found to be higher when a vague prior was used, a further indication that the model where the uniform prior is used is the preferred one.

REFERENCES

Chaloner, K., & Larntz, K. 1992. Bayesian design for accelerated life testing. *Journal of statistical planning and inference*, 33, 245 – 259.

Erkanli, A., & Soyer, R. 2000. Simulation-based ddesign for accelerated life tests. *Journal of statistical planning and inference*, 90, 335 – 348.

Escobar, L. A., & Meeker, W. Q. 2006. A review of accelerated test models. *Statistical science*, 21(4), 552 – 577. Mazzuchi, T. A., Soyer, R., & Vopatek, A. L. 1997. Linear Bayesian inference for accelerated Weibull model. *Lifetime data analysis*, 3, 63 – 75.

Nelson, W. 1972. Graphical analysis of accelerated life test data with the inverse power law model. *Ieee transactions on reliability*, R-21(1), 2 – 11. Singpurwalla, N. D., Castellino, V. C., & Goldschen, D. Y. 1975. Inference from accelerated life tests using Eyring type re-parameterizations. *Naval research logistics*, 22(2), 289 – 296.

Soyer, R., Erkanli, A., & Merrick, J. R. 2008. Accelerated life tests: Bayesian models. *Encyclopedia of statistics in quality and reliability*, 1, 20 – 30.

Van Dorp, J. R., & Mazzuchi, T. A. 2005. A general Bayes Weibull inference model for accelerated life testing. *Reliability engineering and system safety*, 90, 140 – 147.

West, M., Harrison, J., & Migon, H. S. 1985. Dynamic generalized linear models and Bayesian forecasting. *Journal of the American statistical association*, 80, 73 –97.

Alternative Techniques of Constructing Empirical Bayes (EB) Confidence Intervals for Hierarchical Negative Binomial Generalized Linear Model

Mohanad F. Al-khasawneh
Department of Mathematics, Statistic and Physics, Qatar University, Doha,2713 Qatar - alkhasawneh@qu.edu.qa

Abstract

The negative binomial (NB) distribution is commonly used to model count data in many areas of research. We use a parametric hierarchical empirical Bayes (EB) approach for estimating the parameters of (NB) generalized linear models (GLiM). The NB re-parameterization in terms of dispersion and proportion parameters is used by using a beta prior on the latter parameter. We link the covariate information to the prior mean of the proportions via a parametric link functions. We construct confidence intervals for the hyperparameters and the dispersion parameter using the beta-negative binomial marginal likelihood. We construct EB confidence intervals for the mean response parameters and the proportion as well using different techniques. We apply our procedure to real data set to emphasis its utility. A simulation study is then carried out to investigate the performance of the proposed interval EB techniques as compared to the usual NBGLiM estimates in terms of biases and MSEs.

Keywords: Negative binomial, Hierarchical empirical bayes, Estimation, Generalized linear models.

This page intentionally left blank

CPS07:
Survey Method (1)

Implementation of Big Data for Official Statistics in Indonesia

Setia Pramana
Institute Of Statistics, Jakarta, Indonesia – setia.pramana@stis.ac.id
Jonggun Lee
UN Global Pulse Jakarta Indonesia– jonggun.lee@un.or.id
Ricky Yordani, Budi Yuniarto, and Robert Kurniawan
Institute Of Statistics, Jakarta, Indonesia
Imaduddin Amin
UN Global Pulse Jakarta Indonesia

Abstract

Big Data is an umbrella term for explosion in the quantity and diversity of high frequency digital data and it is not usually coming from traditional sources. The speed and frequency by which data is produced and collected—by an increasing number of sources—is responsible for today's data deluge: the amount of available digital data is projected to increase by an annual 40%. Most of this data is being generated at no cost, in real-time, merely as a bi-product of people going about their daily lives. For the past decade, private sector firms have been driving the innovation of mining of such digital trails to understand the needs of their customers and the current market trends. Also they have been changing and improving their decision-making processes based on the mined information.

"Big Data for Development" is a concept that refers to the identification of sources of Big Data relevant to policy and planning of development programmes. It differs from both "traditional" development data and what the private sector and mainstream media call Big Data. Potential applicability of "Big Data for Development" at the most general level, when it is properly analysed, these new data can provide snapshots of the well-being of populations at high frequency, high degrees of granularity, and from a wide range of angles, narrowing both time and knowledge gaps.

This research discussed several possible implementations of Big Data to the official statistics in Indonesia. Furthermore, two case studies would be discussed: (1) developing a statistical model to nowcast food prices using multiple sources of data including social media i.e, crowdsourcing, and (2) predicting inter-city commuting patterns using social media, i.e., twitter.

The result shows similar trend between crowdsourcing approach and BPS Survey for all commodities. In case of commuting pattern, similar pattern between the twitter approach and the commuter survey 2014 is observed. From the study it is shown that the Big Data approach can provide faster, real-time and less expensive information on food prices and commuting behaviour. Similar approach can be implemented for other official statistics.

Keywords: Big data; Official statistics; Nowcasting; Commuting behaviour; Food price.

A Modified Dual to Ratio-Cum-Product Estimator for the Population Mean Using Auxiliary Information under Stratified Random Sampling

Rohini Yadav
Department of Statistics, Amity Institute of Applied Sciences, Amity University, Noida- 201313 (U.P.), India
Email: rohiniyadav.ism@gmail.com

Abstract

This paper deals with the problem of estimating the population mean of the study variable y with the help of the auxiliary variables under stratified random sampling. The estimator of dual to ratio-cum-product estimator for the population mean of the study variable has been suggested. The properties of the suggested estimator have been studied under large sample approximation. It has been shown that the suggested estimator is more efficient than other considered estimators. To judge the merits of the proposed estimator, an empirical study is given in the support of the present study.

Keywords: Auxiliary variable, Stratified random sampling, Dual to ratio estimator, Bias and mean squared error.

This page intentionally left blank

CPS08:
Macroeconomic Statistics (2)

Tourism Income Equality: Evidence of Bali Province

Central Bank of Indonesia
Stephanie Gunawan
Agni Alam Awirya
Putriana Nurman

Abstract

Amid the global economic dynamics, Bali Province is still able to show a good performance, even better than the performance of the Indonesian economy as a whole.The condition is driven by the dominance of the tourism industry which contributes more than 30% to the economy of Bali.Nevertheless, the positive development has not yet been felt by Bali citizens evenly, as seen from regency/city income per capita in Bali which is still not evenly distributed.Therefore, a study on the economic impact of tourism on Bali's economy as a whole and about equity in the Bali is needed.This study uses primary data survey to foreign tourists and businesses using Money Generating Model, as well as the gap analysis regencies/ cities and utilization of tourism opportunities.

The results of this study indicate that the economic impacts of tourism on the economy of Bali can be viewed from three aspects, namely sectoral output gain, output of labour, and output of wages/salary. In addition, there is a gap between regencies/cities in Bali, where preferences of foreign tourists visiting Bali and the business operators' opinion state that Badung, Gianyar and Denpasar contribute better to the economic growth of the local area. Nevertheless, there is room for improvement for equal income distribution as the results of tourism activities to peripheral region (edge region) which 'might' be reviewed on its implementation by the regional government of Bali.Therefore, the study recommends the need for increased government role in realizing the utilization of the tourism industry output to be distributed more evenly by providing jobs, the expansion of the development of tourist areas, promotions, support to innovation, and financing assistance.

Keywords: Regional economy, Money generating model, Tourism, Equality

1. BACKGROUND

Tourism's contribution to government revenues can be broken down into two, namely: direct and indirect contributions. Direct contribution is derived from tax revenue collected from the tourism workers and tourism businesses in the tourist area received directly by the department of revenue for a destination. While the indirect contribution of tourism to government revenues come from taxes or customs duties on goods imported and the tax charged to tourists who visit. Contributions of tax from Hotels and Restaurants sector (PHR) to total national GDP (Gross Domestic Product) in the third quarter of 2012 was 13%, but the contribution of PHR sector to the GDP in the province of Bali province was above 30% in the first quarter.

The majority of foreigners entering Indonesia through Ngurah Rai International Airport in Bali as the most popular vacation spot for foreign tourists in Indonesia.This island is the residence of the majority of Indonesian Hindu minority community and it is offering various types of Hindu Bali tourism related to art and culture and also a vibrant nightlife and beautiful rural areas.The second arrival point is Soekarno-Hatta International Airport, located close to the capital Jakarta.

Tourist arrivals to Indonesia from year 2007 - 2012 has increased from 5.51 million in 2007 to 8.04 million in 2012, an increase of 45.91% and total visits there, 3.87% is going to Bali. This shows the role of Bali as the main gateway to the Indonesian tourist visits is very important. Indonesia tourism development strategy also puts Bali as a tourism gateway and national food support.

However, the development of tourism in Bali remains concentrated in the region of South Bali especially Kuta and surrounding areas. BPS data in 2012 showed that more than 65%-star hotel located in Badung District in the southern part of Bali. Concentration of tourism activities in Bali also cause unevenness of development in Bali. The level of human development index Badung and Denpasar in Bali are the highest values respectively at 77.26 and at 80.45 in 2012. Therefore, efforts to equalize the yields of tourism in Bali is a challenge for policy-making.

This study aims to estimate the economic impact of tourism on Bali's economy as a whole. In addition, this study will also analyze the results of the income distribution of tourism in Bali as well as identify opportunities of a more equitable utilization.

The results are expected to be input for stakeholders in Bali, in decision making of Bali development in the future. In addition, this research is also expected to be useful as a reference for the development of tourist destinations in the regions that share similar characteristics with Bali.

2. PREVIOUS RESEARCH

Equality of tourism outcome has been a subject of previous studies. Haddad et al (2013) conducted an analysis of the regional imbalances in the Brazilian tourism. The results showed that the Domestic tourism can be an important channel for efficient natural resources allocation and for regional disparities reduction. Li et al (2016) showed that the development of tourism reduce regional inequality in China. Juan and Piboon Rungroj (2007) conducted a study on the Bai Ethnic Village which offers travel destinations in the form of a homestay. The results showed that although they have the potential for increased revenue and employment, the development of the region would lead to a worsening of income distribution pattern.

3. METHODOLOGY

Data

The data used in the analysis of this study are primary data and secondary data. Primary data were collected by distributing questionnaires to the respondents. There are two kinds of foreign tourists' respondents, with the number of respondents of 500 people and businesses as much as 200 respondents. The sampling method used was non probabilistic quota sampling. For business respondents, the largest proportion of respondents from regions that are further away from the center of tourism activities like Karangasem or Jembrana. As for foreign tourists respondent survey was conducted at the departure gate of Ngurah Rai international airport. The survey period is the month of July 2012. Secondary data were obtained from government publications and other private entities (journals, periodicals, textbooks and so forth).

Data Analysis Tools

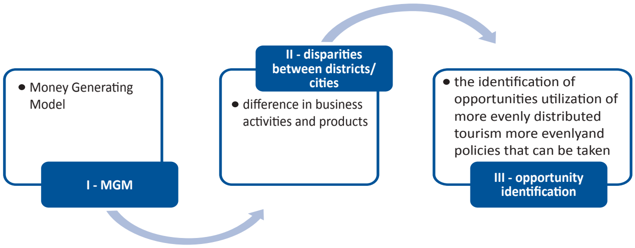


Figure 3.1. Stages of Data Analysis Tools

Analysis of the data that has been obtained is through the following three stages:

1. Method of Generating Money Model (MGM) estimates the economic impact of tourist spending. The formula used to estimate economic impact is as follows:

The impact of tourism on the economy = **The number** x **The average spending of tourists** x **Multiplier**

Multiplier figure is obtained from the input-output table of multiplier calculation of Bali in 2005 as has been done by Awirya (2009).

2. Analysis of the gap between districts / cities.This analysis is done by connecting the different business activities and products generated from tourism. With regard to the comparison between the preferences of tourists: travel destinations or info and entrepreneurs point of view: territoriality.
3. Identify opportunities of tourism utilization to be distributed (more) evenly and policies that (can be) taken.Al is obtained by comparing info about opinions of entrepreneurswith government involvement or through other news sources.

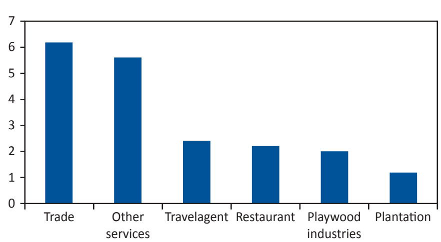
4. RESULT ANALYSIS

Economic Impact of Tourism Activities

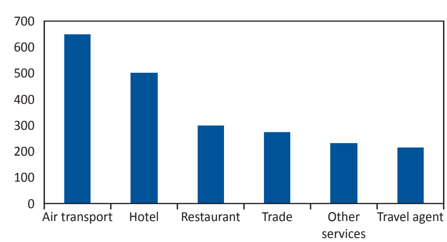
The economic impact of tourism on the economy of Bali can be seen from several things including the sectoral output, the number of employment as well as the establishment of wage/salary. The biggest impact of traveler spending to the increased sectoral output is of air transport.Sequentially from high to low, the sectors that experience the largest increase in output is air transport sector, hospitality sector, restaurants and food stalls, trade, other services, and travel agencies. Increased output 6 this sector could amount to 63.08% of the total increase in output.

As for the impact on the formation of the highest labor is the trade sector.The order of the economic impact of the results of tourism activity on the formation of labor is trade, other services sector, the travel agency sector, the sector of restaurants and food stalls, timber and plantation industries. The six sectors can absorb 86.11% of the total increase in the existing workforce.

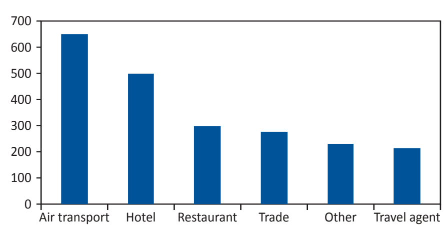
For the economic impact of the results of tourism activity on the formation of wages/salaries is currently on air transport to the highest order. After that, the sequence is other services, trade, travel agency, and sea freight. 6 sectors that influence the formation of the wage/salary is able to account for 84.22% of the total output.Here is a graphic visualization of the results of economic impact of tourism on the economy of Bali.



Graph 4.1. Labor Impact of Tourism



Graph 4.2. Output Impact of Tourism

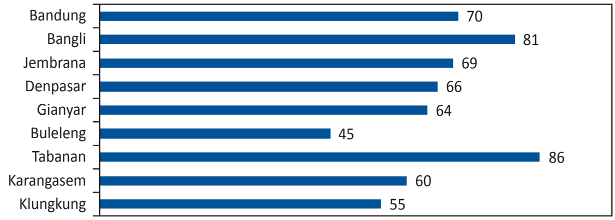


Graph 4.3. Wage Impact of Tourism

Inequality Tourism Activities and Peripheral Tourism Development Opportunities

The result of this analysis refers to the results of a questionnaire distributed to tourists and tourism businesses. The analysis shows that foreign tourists and businessmen stated that the territory of South Bali is still the center of tourism in Bali.However, there is little difference is that businesses found Gianyar regency become a major tourist destination after the Badung regency with a share of 42% and 17% of the total tourists coming, while tourists found only Badung Regency which became the center of tourism in Bali with a portion 60% of the total rating.

The survey results also shows that favorite tourist spot for foreign tourists are the beaches (stated by 23.19% of respondents), followed by shopping center (stated by 15.08% of respondents).Cultural tourism only ranks the fifth of travelers favorites. The results of the survey to businesses indicate that the share of annual turnover relating to the business activities is an average of 66%. The biggest share of businesses is in Badung with a share of 86%.This shows that only less than 15% of business turnover that is not derived from tourism activities. For areas that are far from the center of tourism activities like Karangasem the value is below the average that is only 60%. This shows that there isinequality from tourism yiled in Bali.



Graph 4.4. Share Omzet from Tourism Activity

The result of further analysis shows that there are some major points of inequality between regions as a result of activities in Bali that are:

1. The imbalance of development by policy makers in prioritizing areas of only southern Bali (stated by 51% of business operators).
2. The existence of an exclusive group who enjoy tourism activities and being in only certain regions (stated by 19% of businesses respondents).
3. The ability of each region to absorb the PAD is not the same as a result of tourism activities are centered so inequality is becoming increasingly broad (stated by 15% of businesses respondents).
4. Utilization of manpower from areas far from the center of tourism activity is still low so as to encourage high unemployment (stated by 9% of business respondents).

Meanwhile, expectations of businesses to policy makers in an order to equalize the results of tourism to all areas of Bali are as follows:

1. Opening jobs, especially in areas far from the center of Bali's tourism activities in the southern part (stated by 29% of respondents of business operators).
2. Increase the dissemination and promotion of tourism especially in the pheriperal area of tourism thus increase tourist visits to the area.Promotion also includes efforts to introduce products of the type of local area labor intensive products as to encourage employment (stated by 24% of respondents).
3. Balancing the development of infrastructure and facilities related to access to tourist destinations and facilities for supporting tourism in the pheriperal region (stated by 22% of respondents).
4. Simplify bureaucratic establishments in the peripheral region of the tourism industry in particular businesses that are environmentally friendly and still maintain the preservation of local culture. In addition, tt is also supported by the development of small and medium enterprises with the help of a credit program (stated by16% of respondents).

5. CONCLUSIONS AND RECOMMENDATIONS

Conclusion

Conclusions from the study regarding the equitable distribution of tourism study case of Bali are as follows:

1) The economic impact of tourism activities on the economy of Bali can be viewed from three aspects, namely sectoral output gain, output of labor, as well as the output of wages / salary, which significantly affects the performance of certain sectors.The air transport sector in Bali is the sector with the largest increase in output and increase in revenue due to spending foreign tourists.Meanwhile, the trade sector experiences the highest employment.
2) There is a gap between districts / municipalities in Bali region as indicated by the large share of the turnover of the tourism businesses in Badung compared to other areas. There are still opportunities for even distribution of tourism in Bali.Efforts to develop tourist destinations in the peripheral region may be one option to equalize the tourism activities in Bali.

Recommendations

1) Open business opportunities (jobs) that employ more local employment (labor intensive) (especially Bali).
2) Expand the development of another tourist areas which have not yet 'untapped' well.
3) Promoting a more vigorous and concrete actions, such as 'direct' traffic (especially at international events) on a new tourist destination (eg: Village Tourism).
4) Put forward and innovate local works, such as empowering local artists, local produce market, as well as developing the built-assisted SMEs featured in each area.
5) Allocate / distribute aid loans or CSR program that is better targeted and encourage the development of local products.

REFERENCES

Awirya, Agni Alam. 2009. Peningkatan Ekonomi Bali melalui Pengembangan Pariwisata. Jurnal Ekonomi dan Pembangunan. LIPI. Vol. XVII(1)

Haddad EA, Porsse AA and Rabahy WA (2013) Domestic tourism and regional inequality in Brazil. Tourism Economics: the Business and Finance of Tourism and Recreation 19(1): 173–186.

Juan, H. and Piboonrungroj, P. 2007. Tourism Development: A Case study of Homestay Service (Nongjiale) in Bai Ethnic Village. In: Kaosa-ard, M. ed. Mekong Tourism: Blessing for All?, Bangkok: White Lotus. pp. 138-166

Li, Hengyun; Jason Li Chen; Gang Li and Carey Goh (2016) Tourism and regional income inequality: Evidence from China: Annals of Tourism Research 5881–99.

Participation of Balinese Toward Tourism Do Government and Tourism Industries Affect Participation?

Eka N. Kencana
Department of Mathematics, Udayana University, Badung Indonesia - i.putu.enk@unud.ac.id

Abstract

This essay is aimed to study the influencing factors for community participation in tourism activities. Due to an important element for implementing tourism sustainability, participation of local people in tourism activities and its development had attracted many scholars. However, researchers attempted to analyze the causes of participation are very rare. In order to contribute for this gap, we studied the government's role and tourism industries behavior that are proposed as two antecedent factors for Balinese people participating in tourism at Kuta and Nusa Dua, two famous destinations in Bali. Applying hierarchical construct in structural equation model (SEM), data from 64 communities leaders at both destinations were gathered in June 2015 and analyzed. The results showed government's role significantly affects local participation and behavior of tourism industries. However, the behavior of industries did not show significant effect on local participation. In addition, participation of local people at both areas had effect in increasing the benefit and reducing the cost related to tourists' activities, significantly.

Keywords: Balinese, Hierarchical construct, Participation, SEM, Sustainability

1. INTRODUCTION

The latest paradigm for doing national development is its sustainability. This principle becomes mainstream for every country around the world do their national development. In spite of its popularity among policy makers, politicians, researchers, and other agents, there's still lot of debates regarding understandability of this concept. According to United Nations [1], variation in interpretation regarding sustainable development is the main factor for the less successful of its implementation although development agents such as governments, researchers, and the international organizations announce this paradigm intensively.

Sustainable development has been started its position in national development discourses and attracts many researchers since the publication of document entitled "Our Common Future" by Brudtland Commission, a task force developed by World Commission on Environment and Development (WCED) in 1987. In this report, the commission wrote

> "We see instead the possibility for a new era of economic growth, one that must be based on policies that sustain and expand the environmental resource base. And we believe such growth to be absolutely essential to relieve the great poverty that is deepening in much of the developing world."[2]

Refers to above note, efforts to sustain and expand the environmental resources become an important issue in national development process. In spite of the simplicity definition for sustainable development by WCED which stated "Sustainable development is development which meets the needs of the present without compromising the ability of future generations to meet their own needs.", according to United Nations in its report "The Millennium Development Goals Report 2014", threats for environmental sustainability tend to increase. For instance, globally carbon dioxide (CO_2) emissions increase almost 50 percent in 2011 compared to 1990 level [3].

In sustainable development process, people are the central agents. The national development will sustain only if its focus relies on people empowerment. According to Kates *et al.* [4], it's almost impossible to realize people empowerment without focusing on three dimensions of sustainable development, i.e. (a) economic, (b) socio-cultural, and (c) environmental dimension. We believe to enhance the focus of development agents, roles and responsibilities of these agents in the development process have to be re-coordinated. Coordination is a key factor in determining the successfulness of sustainable development in various sectors [5].

Tourism as an industry also adopts the concept of sustainable development. The World Tourism Organization stated sustainable tourism development is a development that is conducted to fulfill visitors' and communities' needs while keeps and expands the development opportunities in the future [6]. Refers to this definition, it is clear visitors and local communities are the primary concern in tourism development activities. In the last decade, tourism has big impacts on Indonesian's economic and social development. According to Mr. Arief Yahya, Tourism Minister of Indonesia, direct contribution of tourism for Indonesian GDP in 2014 as much as 4.01 percent. In addition, Indonesian tourism also generated US$11.17 million and 10.32 million people works in tourism and related sectors [7]. According to World Travel & Tourism Council, Indonesian rank for relative importance of travel and tourism total contribution to Gross Domestic Product (GDP) is 17 out of 184 countries around the world [8].

For Bali and Balinese people, tourism has a long story. Tourism in this small island had been started in 1969 when the former President of Indonesia, Soeharto, announced the First Phase of 5 Year Plan or *Rencana Pembangunan Lima Tahun* (REPELITA) Tahap I. The number of foreign tourist visited Bali in 1969 is recorded as much as 11 278 visitors or approximately 13.10 percent from tourists' visit to Indonesia. These figures increase dramatically to 4 001 835 at the end of year 2015, approximately 38.45 percent from foreign visitors to Indonesia that is recorded 10 406 759 visitors. Figure 1 displays number of tourists visited Bali and Indonesia for period 1969-2014.

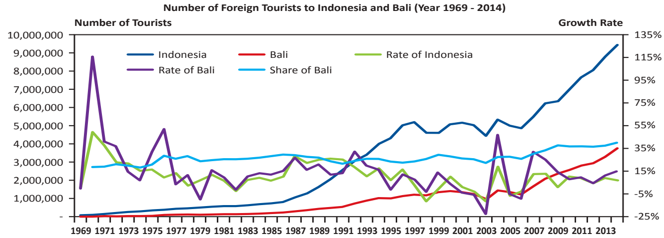


Figure 1. Number of Foreign Tourists to Bali & Indonesia, 1969-2014 (Source: Bali Tourism Office, 2015)

In spite of tourism and its related sector's to contribute in Bali's economic, there is no doubt some negative impacts also emerge from tourism development, especially for socio-cultural and environmental dimensions. Our research that is conducted in 2014 showed people at Kuta and Nusa Dua areas, two central tourist destinations in Bali, felt unsatisfied regarding their safety and comfort. Local people judged tourism is the cause for the raising of criminality and other social destructions at their villages. If these negative impacts could not be solved wisely, these will affect communities' evaluation regarding the benefit and burden cost of tourism.

Community-based tourism (CBT), in a simplified manner, can be understood as local people taking care of their natural and socio-cultural resources in order to gain benefits from tourism activities [9,10]. From this perspective, we assure participation of local people on tourism in their own backyard is vital to realize sustainable tourism. This article is aimed to elaborate the effects of government and tourism industries at Kuta and Nusa Dua, Bali on local community participation on tourism activities at both areas. In addition, this work is directed to study participation's effect on benefit and cost of tourism activities.

2. LITERATUR REVIEW

Community-based development and its variant, community-driven development, according to Mansuri & Rao [11] is a development in which the communities have direct controls over project decisions. It has potential gains for community such as enhancing project sustainability, poverty reduction, increasing efficiency and effectiveness, empowering poor people and many more. These gains arise from changing the passive roles of community to be actively involved and voiced in development processes. Community-based development relies on community to use and organize their social capital to (actively) participate in development processes.

Community participation is a concept, which is almost impossible to define it precisely. First, participation itself does not have a universally valid definition because of it is used widely and scope is very comprehensive

[12,13]. Secondly, United Nations (1985) cited in Tosun [13] stated the term community participation "can best be understood in the context of specific country and its political and socio-economic system". This suggests that participation should be approached in terms values and norms of the participants. In this context, local culture is an important media in determining successfulness of community participation. Finally, community participation is not a flat concept but rather is a categorical term. According to Tosun, participation can be group in three categories namely: (a) coercive, (b) induced, and (c) spontaneous participation [13].

According to Connell (1997) cited in Okazaki [14], community participation in tourism planning "is not only about achieving the more efficient and more equitable distribution of material resources: it is also about the sharing of knowledge and the transformation of the process of learning itself in the service of people's self-development. Viewed from Connell's perspective, equitable distribution regarding the benefit and cost of tourism development was not the only objective, but sharing the knowledge and learning process to local people regarding tourism development also be important goals in sustainable tourism development. Two actors that are responsible for sharing tourism knowledge and educated the local people are local government and tourism industries.

Government and tourism industries have important roles to realize sustainable tourism because government has political power and business sectors control capital needed in tourism development. Thus, to empower the local community through sharing the knowledge and opportunities in tourism activities, government and business' roles become very important. Our latest study conducted at Kedonganan Beach, center of sea food culinary tourism in Bali, showed local government roles that is formed by distributive (financial and infrastructure) and regulative roles had significant effect on people's economic condition. However, these roles did not affect socio-cultural dimensions and communities' satisfaction [15].

3. RESEARCH METHOD

Instrument, Population, and Research Sample

For assessing the causal relationships among latent variables in our model, we applied quantitative modeling by using variance-based structural equation model (PLS-SEM). An instrument was built to collect data from 64 community leaders at Kuta and Nusa Dua. Questionnaire with 5 Likerts' scaled option for every closed item was designed. We choose community leaders at both areas as data source because they represent the local people. Table 1 lists number of respondents at both areas:

Table 1. List of Respondents

AREA	DISTRICT	NUMBER OF VILLAGES	NUMBER OF RESPONDENTS (PEOPLE)				
			VILLAGE'S LEADER	RELIGIOUS LEADER	LOCALS REPRESENTATIVE	OTHERS	TOTAL
NUSA DUA	KUTA SELATAN	5	5	8	2	6	21
KUTA	KUTA	5	5	6	4	9	24
	KUTA UTARA	4	4	4	3	0	11
	MENGWI	3	3	5	0	0	8
BOTH AREAS		17	17	23	9	15	64

Conceptual Research Model and the Hypotheses

To answer the research questions, we design our research model and build five hypotheses as depicted in Fig. 1. In our model, tourism industries behavior and local community participation is positioned as endogenous as well as exogenous constructs. Five hypotheses were built to elaborate causal relationships between latent variables, i.e.:

H1 : Government roles positively affect industries behavior at Kuta and Nusa Dua areas for promoting tourism sustainability;

H2 : Tourism industries behavior at Kuta and Nusa Dua areas positively affect local community participation for promoting tourism sustainability;

H3 : Government roles positively affect local community participation for promoting tourism sustainability;

H4 : Local community participation positively affects the positive impacts of tourism development at Kuta and Nusa Dua areas; and

H5 : Local community participation negatively affects the negative impacts of tourism development at Kuta and Nusa Dua areas.

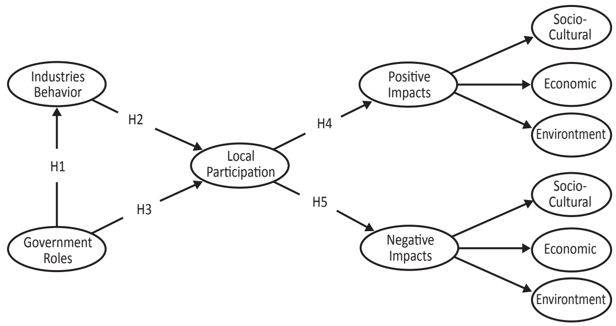


Figure 2. Conceptual Research Model

Data Analysis

Before the questionnaire was used to collect data, we tested it on pilot study conducted at Sanur Beach. Leaders at this area as much as 28 people were randomly selected as our respondents. Collected data were analyzed to assess item's validity and construct's reliability. According to Nunnaly [16], set of items are considered has internal consistency if its Cronbach's alpha is equal or greater than 0.70 as threshold value although for exploratory research, as long as the alpha coefficient value is greater than 0.60 is acceptable [17]. Second criterion to assess quality of questionnaire is item's validity. An item is assumed valid if its correlation with total items of the same construct is greater than 0.30 as the lowest limit [18]. If the questionnaire proved qualified enough, then research data were collected and analyzed by applying PLS-SEM.

4. RESULTS

Profile of Respondents

Descriptively, our respondents are male (87.5 percent) and has been completed their diploma's education. Their ages in the range of 28 – 69 years old with the average is 48.13 years. All of our respondents have been living at Kuta or Nusa Dua areas at least for 15 years with the average are 47.18 years. From these facts, we concluded that our respondents were properly representative of local people at Kuta and Nusa Dua areas.

The Quality of Questionnaire

As mentioned before, a pilot test was conducted to assess quality of questionnaire by distributing it to 28 community leaders in Sanur area. Five latent variables with 38 items in total were examined their reliability and items' validities. Table 2 lists validity and reliability's test results using SPSS 20:

Table 2. Validity and Reliability's Test Results

Latent Variable	Item		Correlation
Positive Impacts α = 0.773	TB1	Tourism creates job for local community	0.672
	TB2	Tourism creates business opportunitiesfor community	0.574
	TB3	Tourism increases income for community	0.592
	TB4	Tourism increases quality of public infrastructure	0.555
	TB5	Tourism increases quantity of public infrastructure	0.614
	TB6	Stronger appreciation of the public on their traditional values	0.220
	TB7	More protected and preservation of Hindu shrines	0.362
	TB8	The growing of public concern for environments	0.479
Negative Impacts α = 0.906	TC4	The increasing of waste and trash that is unmanaged	0.658
	TC5	Deterioting of public spaces	0.661
	TC6	The increasing of water pollution	0.801
	TC7	The increasing of air pollution	0.658
	TC8	The decreasing of life comfort of localpeople	0.698
	TC9	The decreasing of agricultural land because of conversion	0.530
	TC10	The eroding of traditional activities	0.582
	TC11	Worsening social behavior of local people	0.544
Negative Impacts	TC1	The rising prices for staple goods	0.482
	TC2	The rising prices of properties and rental houses	0.433
	TC3	The increasing of business competition	0.599
Govern-ment Role α = 0.930	GO1	High commitment to accommodate people voice	0.799
	GO2	Combine top-down and bottom-up approach	0.796
	GO3	Effective regulation has been developed	0.815
	GO4	Effective regulation has been implemented	0.869
	GO5	Enhancing local participation in tourism	0.802
	TC12	The waning of traditional building design	0.521
	TC13	The commercialization of the local culture	0.465
	TC14	Rising crime rates	0.674
	TC15	Increasing public discomfort	0.634
Local Partici-pation α = 0.943	CP1	Local community get involved in tourism planning	0.830
	CP2	Local community get involved in monitoring tourism development	0.816
	CP3	Local organization get involved in tourism planning	0.923
	CP4	Local organization get involved in monitoring tourism development	0.888
Indust ries	IB1	Prioritization of local people for	0.592

Source: Own primary data (2015)

Table 2 shows all latent variables on the model have alpha coefficient greater than threshold value suggested by Nunnaly [16]. The least and the greatest alpha's was found on tourism benefit's and local participation constructs as much as 0.773 and 0.943, respectively. In addition, excepts for TB6 on tourism benefit construct, all items on respective constructs have correlation values greater than 0.30, which indicated theses constructs were appropriately measured. By eliminating TB6 as indicator for tourism benefit construct will change alpha coefficient from 0.733 to 0.799, we decided to remove it from our questionnaire. Noting these facts, we conclude the instrument is qualified to use in final data collecting. The operational model in our research can be depicted in Figure. 3:

Measurement Model Analysis

Prior to assess the relationships between latent variables, we conducted measurement model analysis to evaluate every construct in term of their indicators as suggested by Hair *et al.* [17] and Hox & Bechger [19]. In measurement or outer model analysis, convergent validity of construct is assessed by evaluating its average variance extracted (AVE) and every indicators of respective construct is assessed by its outer loading for reflective indicators or its outer weight for formative indicators. A construct is argued achieves convergent validity if its AVE is greater than 0.50 [20] and its indicators have outer (loadings or weights) values greater than 0.60 [17] or significant [20]. Table 3 lists the AVEs, outer values and their significance level, and composite reliabilities (CRs) for latent variables in our model:

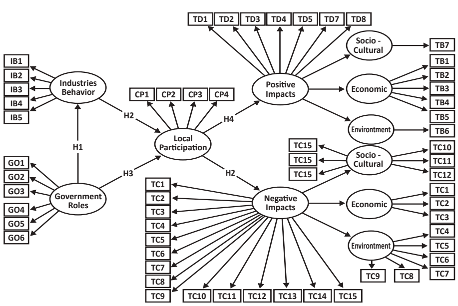


Figure 3. Operational Research Model

Latent Variable	AVE	CR	Item	Outer Value	p-value
Government Role	0.742	0.945	GO1	0.856	0.000
			GO2	0.855	0.000
			GO3	0.877	0.000
			GO4	0.914	0.000
			GO5	0.877	0.000
			GO6	0.784	0.000
Industries Behavior	0.569	0.868	IB1	0.720	0.000
			IB2	0.779	0.000
			IB3	0.819	0.000
			IB4	0.795	0.000
			IB5	0.646	0.000
Positive Impacts	0.537	0.884	TB1	0.888	0.000
			TB2	0.795	0.000
			TB3	0.816	0.000
			TB4	0.728	0.000
			TB5	0.810	0.000
			TB7	0.285	0.102
			TB8	0.638	0.000
Local Participation	0.855	0.959	CP1	0.902	0.000
			CP2	0.892	0.000
			CP3	0.961	0.000
			CP4	0.944	0.000
Negative Impacts	0.442	0.921	TC1	0.578	0.000
			TC2	0.541	0.000
			TC3	0.699	0.000
			TC4	0.714	0.000
			TC5	0.715	0.000
			TC6	0.859	0.000
			TC7	0.745	0.000
			TC8	0.743	0.000
			TC9	0.618	0.000
			TC10	0.617	0.000
			TC11	0.573	0.000
			TC12	0.560	0.000
			TC13	0.489	0.001
			TC14	0.723	0.000
			TC15	0.695	0.000

Table 3. Results For Measurement Model Analysis

On the constructs level, Table 3 showed negative impacts as a latent variable had AVE slightly less than 0.50 as suggested. However, noting that all of its indicators were significant, we concluded this construct is fairly reflected by its indicators. In addition, we decided to keep TB7 as an indicator for positive impact latent variable even though its loading value is less than the threshold and insignificant, because by excluding this item will cause socio-cultural latent has no indicator. Based on these findings, we concluded the inner or structural final model is worth to be analyzed.

Inner or Structural Model Analysis

The second step in applying PLS-SEM is doing an inner or structural model analysis. Inner model is a sub-model in PLS-SEM that reflects the causal relationships among latent variables [17]. Noting PLS-SEM assumes free-distribution for variables on the model, inner model assessment is conducted by applying non-parametrical technique through bootstrapping procedure available in SmartPLS 3.0 [21]. Assessment of inner model evaluates the AVEs and R^2s of endogenous latent variables, significance of path values, and goodness of fit (GoF) of model. According to Chin [22], threshold values for R^2 to say the variance of construct is weak, moderate, or substantial are 0.19, 0.33, and 0.67, respectively. Table IV represents the R^2s, AVE, and number of indicators for latent variables in our final model:

In PLS-SEM, it is possible to make a hierarchical model through the repeated use of manifest variables and a higher-order latent variable can be created by specifying lower-order latent, and so forth. A higher-order latent variable is represented (formatively or reflectively) by all of manifest variables in its underlying lower-order latent variables [23]. Hierarchical construct model that also is known as multidimensional constructs model. In our model as depicted in Figure 2, we had nine first-order and two second-order latent variables. Table 4 lists the assessment results of structural or inner model analysis.

Table 4. Results for Structural Model Analysis

Endogenous Latent		Number of Items	R2	Chin's Criterion for R2	AVE
Second Order	First Order				
Positive Impacts	Industries Behavior	5	0.550	Moderate	0.569
	Local Participation	4	0.263	Weak	0.855
Negative Impacts		7	0.183	Weak	0.537
	Economic	5	0.944	Substantial	0.693
	Environment	1	0.407	Moderate	1.000
	Socio - Culture	1	0.081	Very weak	1.000
		15	0.077	Very weak	0.442
	Economic	3	0.542	Moderate	0.692
	Environment	6	0.843	Substantial	0.643
	Socio - Culture	6	0.700	Substantial	0.542

Source: Analyzed from primary data (2016)

In spite of PLS-SEM does not permit us to assess the global goodness-of-fit (GoF) as found on covariance-based SEM (CB-SEM), Tenenhaus *et al.* [24] provide a formula that can be used as proxy for global GoF, as follows: $\widetilde{\text{GoF}} = \sqrt{\overline{AVE}.\overline{R^2}}$ (1)

In eq. (1), $\overline{AVE}$ and $\overline{R^2}$ are weighted average with weight is number of items of enndogenous latent variable. By eliminating AVE and R^2 values for environment and socio-cultural positive impacts in calculating eq. (1) because both constructs had only one item (Vinzi *et al.*, 2010 cited in Henseler & Sarstedt [25]), eq. (1) gave us $\widetilde{\text{GoF}}$ value as much as 0.477 and we argued this value is sufficient to conclude the model is quite good to predict the causal relationship among government roles, industries behavior, communities participation towards tourism development, and impacts of tourism activities. The structural model with its estimates for these relationships is shown on Figure. 4.

5. DISCUSSION

Refers to inner model's results, we found four out of five hypotheses were accepted. Local government roles proved significantly affect tourism industries behavior (H1) through tourism regulation had been made and implemented, and significantly affect local community participation (H3) by showing its commitment to hear people voice, enhancing local community participation, and empowering local organizations. These findings support previous research by Pillora & McKinlay [26] that stated "… it is important to engage communities in decision-making over the delivery of important services at the local level." According to Organisation for Economic Co-Operation and Development (OECD), good governance is a concept that has come to regular use in various discourses as well as in sustainable development. Good local governance that should be initiated by local government will assures the corruption is minimized, the views of minorities are considered, and their voices are heard in decision-making process [27]. From the acceptance of H1 and H3, we concluded the local government roles had been properly regulated tourism industries at Kuta and Nusa Dua areas and positively affect people participation as the cornerstone of good local governance. In spite of some regulations have been made related to tourism industries, industries behavior did not significantly affect local participation. We rejected H2 that stated industries behavior positively affect local participation for promoting tourism sustainability. This finding was not inline with Ertuna & Kurbas' results whom studied local community involvement in rural tourism development. One finding of their research is local business seems to stimulate the rural communities to participate in tourism activities [28]. We argue at least two causes regarding this difference. First, type of business owner. Most of tourism business at Kuta and Nusa Dua areas were owned and controlled by non-local people while Ertuna & Kurbas studied business owned by local people. Second, while Ertuna & Kurbas studied new business had been developed by local people, most of tourism industries at Kuta and Nusa Dua areas were in mature stages. Owner and stages of tourism businesses will affect their perception regarding local communities and vice versa.

In addition, we also found that H4 and H5 hypotheses were accepted. Local community participation proved significantly affects tourism impacts on both sides; the positive and negative impacts. On the positive impacts of tourism, the more local people get involved on tourism development the more positive impacts they may feel and the most aspect they perceive was reflected on economic benefit follows by environtment benefit. Besides of it, local participation will reduce the negative impacts of tourism development. Our findings about the effects of local participation on tourism impacts are inline with some similar researches. Yaman and Mohd [29] in their work stated by empowering local people, community-based tourism that is characterized by participation of local people will reduce negative impacts and concentrate the benefits of tourism locally. As noted by Tosun that the more local communities benefit from tourism, the more likely they will be to protect area's natural and cultural heritage [30].

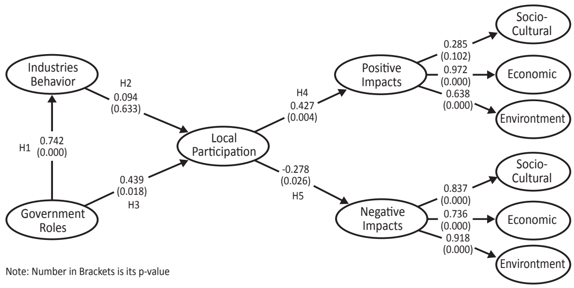


Figure 4. Structural Model with Estimates

From the causal relationhips between positive impacts of tourism as first-order latent with its second-order latent variables, we found socio-cultural latent was not significantly affected. This means that tourism does not yet give some benefits for the protection and preservation of Hindu shrines at both areas. However, economic and environtment dimensions are positively affected by positive tourism impacts. Job opportunities for local people and the increasing of income are two economic's indicators that were mostly influenced. Furthermore, reflections of negative impacts of tourism were significant on all of its dimensions where environtmental and economical aspects were the biggest and the least influenced. Our findings in evaluating tourism impacts at Kuta and Nusa Dua are inline with previous research by Leksakundilok, noted that in promoting community- based (eco)tourism, development agents should not focus only on negative impact to environtmental but also important to consider the impacts for social and cultural fabric of local communities [31].

6. CONCLUSION AND RECOMMENDATION

Community participation is key factor in promoting sustainable tourism. In spite of its importance, the determinants of community participation and their effects are quite scarce. In this research we showed local government roles significantly affect local community participation and tourism industries behavior whereas industries behavior does not affect people participation regarding the sustainability of tourism at Kuta and Nusa Dua areas, southern of Bali. The formal and informal leaders at both areas perceive local government has been successfully regulates tourism industries to do their businesses properly relates to implement sustainable tourism development. Furthermore, village leaders also appreciate local government in promoting participatory tourism development by combining government wants and communities needs, hearing and considering people voice, and empowering local bodies to take parts in tourism activities. However, behavior of tourism businessess at these area does not successfully affect local participation. We believed this insignificant effect arose from facts i.e. (a) most of tourism businessess are owned and controllod by non-local people, and (b) tourism industries at these area are in mature stages.

Refers to the findings of our work, we suggest tourism development agents at Kuta and Nusa Dua areas to focus their efforts and resources, in building sustainable-community-based tourism (SCT), on positive and negative impacts of tourism in all of three sustainable dimensions. In order to enhance and to empower local people capabilities in tourism activities, social structures and cultural fabrics are important aspects to consider as well as to increase their abilities to apply soft technologies such as knowledge, information, financial, and entrepreneurial capabilities in tourism activities and development.

REFERENCES

Bass, S., Dalal-Clayton, B. & Pretty, J., 1995. *Participation in Strategies for Sustainable Development*. London, UK: International Institute for Environment and Development.

Chin, W.W., 1998. The partial least squares approach to structural equation modeling. In G.A. Marcoulides, ed. *Modern Methods for Business Research*. Mahwah, NJ: Lawrence Erlbaum Associates. pp.295-358.

The Use of Passenger Exit Survey to Estimate Tourist Spending as Driver of Regional Economy

Agni Alam Awirya*
Bank Indonesia, Sibolga Representative Office, Indonesia – agni_aa@bi.go.id
Elisabeth R.S.Y. Silitonga
Bank Indonesia, Pematangsiantar Representative Office, Indonesia – elisabeth_rs@bi.go.id
Meita Elshinta Siagian
Bank Indonesia, Pematangsiantar Representative Office, Indonesia – meita_es@bi.go.id

Abstract

Tourist visit can be considered as one of the economic driver in the intended travel destination. The tourist spending, in form of transportation, accommodation, food and beverages and any other spending, have impact in increasing demand that also increases Gross Regional Domestic Product (GRDP) and employment. Attracting the high-spending tourists can be a strategy to improve the local economy. The purpose of this study is to estimate tourists' characteristics that determine tourist spending in North Sumatra and the economic impacts of tourist spending in North Sumatra. The study result could be an input in formulating promotional strategy in tourist destination for the segmented tourists. The results of analysis using data of Passenger Exit Survey (PES) of North Sumatra in 2014 indicates that most tourist spending are allocated to Hotel and Restaurants subsector and this subsector also generates the highest increase in output. The analysis results also reveal that tourist who has leisure purpose might spend more, meanwhile length of stay and numbers of travel partner negatively affect the higher spenders. On the other hand, age, sex and frequency of visit are insignificantly to the tourist spending.

Keywords: Tourist; Spending; Economic effect.
JEL Classification: C83, C21, Z32.

1. INTRODUCTION

The global economic slowdown that began from the crisis «Subprime Mortgage» in America in 2008 have an impact on national and regional economic slowdown, especially given the current economy is still dominated by sectors based on natural resources or commodities. It is also experienced by North Sumatra whose economy is still dominated by commodity-based sectors, namely oil, rubber and coffee. Therefore, to boost the economy, the government should start looking for an alternative of non-commodity based sectors, one of which is the tourism industry. The tourism industry in Indonesia has been proven to not be affected by the global economic slowdown, as seen from indicators of tourist visits (tourists), the average spending of tourists per visit, and revenue from foreign tourists that still show an increasing trend since the global crisis to date.

Correspondingly, to support the tourism industry in North Sumatra, the government has launched the Lake Toba to became one of the 10 strategic important national tourism areas of Indonesia. The declaration of this should be supported by policies to improve the performance of the tourism industry that attracts tourists, especially foreign tourists to visit North Sumatra. The expected increase is not only from the number of tourists, but also from an increase in the number of tourist spending, given the tourism industry through tourist spending is believed to provide a substantial contribution both to national growth (Brida & Pulina 2010; Figini & Vici, 2010; Lee & Chang, 2008) as well as regional growth (Cortes-Jimenez, 2008; Paci & Marrocu, 2014). In addition, tourism industry also has been linked to a variety of service industries such as accommodation, restaurants, food processing industries, transportation, entertainment industry and other services (Paton, 1985: 64).

Therefore, it is necessary for the policy to encourage the increase of tourists in North Sumatra especially travelers with high spending levels. In order to support this, there should be a research conducted that aims to determine the factors that affect the tourist expenditure in North Sumatra and its impact on the economy. This study was conducted based on survey results of Passenger Exit Survey (PES) conducted by Bank Indonesia Representative Office of North Sumatra in 2014.

Research on the factors affecting tourist expenditure has been studied in several papers, such as Kim (2011) conducted research in Macau, China and concluded that the tourists who were married, had a higher education background and household income tend to have higher expenditure. Marcusse (2011) found that the type of accommodations, length of stay, travel party size, destination, activities, package tours, income and first-time versus repeat visitors will significantly affect the tourist spending in Denmark.

This study also took the topic of traveler spending, especially in analyzing factors that become determinants for travelers greater spending in North Sumatra. In addition to estimating the determinants of the size of traveler spending, this analysis also expanded on the initial estimate of the impact of tourist spending for the improvement of economic output as well as an analysis of the characteristics of the tourists who come to North Sumatra.

Results are expected to be used as input in the policy development of the tourism industry, especially in North Sumatra. Knowledge about the characteristics of travelers is expected to be a better base in an attempt to increase the number of tourist visits. Target travelers with greater spending may impact more effectively to the improvement of the regional economy.

2. LITERATURE REVIEW

Tourism Definition

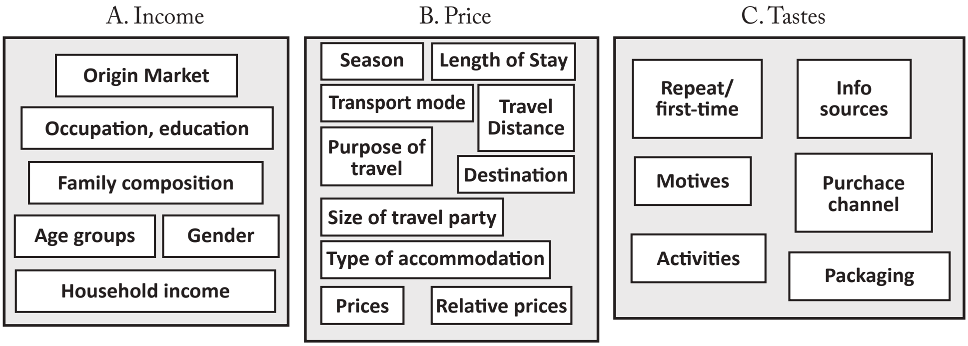
Tourism can be defined as "the activity of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business or any other purpose" (WTO, 1999). Anyone who spends at least one night away from home is thus classified as a tourist. When travel takes place within the country of residence, it is referred to as domestic tourism (WTO, 1999). While international tourism usually refers to all tourists from foreign countries. This paper uses the term "tourism" for tourists from foreign countries. According to Economic Commission of the League of Nations, tourist is defined as any person travelling for a period of 24 hours or more in a country other than that in which he usually resides.

Tourist Spending Definition

Tourist spending is defined as the total consumption expenditure made by a tourist or on behalf of a tourist for and during his/her trip and stay at destination. It encompasses a wide variety of items, ranging from the purchase of consumer goods and services inherent in travel and stays to the purchase of small durable goods for personal use, souvenirs and gifts for family and friends. Tourist spending can be measured in at least four different ways: Per person per night, per person per stay, per travel party per night, or per travel party per stay (Kozak et al., 2008).

Determinants of Tourist Spending

Tourist Spending are a function of income, prices (tourism prices and relative prices) and tastes, as shown in Figure 1. This is in accordance with economic theory, which suggests that the quantity demanded of a good or service is a function of the relative price of the commodity, the consumer's income and the consumer's tastes (Farrell, 1953-54; Downward and Lumsden, 2000).



Source: Marcussen, 2011

Figure 1. Determinants of Tourist Spending

Economic Impacts of Tourist Spending Full Ship

The tourist spending is considered to affect economics through:

a. Foreign Exchange Earnings: tourist visit to a tourism destination enhances the foreign exchange business for transactions during visit.

b. Contributions To Government Revenues, direct contributions are generated by taxes on incomes from tourism employment and business, meanwhile indirect contributions come from taxes and duties on goods and services supplied to tourists.

c. Employment Generation: tourism can generate jobs directly through hotel, restaurants, nightclubs, taxis or souvenir sales and indirectly through the supply of goods and services needed by tourism-related business.

d. Infrastructure Development: tourism can induce the local government to improve infrastructure, such as better water and sewage systems, roads, electricity, telephone and public transportation.

e. Development of Local Economies: tourism can improve local revenues through informal employment such as informal guides, rickshaw drivers or street vendors.

The previous studies explore variables that influence tourist spending in some countries as follows:

Table 2.1. Previous Studies on Determinants of Tourist Spending

No	Tittle	Object of Research	Methodology	Conclusion
1	Factors Affecting The Expenditure of Domestic and Foreign Tourists (Jurdana, Frleta (2016))	Rijeka and Opatija, Croatia	Descriptive Statistic, Regression	The increase of family income and tourists' satisfaction with facilities will positively and significantly affect the foreign tourists spending. On the other hand, the increase of duration of stay will negatively affect the foreign tourists spending. While education level, gender, past behaviour, intention to return and satisfaction with information were insignificant in model.
2	Determinants of Tourist Spending – in cross-section studies and at Danish destinations (Marcussen, 2011)	Denmark	Descriptive Statistic, Regression	Type of accomodation, length of stay, travel party size, destination, activities, package tours, income and first-time versus repeat visitors will significantly affect the tourist spending in Denmark.
3	Determinants of visitor expenditures in mountain tourism (Fredman, 2008)	Swedish mountain, Sweden	Descriptive Statistic, Regression	Duration of stay, household income, choice of activity, occupation, participation in an organized trip and choice of accommodation have an effect on expenditures at the destination, while gender, mode of travel, distance and 'significance of activity' are additional determinants of expenditures outside the destination.
4	Factors affecting the travel expenditure of visitors to Macau, China (Kim, 2011)	Macau, China	Descriptive Statistic, Tobit Analysis	Tourists who were married, had a higher education background and household incomeand were in a large party tended to spend more on total travel-related expenditures.
5	Peningkatan Ekonomi Bali Melalui Pengembangan Pariwisata (Awirya, 2009)	Bali, Indonesia	Descriptive Statistic, Money Generating Model	The highest increasing output pushed by tourism consumption are hotel and restaurant sub sector, followed by textile industry sub sector.

3. METHODOLOGY

Sample and Data Collection Procedure

Data used in this study can be classified as primary and secondary data. The secondary data used are input-output tables of North Sumatra in 2005, published by Badan Pusat Stastistik, a central bureau of statistics in Indonesia. Primary data are generated from passenger exit survey (PES). The respondents are foreign pessengers in International Departure gate in Kualanamu International Airport, North Sumatra. The foreign passengers are assumed as tourists who visit North Sumatra. The PES uses questionnaire to obtain information such as age, sex, nationality, country of residence, name of entry and exit point to/from Indonesia, frequency of visit, main purpose of visit, type of accomodation, total and type of spending, length of stay and kind of souvenir spending. The survey used in this study is based on PES of North Sumatra in October 2014, in which 452 questionnaires were gathered. To select the samples for this study, non probabilistic sampling is used.

Data Analysis

Data analysis in this study is divided into three parts which are descriptive statistics, economic impact of tourist spending and regression analysis.

The first part highlights the characteristics of tourist in terms of gender, age, occupation, origin and other characteristics. Those characteristics are presented in form of bar or pie charts.

The second part estimates the effects of tourist spending on the increase of output (GDP) using input-output model (I-O model). Awirya (2009) used I-O model approach as part of money generating model to analyse economic impact of tourism in Bali. Saayman and Naude (2011) also used I-O model to analyse economic impact of tourist spending. To estimate the effect of tourist spending on the increase of output, the effect output is calculated as:

$$Y = (I - A)^{-1} \times D$$

where :

Y : Increase of output
$(I - A)^{-1}$: Leontief inverse
I : identity matrix
A : the input-output coefficients matrix
D : Tourist spending

The last part uses regression analysis to identify determinants of tourist spending [Marcussen (2011); Fredman (2008); Jurdana and Frleta (2016)]. This study uses logistic regression for analysis. The dependent variable in this study is tourist spending per day. For statistical modeling, the variable is a binary variable based on their median values, where :

1 : for tourist spending more than median value, and

0 : for toursist spending less than median value.

The independent variables are tourist characteristics in terms of age, sex, visit purpose, length of stay, visit frequency and number of tourist companions (travel partner). Data are obtained from Passenger Exit Survey of North Sumatra in 2014 and formulated as follows :

$$\text{Probspending} = f(\text{age}, \text{sex}, \text{purp}, \text{days}, \text{visited}, \text{with})$$

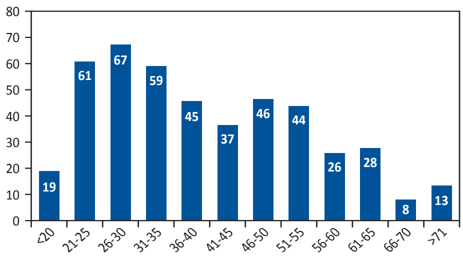
Where :

Age : age of tourist (in years)
Sex : gender of tourist → Male (1) and female (0)
Purp : visit purpose → pleasure (1) and other purpose, such as business (0)
Days : length of stay (in days)
Visited : frequency of visit, including at the time of survey
With : number of companions/partners (persons)

4. RESULTS

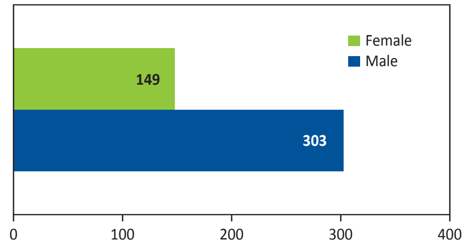
Descriptive Statistics of Tourist Spending

Based on PES results, tourists visit North Sumatra are varied in ages, in which majority (> 90%) are in productive ages (Graph 1). In terms of gender, 67% tourists are male (Graph 2).



Source: PES, 2014

Graph 1. Tourist Age Range

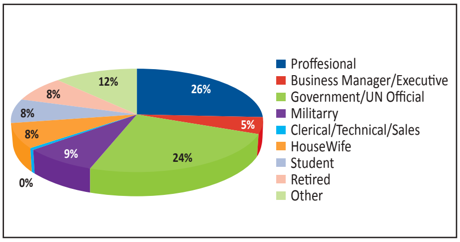


Source: PES, 2014

Graph 2. Tourist Gender

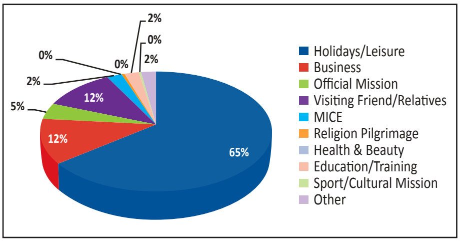
The PES classified tourist occupation as professionals, government/UN official, military, business manager/executive, clerical/technical/sales, housewife, student, retired and others. The survey result shows most tourists works as professional (26%) and work in government sector/UN Official (24%), see Graph 3.

Identifying the purpose of visit to North Sumatra, mostly tourists (65%) visited for holidays/leisure purpose. Meanwhile, 12% tourists visited North Sumatra for business purpose, 12% for visiting friends/ relatives and the remaining for other purposes, such as official mission, MICE (meeting, incentives, convention, exhibition), education and training, sport/cultural mission (see Graph 4).



Source: PES, 2014

Graph 3. Tourist Occupation



Source: PES, 2014

Graph 4. Purpose of Visiting of Tourist

Economic Impact of Tourist Spending

Based on PES data used, the average tourist spending amounts to IDR 1.911.000 per day. Using statistical data published about total tourists and average length of stay, we estimate total tourist spending in North Sumatra in 2014. Total tourists visited North Sumatra in 2014 are 270.837 persons and the average length of their stays are 1,8 days. By multiplying number of tourists, tourist spending and length of stay, total tourist spending in North Sumatra in 2014 amounts to IDR 931,63 billions or equivalent to USD75,13 millions.

The largest part of tourists spending is distributed to Hotels and Restaurants subsector (51,1% of total spending). Other allocations of spending are distributed to Trade subsector (26,7%), Other Services (9,2%), Road Transport (8,2%), Air Transport (2,5%), Sea Transport (1,8%) and Communications (0,5%).

In particular, to estimate the effects of tourist spending on the increase of output, the Leontief inverse matrix is multiplied by tourist spending. The results revealed that Hotels and Restaurants subsector generates the highest increase in output. The next rank is Trade subsector that generates increase in output as much as 69% of the increase in output of Hotels and Restaurant. We use relative scale to accommodate I-O table in 2005. In line with the results, Awirya (2009) also reveals that Hotels and Restaurant and Other services subsector are in the top 5 economic subsectors that generates highest impact of tourism in Bali. Table 4.1 reveals subsectors that generate highest increase in output as the impact of tourist spending.

Table 4.1

Subsector	Increase in Output
Hotels and Restaurants	Highest
Trade	69% of the increase in output of Hotel and Restaurants
Food and Beverages Industry	40% of the increase in output of Hotel and Restaurants
Other services	29% of the increase in output of Hotel and Restaurants
Oil Palm Industry	26% of the increase in output of Hotel and Restaurants

All subsectors that generate the highest impact in output are related directly to tourism, except the oil palm industry. The increase in output of oil palm industry is estimated as the indirect effects resulting from various rounds of re-spending of the related industry's (Hotel and restaurants, trade, food and beverages) receipts in oil palm industry as a backward-linked industry. The result is also in line with the position of oil palm industry as one of the main industry in North Sumatra.

Regression Analysis

The logistic regression reveals that three independent variables significantly affects the probability of high spending; those are the purpose of visit (purp), length of stay (days) and number of tourist partner (with). The significance level of 'days' amounts to 1%, meanwhile 'with" amounts to 10%.

Table 4.2

Dependent Variable: PROBSPEND3
Method: ML -Binary Logit (Quadratic hill climbing)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AGE	0.004722	0.007300	0.646857	0.5177
SEX	0.239979	0.224332	1.069745	0.2847
PURP	0.757742	0.239861	3.159086	0.0016
DAYS	-0.281500	0.050437	-5.581199	0.0000
VISITED	0.002443	0.018500	0.132033	0.8950
WITH	-0.206987	0.109567	-1.889141	0.0589
C	0.767087	0.470432	1.630603	0.1030
McFadden R-squared	0.111991	Mean dependent var		0.500000
S.D. dependent var	0.500554	S.E. of regression		0.463246
Prob (LR statistic)	0.000000			

The purpose of visit has prob value 0,0016, smaller than its significance level of 1% with positive coefficient. This result indicates that tourist who has leisure purpose will spend more than tourist with other purposes of visit.

The length of stay also has prob value 0,0000, smaller than its significance level of 1% with negative coefficient. This result indicates that the longer length of tourist stay, the lower the tourist spending. As the length of stay increase, the total tourist spending also increase, but the average of tourist spending per day will be lower. Jurdana and Frelta (2016) and Imler (2011) found similar result that showed negative effect of length of stay on tourist spending.

The number of tourist partner when visit has prob value 0,0589, smaller than its significance level of 10% with negative coefficient. This result show that the more persons who accompanied tourist, the lower the tourist spending. The average of tourist spending might be lower because certain expenditures can be shared, such as hotels.

On the other hand, three other variables are insignificant in affecting tourist spending. Fredman (2008) and Imler (2011) also revealed that gender is insignificantly affect tourist spending. Meanwhile, Marcussen (2011) concludes that there is no significance difference between first and repeat visit in tourist spending.

5. CONCLUSIONS

The average of tourist spending in North Sumatra in 2014 amounts to IDR 1.911.000 per day or amounts to IDR 931,63 billions for a year or equivalent to USD75,13 millions. Most of tourists visited North Sumatra are in productive ages and 67% are male. The survey results also show that most tourists works as professional and work in government sector/UN Official, while the purposes of visit are mostly for holidays/leisure purpose.

Most of tourist spending is allocated to Hotels and Restaurants and Trade subsector. The calculation of economic impact using Leontief inverse matrix shows that Hotels and Restaurants subsector generates the highest increase in output, followed by Trade that generates increase in output as much as 69% of the increase in output of Hotels and Restaurant.

The regression analysis reveals that tourist who has leisure/holiday purpose might spend more than other purpose of visit. Meanwhile, length of stay and numbers of travel partner are negatively affect higher spending. The other variables, those are age, sex and frequency of visit are insignificantly affect the tourist spending.

This study findings are relevant for both policy makers and business managers in tourism sector. As most tourists visit North Sumatra for leisure and frequency of visit has insignificant effect on tourist spending, we suggest that the promotional strategy will stress the beauty and comfort of tourism destination, instead of improvement of amenities for business. At the end, all efforts for boosting tourist spending will impact the output.

Considering the limitations of the study, it has to be pointed out that the models in this study did not consider all possible variables. Consideration of other relevant variables (type of accommodation, transportation mode, etc) might explain more. For the further research, it is recommended to conduct the study using larger sample and longer time period (peak and low season).

REFERENCES

Awirya, Alam Agni. 2009. Peningkatan Ekonomi Bali Melalui Pengembangan Pariwisata. *Multiplier Bali Tourism*.

Brida, J. G., & Pulina, M. (2010). A literature review on the tourism-led-growth hypothesis. Working paper 17.

Cortes-Jimenez, I. (2008). Which type of tourism matters to the regional economic growth? The cases of Spain and Italy. International Journal of Tourism Research, 10, 127-140.

Downward, P., and Lumsden, L. 2000. The Demand for Day-Visits: An Analysis of Visitor Spending. *Tourism Economics, 9*(1). 67-77.

Farrel, M.J., 1953-54. Some aggregation problems in demands analysis. *Review of Economic Studies, 21*. 193-203.

Fredman, Peter. 2008. Determinants of visitor expenditures in mountain tourism. *Tourism Economics, 14* (2). 297–311.

Figini, P., & Vici, L. (2010). Tourism and growth in a cross-section of countries. Tourism Economics, 16(4), 789-805.

Imler, Alexis. 2011. Evaluating Nature of Expenditures and Economic Impact of Tourism Spending On Nature-Based Activities in South Carolina Coastal Economies. *All Theses*. Paper 1210.

Jurdana, Dora Smolčić., and Daniela Soldić Frleta. 2016. Factors Affecting The Expenditure of Domestic and Foreign Tourists. *Tourism & Hospitality Industry, Congress Proceedings*, pp. 418-431.

Kim, W.G., Kim, T., Kim, S.H., Park, S.S., Gazolli, G., and Yumi Park. 2011. Factors affecting the travel expenditure of visitors to Macau, China. *Tourism Economics, 17* (4). 000–000.

Kozak, M., Gokovali, U., Bahar, O. 2008. Estimating the Determinants of Tourist Spending: A Comparison of Four Models. *Tourism Analysis, 13*(2).143-155.

Lee, C. C., & Chang, C. P. (2008). Tourism development and economic growth: a closer look at panels. Tourism Management, 29(1), 180-192.

Marcussen, Carl H. 2011. Determinants of tourist spending in crosssectional studies and at Danish destinations. *Tourism Economics,17* (4), 833–855.

Paci, R., & Marrocu, E. (2014). Tourism and regional growth in Europe. Papers in Regional Science, 93(Suppl. 1), S25-S50.

Paton, S.G. 1985. Tourism and Local Economic Development: Factory Outlets and the Reading SMSA. Growth and Change, 16(3):64-73, July.

Saayman, M., Saayman, A., Naude`, W. 2011. Economic and Sectoral Effects of Tourist Spending in South Africa: Regional Implications. *Paper prepared for the ABAS International Conference,* 10 – 12 July, Prague, Czech Republic.

World Tourism Organization (WTO). 1999. *International Tourism A Global Perspective*. Madrid, Spain

CPS09: Probability Theory & Statistical Modelling (2)

Sample Size Determination for Attaining Maximum Power under Cost Constraints

Bhargab Chattopadhyay*
Indian Institute of Information Technology Vadodara, Gandhinagar, India – bhargab@iiitvadodara.ac.in

Pradipta Ray
The University of Texas at Dallas, Richardson, USA – prr105020@utdallas.edu

Abstract

RNA abundance profiling assays like Quantitative Polymerase Chain Reaction (qPCR) Ribonucleic acid sequencing (RNA-seq) experiments are high-throughput, massively parallel assays for querying RNA species abundance in individual or pools of cells, and have enormous importance in varied fields like evolution, pathology, and drug development. However, such experiments are typically performed by a priori determination of sample sizes during experimental design, often leading to underpowered or overpowered statistical hypotheses testing because of too few or too much experimental replicates respectively. Recently, algorithms like SCOTTY have been developed to provide a two-stage experimental design framework for RNA-seq experiments. We extend this approach by designing a sequential experimental design framework for RNA abundance quantification experiments for a single gene (like qPCR). We show how resources can be optimally allocated with respect to maximizing power for particular cost constraints for such experiments, developing a closed form solution for single gene hypothesis testing on RNA abundance quantification.

Keywords: Testing for equivalence; Testing for superiority/non-inferiority; Power; Sequential analysis

1. INTRODUCTION

Dramatic reductions in cost of RNA-seq have resulted in a shift from experimental design paradigms with a minimal number of replicates to designs which seek to minimize Type I or Type II errors. Several studies have analyzed sample size considerations for power of statistical hypothesis testing in RNA abundance profiling assays, such as Li *et al* (2013); Hart *et al* (2014); and Wu *et al* (2015). However, these approaches do not factor in budget constraints for experimental design. Increase in sample sizes causes the power of a test to increase (due to reduction in variability) but also incurs larger sampling cost, leading to a trade-off in design of experiments. Ching *et al* (2014); and Liu *et al* (2014) provide frameworks for optimizing power in the design of experiment under the constraint of a capped budget. However, Liu *et al* uses the notion of cost per 1% power, while Ching *et al* use simulated datasets to estimate variation in datasets, highlighting an important notion that dataset variability is difficult to estimate *a priori*. Busby *et al* (2013) proposed an algorithm, SCOTTY, based on a two-stage procedure which can be used to find the optimal sample size in order to attain the maximum power of a test in a case control setup under a fixed cost. Such two-stage procedures are a special case of sequential analysis and have been well studied in Ghosh and Sen (1991); Ghosh et al. (1997); Mukhopadhyay and De Silva (2009) and others. The two-stage procedure is less economical than the group-sequential sequential procedure in terms of optimal size (Ghosh and Sen, 1991) as a two-stage procedure is based on smaller sample size than the purely sequential or group sequential procedures. In this article, we maximize the power of the superiority or inferiority tests related to a single gene belonging to more than two cell types under a cost constraint considering a fixed read depth using a group sequential procedure using an exact, closed form formulation of the power. This test can be used for single gene abundance quantification assays like qPCR, but we consider this as an important first step in formulating procedures for maximizing power with respect to a set of genes picked *a priori* or through sequential analysis, under a budget constraint.

2. TESTS FOR EQUIVALENCE AND NON-INFERIORITY/SUPERIORITY

For the i^{th} group suppose $X_{i1}, \dots, X_{in}$ be independent and identically distributed random variables, not necessary normal, with means μ_i and variances , thus in total there are $n = \sum_{i=1}^{k} n_i$. Let us consider a parameter, which is a function of K population means, $\partial = \sum_{i=1}^{k} c_i \mu_i$, (c1,….. ck) are known coefficients.

An estimator of the parametric function $\partial_{is} \hat{\partial}_n = \sum_{i=1}^{k} c_i \bar{X}_i$, $\bar{X}_i$ being the sample mean of i^{th} group mean. If γ_i's are sample size allocation ratios then $n_i = \gamma_i n_i$ for I = 1,2,…..,K with γ_i = 1

The variance of $\hat{\partial}_n$ is

$$Var(\hat{\partial}_n) = \sum_{i=1}^{k} \frac{c_i^2 \sigma_i^2}{n_i} = \frac{1}{n_1} \sum_{i=1}^{k} \frac{c_i^2 \sigma_i^2}{\gamma_i} \tag{1}$$

Thus for not so small n1 , using central limit theorem, we have

$$\frac{\sqrt{n_1}(\hat{\partial}_n - \partial)}{\sqrt{\sum_{i=1}^{k} c_i^2 \sigma_i^2 / \gamma_i}} \xrightarrow{d} N(0,1) \tag{2}$$

Using the central limit theorem, we can test claims regarding the parametric function ∂ without the assumption of normality of the data.

2.1. Testing for Equivalence

Suppose, we wish to test hypotheses for equivalence by comparing ∂, to two equivalence margins with ∂1>∂>∂2 (Lehmann and Romano, 2006; Guo et al., 2011; Luh and Guo, 2015). Thus the null and alternative hypotheses are:

$$H_0 : \partial \le \partial_1 \text{ or } \partial \ge \partial_2 \qquad \text{against} \qquad H_1 : \partial_1 < \partial < \partial_2 \tag{3}$$

In many studies, such as educational testing and drug testing, often we show equivalence of treatments. Then the null hypothesis H_0 defined in Equation (3), is rejected at α % level of significance if,

$$\partial_1 + z_\alpha \sqrt{Var(\hat{\partial}_n)} < \hat{\partial}_n < \partial_2 + z_\alpha \sqrt{Var(\hat{\partial}_n)} \tag{4}$$

Where, Z_α is the upper α point of standard normal distribution.

The approximate power function of the corresponding test of equivalence is given by,

$$P_n(\partial) = \Phi\left(-z_\alpha + (\partial_2 - \partial)/\sqrt{Var(\hat{\partial}_n)}\right) - \Phi\left(-z_\alpha - (\partial - \partial_1)/\sqrt{Var(\hat{\partial}_n)}\right) \tag{5}$$

2.2. Testing for superiority/inferiority

Suppose, we wish to test hypotheses for superiority (or inferiority) by comparing ∂, to ∂_0 (Lehmann and Romano, 2006, Guo et al., 2011). provided that the larger value of the response variable is better, the null and alternative hypotheses are:

$$\begin{array}{ll} \text{(Inferiority)} & H_0 : \partial \le -|\partial_0| \text{ against } H_1 : \partial > -|\partial_0| \\ \text{(Superiority)} & H_0 : \partial \le |\partial_0| \text{ against } H_1 : \partial > |\partial_0| \end{array} \tag{6}$$

Since the superiority test can be extended from inferiority test, so in this article, we consider the inferiority test. The null hypothesis H_0 of the inferiority test defined in Equation (6), is rejected at α % level of significance if,

$$\hat{\partial}_n > |\partial_0| + z_\alpha \sqrt{Var(\hat{\partial}_n)} \tag{7}$$

Where, Z_α is the upper α point of standard normal distribution.

The approximate power function of the corresponding test of equivalence is given by,

$$P_n(\partial) = \Phi\left(-z_\alpha + (|\partial_0| + \partial)/\sqrt{Var(\hat{\partial}_n)}\right) \tag{8}$$

3. SAMPLE SIZE ALLOCATION

In experimental research, limited funding is allotted beforehand to carry out the sampling process. Thus under a cost constraint we have to carry out the test of equivalence or inferiority (or superiority) with minimum errors. Note that we fix the probability of type I error to α , so the idea is to minimize the probability of type II error or equivalently maximizing the power under cost constraints. Recall from Equations (5) and (8) in test of equivalence and test of inferiority (or superiority), the power gets maximized if $Var(\hat{\partial}_n)$ gets minimized.

Let us consider that a fixed amount A_0 is allotted for the sampling process and also consider that a_1 a_2 …., a_k denotes the cost of sampling one observations in groups 1,2,….. K respectively. Thus,

$$A_0 = \sum_{i=1}^{K} a_i n_i = n_1 \sum_{i=1}^{K} a_i \gamma_i$$

Thus, the allotted sample size for group 1 under the cost constraint is

$$n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i} \tag{9}$$

Using the expression in Equation (9), the variance of ∂_n is

$$Var(\hat{\partial}_n) = \frac{1}{A_0}\left(\sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{\gamma_i}\right)\left(\sum_{i=1}^{K} a_i \gamma_i\right) \tag{10}$$

which will be minimized under the optimal allocation ratio

$$\gamma_i = \frac{|c_i|\sigma_i}{|c_1|\sigma_1\sqrt{a_i/a_1}} \tag{11}$$

Thus with the the optimal allocation ratio in Equation (11), the optimal sample sizes for the K groups are,

$$n_{0i} = \frac{A_0 |c_i| \sigma_i}{\sum_{i=1}^{K} |c_i| \sigma_i \sqrt{a_i a_i}}, \quad \text{for } i = 1,2,\ldots,K \tag{12}$$

and the minimized variance of ∂_n under the optimal allocation ratio is

$$V_0 = \frac{1}{A_0}\left(\sum_{i=1}^{K} \sqrt{a_i}\,|c_i|\sigma_i\right)^2 \tag{13}$$

Thus, maximum power is attained, under the cost constraint A_0 if the minimum variance defined in Equation (13) is attained or in other words if the optimal sample allocation ratio is used. However the optimal allocation ratio as defined in Equation (11) depends on population variances of the *K* groups. In practice, the values of population variances of each group are unknown, hence we cannot compute the allocation ratio and thereby the sample size required to obtain maximum power of the test. In the next section, we develop a new approach to avoid using a potentially poor estimate of sample size, often obtained by using supposed population standard deviations, which are potentially poor estimates to plan sample size. Our method ensures that we are informed by actual data from the population of interest.

4. SAMPLE SIZE DETERMINATION

As opposed to use of supposed population values, the new approach that we momentarily develop, do not fix the sample sizes in advance, rather we estimate in stages. In statistics, this procedure is called sequential procedure. In sequential procedure, in the first stage, a small sample called a pilot sample is observed and then the parameter(s) are estimated to check a pre-defined condition in a pre-specified rule, known as the stopping rule. Further sampling of observations is carried out if the pre-defined condition is not met and further sampling is stopped once the pre-defined condition is satisfied. At a particular stage, if the pre-defined condition is not met, the researcher collects one or more additional observation(s) and then estimates the parameter of interest. This process is repeated until the pre-defined condition is met. For details about the general theory of sequential estimation procedures, readers are referred to Sen (1981), Ghosh and Sen (1991) and others.

As discussed, the optimal sample sizes for the K groups, n_{io} for $i = 1,2,\ldots, K$ is unknown because it depends on corresponding population standard deviations σ_i which is itself unknown in practice. Thus, in order to estimate n_{io}, an estimator of σ_i is required, henceforth we will use s_{iN} as an estimator of the population standard deviation based on N_{io} observations drawn from each of the K groups, where, N_{io} is the estimated optimal final sample size for the i^{th} group. We now develop an algorithm to find an estimate of the optimal sample size via the sequential estimation procedure.

Stage 1: First m (>2) randomly selected observations are collected from each of the K groups, often called pilot sample size and set $N_{io\,=\,m}$ for each of the K groups. From these m observations from each K groups, we estimate the corresponding population standard deviation. Let us define, S_{il}^2 as the sample standard deviation computed from m observations collected from i^{th} group in the first stage. If for i^{th} group (i= 1,2.,K), $m \ \geq \ A_0|C_1|S_{iNio}$ / $(\sum_{i=1}^{k} c_1|\, s_{lNl}\, \sqrt{a_1} a_1)$ stop sampling from the i^{th} group and set the sample size for the i^{th} group equal to $N_{io\,=\,m}$. If for the i^{th} group (i= 1,2.,K), $m < A_0|C_1|S_{iNio}$ / $(\sum_{i=1}^{k} c_1|\, s_{lNl}\, \sqrt{a_1} a_1)$ then proceed to the next step.

Stage 2 : Obtain m' additional observations from the remaining groups. Thus there are $N_{io} = m + m'$ observations from each of the remaining groups (i.e., the pilot sample size and an additional m' observations per group, for a total sample size of $\sum_{i=1}^{k} N_{io}$. If $m + m' \geq A_0|C_i|S_{iNio}$ / $(\sum_{i=1}^{k} c_1|\, s_{lNl}\, \sqrt{a_1} a_1)$, stop further sampling and set the final sample size for the i^{th} group equal to $N_{io} = m{+}m$ and if $m + m' < A_0|C_i|S_{iNio}$ / $(\sum_{i=1}^{k} c_1|\, s_{lNl}\, \sqrt{a_1} a_1)$ then continue the sampling process for that group by sampling m' more observations from that group. This data collecting process after the pilot sample stage continues until the sample size condition is satisfied, that is, $N_{io} \geq A_0|C_i|S_{iNio}$ / $(\sum_{i=1}^{k} c_1|\, s_{lNl}\, \sqrt{a_1} a_1)$ At this stage, we stop further sampling and report that the final sample size $N_o = \sum_{i=1}^{k} N_{io}$.

At each stage of the algorithm, we check whether the sample size collected up to that stage is at least as large as the estimated value of N_{io} using observations from the i^{th} group collected until that stage. Based on the algorithm just outlined, a sampling stopping rule can be defined as follows:

*N*io is the smallest integer n_i *(≥m)* such that $n_i \geq A_0|C_i|S_{ini}$ / $(\sum_{i=1}^{k} c_l|\, s_{ini}\, \sqrt{a_1} a_1)$

The total cost of sampling $\sum_{i=1}^{k} N_{io}$ observations, N_{io} being the estimated final optimal sample size for the i^{th} group computed using Equation (14) is A_0. This is proved in the following lemma.

Lemma : The total cost sampling $\sum_{i=1}^{K}N_{io}$ observations is A_0.

Proof : We note that N_{io} is the estimated final optimal sample size for the i^{th} group. We know from Equation (14), that

$$\frac{A_0|C_i|s_{iN_{io}}}{\left(\sum_{i=1}^{k}|C_i|s_{iN_{io}}\sqrt{a_i\alpha_i}\right)} \le N_{io} \le \frac{A_0|C_i|s_{iN_{io}-1}}{\left(\sum_{i=1}^{k}|C_i|s_{iN_{io}-1}\sqrt{a_i\alpha_i}\right)} \qquad (14)$$

Multiplying by a_i on all three sides in Equation (15) and summing them over we get,

$$\frac{A_0\sum_{i=1}^{k}|C_i|a_i s_{iN_{io}}}{\left(\sum_{i=1}^{k}|C_i|s_{iN_{io}}\sqrt{a_i\alpha_i}\right)} \le \sum_{i=1}^{k} a_i N_{io} \le \frac{A_0\sum_{i=1}^{k}|C_i|a_i s_{iN_{io}-1}}{\left(\sum_{i=1}^{k}|C_i|s_{iN_{io}-1}\sqrt{a_i\alpha_i}\right)}$$

Thus, we can say using Sandwich theorem that $\sum_{i=1}^{k} a_i N_{io} = A_0$, the total cost of sampling $\sum_{i=1}^{k} N_{io}$ observations.

5. CONCLUSION

Our approach provides a framework for formulating a sequential procedure for transcript abundance quantification assays. We show that our method can optimally allocate budgetary resources in a sequential sampling scenario within the context of maximizing power for statistical hypothesis testing. We aim to expand from a single gene framework to perform similar analysis for a set of genes. We are in the process of building an open source tool to perform this, bringing a model-based, computationally tractable, sequential analysis framework to high-throughput transcriptome assays. The development of such a universal, plug and play framework will be enormously beneficial to the transcriptomics community, where experimental design is still largely performed based on *ad hoc* considerations, overlooking statistical hypothesis testing needs.

REFERENCES

Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R., and Marth, G. T. (2013). Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics, 29, 656–657.

Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. Rna, 20-11, 1684-1696.

Ghosh, B. K. and Sen, P. K. (1991), Handbook of sequential analysis, vol. 118, CRC Press. Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). Sequential estimation, Wiley (New York).

Guo, J.-H., Chen, H. J., and Luh,W.-M. (2011). Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. British Journal of Mathematical and Statistical Psychology, 64, 439–461.

Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A., & Kocher, J. P. (2013). Calculating sample size estimates for RNA sequencing data. Journal of computational biology, 20-12, 970-978.

Lehmann, E. L. and Romano, J. P. (2006). Testing statistical hypotheses, pp 81-81, Springer Science & Business Media.

Li, C. I., Su, P. F., & Shyr, Y. (2013). Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. BMC Bioinformatics, 14-1, 357.

Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? Bioinformatics, 30(3), 301-304.

Luh, W.-M. and Guo, J.-H. (2015). Sample size planning for the noninferiority or equivalence of a linear contrast with cost considerations. Psychological Methods, 21, 13–34.

Mukhopadhyay, N. and De Silva, B. M. (2009), Sequential methods and their applications, CRC Press. Sen, P. K. (1981), Sequential nonparametrics: Invariance principles and statistical inference, Wiley NewYork.

Wu, H., Wang, C., & Wu, Z. (2015). PROPER: comprehensive power evaluation for differential expression using RNA-seq. Bioinformatics, 31-2, 233-24

Tracing and Applying the Formula in the Equivalent Methods of Two Way ANOVA in Nonparametric Statistics

Fitri Catur Lestari*
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – fitricaturlestari@stis.ac.id

Abstract

There were three tests in nonparametric statistics that were equivalent with two-way ANOVA in parametric statistics namely Bredenkamp, Hildebrand and Kubinger. Mentioning Friedman method as two-way ANOVA equivalence was incorrect because there was not interaction term between the row and column factor as in two-way ANOVA. This matter would be clear if tracing the original formula of the methods. The purpose of this research were to trace the original formula of Bredenkamp, Hildebrand, Kubinger and test statistics and to apply the equivalent ones in the pharmacy field. The literature study and the secondary data were used in this research. The secondary data involved the effect of brand factor and preparation method factor in Chinese herb lead content. The research results were (1) the relevant tests with two-way ANOVA on nonparametric statistics were Bredenkamp, Hildebrand and Kubinger to know the difference of row factor, column factor, and an interaction between the row and column factor. The formulas in Bredenkamp, Hildebrand and Kubinger tests are derived from two-way ANOVA formulas, especially factorial experimental design. Whereas Friedman test is derived from the formula of Randomized Block Design (2) By Bredenkamp method, the interaction effect and preparation method was not significant but brand factor was significant and by Hildebrand and Kubinger method, the interaction effect, brand factor and preparation method was not significant on alpha five percent.

Keywords: Statistic Nonparametric; Bredenkamp; Hildebrand; Kubinger; Two-Way ANOVA.

1. INTRODUCTION

Several assumption have to be fullfilled before use parametric statistics method. Contrarily nonparametric statistics method doesn't require them. The other advantages of nonparametric statistics method are can be used for small sample, lower assumption were needed, can be used in categorical data (Siegel,1988). The conditions that nonparametric statistics method can be used are: the unknown pattern of population distribution, nominal or ordinal data, small sample size and the characteristic of the population distribution is unknown surely (Djarwanto, 2003).

In nonparametric statistics method there are a lot of tests that are similar with tests in parametric statistics method. There were three tests in nonparametric statistics that were equivalent with two-way ANOVA in parametric statistics namely Bredenkamp, Hildebrand and Kubinger. Mentioning Friedman method as two-way ANOVA equivalence was incorrect because there was not interaction term between the row and column factor as in two-way ANOVA. This matter would be clear if tracing the original formula of the methods. The purpose of this research were to trace the original formula of Bredenkamp, Hildebrand, Kubinger and test statistics and to apply the equivalent ones in the pharmacy field. The literature study and the secondary data were used in this research. The secondary data involved the effect of brand factor and preparation method factor in Chinese herb lead content.

2. LITERATURE STUDY

For tracing the formula of Bredenkamp, Hildebrand and Kubinger test, it is important to do some literature studies about two way anova. Two way anova is a test to know row effect, column effect and interaction effect to response variable.

A. Model Two Way Anova

Model used in two way anova (Neter et al., 1990) is $Y_{ijk} = \mu_{...} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ (2.1); Y_{ijk}= k-th observation for i-th factor A and j-th factor B; $\mu_{..}$ = constant; α_i = effect of factor A; β_j = effect of factor B; $(\alpha\beta)_{ij}$ = effect of interaction between factor A and B; ε_{ijk} = *error;* i=1,2,...,a; j=1,2,...,b; k=1,2,...,n; N=abn; assumptions: $\sum_{i=1}^{a}\alpha_i = 0$; $\sum_{j=1}^{b}\beta_j = 0$; $\sum_{i=1}^{a}(\alpha\beta)_{ij} = 0$; $\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$; ε_{ijk} ~NID$(0,\sigma^2)$ (2.2)

B. Parameter Estimation, Sum Square (SS) and Computation Formula

Based on model (2.1) the equation is: $\varepsilon_{ijk} = Y_{ijk} - \mu_{...} - \alpha_i - \beta_j - (\alpha\beta)_{ij}$; If sum of error was squared, then the equation will be: $Q = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{ijk}^{\ 2} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \mu_{...} - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$ (2.3)

By Least Square Method (LSM), the estimators of the parameters are listed below:

$\hat{\mu}_{...} = \overline{Y}_{...}$; $\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{...}$; $\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{...}$; and $(\hat{\alpha\beta})_{ij} = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$. Sum Square of A factor effect (SSA)= $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{i..} - \overline{Y}_{...})^2 = \frac{\sum_{i=1}^{a}Y_{i..}^{\ 2}}{bn} - \frac{Y_{...}^{\ 2}}{abn}$; B factor effect (SSB)= $\sum_{i=1}^{a}\sum_{b=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{.j.} - \overline{Y}_{...})^2 = \frac{\sum_{j=1}^{b}Y_{.j.}^{\ 2}}{an} - \frac{Y_{...}^{\ 2}}{abn}$; A and B interaction effect (SSAB)= $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} - \overline{Y}_{...})^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}Y_{ij.}^{\ 2}}{n} - \frac{\sum_{i=1}^{a}Y_{i..}^{\ 2}}{bn} - \frac{\sum_{j=1}^{b}Y_{.j.}^{\ 2}}{an} + \frac{Y_{...}^{\ 2}}{abn}$.

3. TRACING THE TEST STATISTIC OF BREDENKAMP, HILDEBRAND AND KUBINGER TEST

A. Bredenkamp Test

All value of Y_{ijk} is transformed to single rank R_{ijk}. If there are the same value of Y_{ijk} , then the rank is average rank from Y_{ijk} that have equal value.

Test statistic for row effect is $\frac{12a}{N^2(N+1)}\sum_{i=1}^{a}R_{i..}^{\ 2} - 3(N+1)$, χ^2_{a-1} distributed (Hühn dan Léon, 1995). It derived from

$$\begin{aligned}
\text{SSA} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R}_{i..} - \overline{R}_{...})^2 = bn\sum_{i=1}^{a}(\overline{R}_{i..} - \overline{R}_{...})^2 = bn\left(\sum_{i=1}^{a}\overline{R}_{i..}^{\ 2} - 2\sum_{i=1}^{a}\overline{R}_{i..}\overline{R}_{...} + \sum_{i=1}^{a}\overline{R}_{...}^{\ 2}\right) \\
&= bn\left(\sum_{i=1}^{a}\left(\frac{R_{i..}}{bn}\right)^2 - 2\overline{R}_{....}\sum_{i=1}^{a}\overline{R}_{i..} + a\overline{R}_{...}^{\ 2}\right) = bn\left(\sum_{i=1}^{a}\frac{R_{i..}^{\ 2}}{(bn)^2} - 2\left(\frac{R_{...}}{abn}\right)\left(\frac{\sum_{i=1}^{a}R_{i..}}{bn}\right) + a\left(\frac{R_{...}}{abn}\right)^2\right) \\
&= bn\left(\frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{(bn)^2} - 2\frac{R_{...}}{abn}\frac{R_{...}}{bn} + \frac{R_{...}^{\ 2}}{ab^2n^2}\right) = bn\left(\frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{(bn)^2} - \frac{R_{...}^{\ 2}}{ab^2n^2}\right) = \frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{bn} - \frac{R_{...}^{\ 2}}{abn};
\end{aligned}$$

$$\frac{SSA}{S^2} = \frac{\frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{bn} - \frac{R_{...}^{\ 2}}{abn}}{S^2} = \frac{\frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{bn} - \frac{R_{...}^{\ 2}}{abn}}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)}\left(\frac{\sum_{i=1}^{a}R_{i..}^{\ 2}}{\frac{N}{a}} - \frac{R_{...}^{\ 2}}{N}\right)$$

$$= \frac{12}{N(N+1)}\frac{a}{N}\sum_{i=1}^{a}R_{i..}^{\;2} - \frac{12}{N(N+1)}\frac{R_{...}^{\;2}}{N} = \frac{12a}{N^2(N+1)}\sum_{i=1}^{a}R_{i..}^{\;2} - \frac{12}{N(N+1)}\frac{\left(\frac{N(N+1)}{2}\right)^2}{N}$$

$$= \frac{12a}{N^2(N+1)}\sum_{i=1}^{a}R_{i..}^{\;2} - 3(N+1)$$

Test statistic for column effect is $\frac{12b}{N^2(N+1)}\sum_{j=1}^{b}R_{.j.}^{\;2} - 3(N+1)$, distributed (Hühn dan Léon, 1995).

It derived from

$$\text{SSB} = \sum_{i=1}^{a}\sum_{b=1}^{b}\sum_{k=1}^{n}(\overline{R}_{.j.} - \overline{R}_{...})^2 = an\sum_{j=1}^{b}(\overline{R}_{.j.} - \overline{R}_{...})^2 = an\left(\sum_{j=1}^{b}\overline{R}_{.j.}^{\;2} - 2\sum_{j=1}^{b}\overline{R}_{.j.}\overline{R}_{...} + \sum_{j=1}^{b}\overline{R}_{...}^{\;2}\right)$$

$$= an\left(\sum_{j=1}^{b}\left(\frac{R_{.j.}}{an}\right)^2 - 2\overline{R}_{...}\sum_{j=1}^{b}\overline{R}_{.j.} + b\overline{R}_{...}^{\;2}\right) = an\left(\sum_{j=1}^{b}\frac{R_{.j.}^{\;2}}{(an)^2} - 2\left(\frac{R_{...}}{abn}\right)\left(\frac{\sum_{j=1}^{b}R_{.j.}}{an}\right) + b\left(\frac{R_{...}}{abn}\right)^2\right)$$

$$= an\left(\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{(an)^2} - 2\frac{R_{...}}{abn}\frac{R_{...}}{an} + \frac{R_{...}^{\;2}}{a^2bn^2}\right) = an\left(\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{(an)^2} - \frac{R_{...}^{\;2}}{a^2bn^2}\right) = \frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} - \frac{R_{...}^{\;2}}{abn}$$

$$\frac{SSB}{S^2} = \frac{\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} - \frac{R_{...}^{\;2}}{abn}}{S^2} = \frac{\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} - \frac{R_{...}^{\;2}}{abn}}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)}\left(\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{\frac{N}{b}} - \frac{R_{...}^{\;2}}{N}\right)$$

$$= \frac{12}{N(N+1)}\frac{b}{N}\sum_{j=1}^{b}R_{.j.}^{\;2} - \frac{12}{N(N+1)}\frac{R_{...}^{\;2}}{N} = \frac{12b}{N^2(N+1)}\sum_{j=1}^{b}R_{.j.}^{\;2} - \frac{12}{N(N+1)}\frac{\left(\frac{N(N+1)}{2}\right)^2}{N}$$

$$= \frac{12b}{N^2(N+1)}\sum_{j=1}^{b}R_{.j.}^{\;2} - 3(N+1)$$

Test statistic for interaction effect is $\frac{12\,ab}{N^2(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(R_{ij.}^2 - \frac{1}{b^2}R_{i..}^2 - \frac{1}{a^2}R_{.j.}^2) + 3\,(N+1)$, X^2(a-1)(b-1) distributed (Hühn dan Léon, 1995). It derived from

$$\text{SSAB} = \left(\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}R_{ij.}^{\;2}}{n} - \frac{R_{...}^{\;2}}{abn}\right) - \left(\frac{\sum_{i=1}^{a}R_{i..}^{\;2}}{bn} - \frac{R_{...}^{\;2}}{abn}\right) - \left(\frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} - \frac{R_{...}^{\;2}}{abn}\right)$$

$$= \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}R_{ij.}^{\;2}}{n} - \frac{\sum_{i=1}^{a}R_{i..}^{\;2}}{bn} - \frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} + \frac{R_{...}^{\;2}}{abn}$$

$$\frac{SSAB}{S^2} = \frac{\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}R_{ij.}^{\;2}}{n} - \frac{\sum_{i=1}^{a}R_{i..}^{\;2}}{bn} - \frac{\sum_{j=1}^{b}R_{.j.}^{\;2}}{an} + \frac{R_{...}^{\;2}}{abn}}{\frac{N(N+1)}{12}}$$

$$= \frac{12}{N(N+1)}\left(\frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2}{n} - \frac{b\sum_{i=1}^{a} R_{i..}{}^2}{b^2 n} - \frac{a\sum_{j=1}^{b} R_{.j.}{}^2}{a^2 n} + \frac{R_{...}{}^2}{abn} \right)$$

$$= \frac{12}{N(N+1)}\left(\frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2}{n} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{i..}{}^2}{b^2 n} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{.j.}{}^2}{a^2 n} + \frac{R_{...}{}^2}{abn} \right)$$

$$= \frac{12}{N(N+1)}\left(\frac{1}{n}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2 - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{i..}{}^2}{b^2} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{.j.}{}^2}{a^2} \right) + \frac{R_{...}{}^2}{N} \right)$$

$$= \frac{12}{N(N+1)}\left(\frac{1}{\frac{N}{ab}}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2 - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{i..}{}^2}{b^2} - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} R_{.j.}{}^2}{a^2} \right) + \frac{R_{...}{}^2}{N} \right)$$

$$= \frac{12ab}{N^2(N+1)}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2 - \frac{1}{b^2}\sum_{i=1}^{a}\sum_{j=1}^{b} R_{i..}{}^2 - \frac{1}{a^2}\sum_{i=1}^{a}\sum_{j=1}^{b} R_{.j.}{}^2 \right) + \frac{12}{N(N+1)}\frac{\left(\frac{N(N+1)}{2}\right)^2}{N}$$

$$= \frac{12ab}{N^2(N+1)}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} R_{ij.}{}^2 - \frac{1}{b^2}\sum_{i=1}^{a}\sum_{j=1}^{b} R_{i..}{}^2 - \frac{1}{a^2}\sum_{i=1}^{a}\sum_{j=1}^{b} R_{.j.}{}^2 \right) + 3(N+1)$$

$$= \frac{12\,ab}{N^2(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(R_{ij.}^2 - \frac{1}{b^2}R_{i..}^2 - \frac{1}{a^2}R_{.j.}^2) + 3\,(N+1)$$

B. Hildebrand Test

All value of Y_{ijk} is transformed $Y_{ijk} \to Y_{ijk}^* = Y_{ijk} - \overline{Y}_{ij.} + \overline{Y}_{i..}$. Y_{ijk}^* is transformed to single rank. $Y_{ijk}^* \to R_{ijk}$ When forming the rank in those transformations, if there are the same value of Y_{ijk}^*, then the rank is average rank from Y_{ijk}^* that have equal value.

Test statistic for row effect is $\frac{12}{a(N+1)}\sum_{i=1}^{a}(\overline{R}_{i..} - \overline{R}_{...})^2$, χ^2_{a-1} distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSA}{S^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R}_{i..} - \overline{R}_{...})^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} bn\sum_{i=1}^{a}(\overline{R}_{i..} - \overline{R}_{...})^2 = \frac{12}{a(N+1)}\sum_{i=1}^{a}(\overline{R}_{i..} - \overline{R}_{...})^2$$

All value of Y_{ijk} is transformed $Y_{ijk} \to Y_{ijk}^* = Y_{ijk} - \overline{Y}_{ij.} + \overline{Y}_{.j.}$ Y_{ijk}^* is transformed to single rank $Y_{ijk}^* \to R_{ijk}$. When forming the rank in those transformations, if there are the same value of Y_{ijk}^*, then the rank is average rank from Y_{ijk}^* that have equal value.

Test statistic for column effect is $\frac{12}{b(N+1)}\sum_{j=1}^{b}(\overline{R}_{.j.} - \overline{R}_{...})^2$, χ^2_{b-1} distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSB}{S^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R}_{.j.} - \overline{R}_{...})^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} an \sum_{j=1}^{b}(\overline{R}_{.j.} - \overline{R}_{...})^2 = \frac{12}{b(N+1)}\sum_{j=1}^{b}(\overline{R}_{.j.} - \overline{R}_{...})^2$$

Y_{ijk} is transformed $Y_{ijk} \rightarrow Y^*_{ijk} = Y_{ijk} - \overline{Y}_{i..} - \overline{Y}_{.j.} + 2\overline{Y}_{....}$ Y^*_{ijk} is transformed to single rank $Y^*_{ijk} \rightarrow R_{ijk}$. When forming the rank in those transformations, if there are the same value of Y^*_{ijk}, then the rank is average rank from Y^*_{ijk} that have equal value.

Test statistic for interaction effect is $\frac{12}{ab(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R}_{ij.} - \overline{R}_{i..} - \overline{R}_{.j.} - \overline{R}_{...})^2$ X^2(a-1)(b-1), distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSAB}{S^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R}_{ij.} - \overline{R}_{i..} - \overline{R}_{.j.} - \overline{R}_{...})^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} n \sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R}_{ij.} - \overline{R}_{i..} - \overline{R}_{.j.} - \overline{R}_{...})^2$$

$$= \frac{12}{ab(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R}_{ij.} - \overline{R}_{i..} - \overline{R}_{.j.} - \overline{R}_{...})^2$$

C. Kubinger Test

All value of Y_{ijk} is transformed to single rank $Y_{ijk} \rightarrow R_{ijk}$. Then it is transformed to $R_{ijk} \rightarrow R^t_{ijk} = R_{ijk} - \overline{R}_{ij.} + \overline{R}_{i..}$ R^t_{ijk} is ordered become R^*_{ijk}. When forming the rank in those transformations, if there are the same value of Y_{ijk}, then the rank is average rank from Y_{ijk} that have equal value. So, is rank forming from R^t_{ijk} to R^*_{ijk}.

Test statistic for row effect is $\frac{12}{a(N+1)}\sum_{i=1}^{a}(\overline{R^*}_{i..} - \overline{R^*}_{...})^2$, χ^2_{a-1} distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSA}{S^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R^*}_{i..} - \overline{R^*}_{...})^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} bn \sum_{i=1}^{a}(\overline{R^*}_{i..} - \overline{R^*}_{...})^2 = \frac{12}{a(N+1)}\sum_{i=1}^{a}(\overline{R^*}_{i..} - \overline{R^*}_{...})^2$$

All value of Y_{ijk} is transformed to single rank $Y_{ijk} \rightarrow R_{ijk}$. Then, it is transformed to $R_{ijk} \rightarrow R^t_{ijk} = R_{ijk} - \overline{R}_{ij.} + \overline{R}_{.j}$. R^t_{ijk} is ordered become R^*_{ijk}. When forming the rank in those transformations, if there are the same value of Y_{ijk}, then the rank is average rank from Y_{ijk} that have equal value. So, is rank forming from R^t_{ijk} to R^*_{ijk}.

Test statistic for column effect is $\frac{12}{b(N+1)}\sum_{j=1}^{b}(\overline{R^*}_{.j.} - \overline{R^*}_{...})^2$, χ^2_{b-1} distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSB}{S^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R^*}_{.j.} - \overline{R^*}_{...})^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} an \sum_{j=1}^{b}(\overline{R^*}_{.j.} - \overline{R^*}_{...})^2 = \frac{12}{b(N+1)}\sum_{j=1}^{b}(\overline{R^*}_{.j.} - \overline{R^*}_{...})^2$$

All value of Y_{ijk} is transformed to single rank $Y_{ijk} \rightarrow R_{ijk}$. Then, it is transformed to $R_{ijk} \rightarrow R^t_{ijk} = R_{ijk} - \overline{R}_{i..} + \overline{R}_{.j}$. R^t_{ijk} is ordered become R^*_{ijk}. When forming the rank in those transformations, if there are the same value of Y_{ijk}, then the rank is average rank from Y_{ijk} that have equal value. So, is rank forming from R^t_{ijk} to R^*_{ijk}.

Test statistic for interaction effect is $\frac{12}{ab(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R^*}_{ij.}-\overline{R^*}_{i..}-\overline{R^*}_{.j.}-\overline{R^*}_{...})^2$, $\chi^2_{(a-1)(b-1)}$ distributed (Hühn dan Léon, 1995). It derived from

$$\frac{SSAB}{S^2}=\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{R^*}_{ij.}-\overline{R^*}_{i..}-\overline{R^*}_{.j.}-\overline{R^*}_{...})^2}{\frac{N(N+1)}{12}}=\frac{12}{N(N+1)}n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R^*}_{ij.}-\overline{R^*}_{i..}-\overline{R^*}_{.j.}-\overline{R^*}_{...})^2$$

$$=\frac{12}{ab(N+1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{R^*}_{ij.}-\overline{R^*}_{i..}-\overline{R^*}_{.j.}-\overline{R^*}_{...})^2$$

D. The similiarity and Difference of Bredenkamp, Hildebrand and Kubinger test

The similiarity of Bredenkamp, Hildebrand and Kubinger test is on test statistics formula tracing. They were derived from formula in two way anova to examine row, column and interaction effect. While the difference of those tests is on the transformation and rank forming. Furthermore the difference was stated by Hühn dan Léon (1995) that based on repeated researches and various cases, it can be concluded that Hildebrand and Kubinger test have relatively same accuracy to detect row, column and interaction effect. While Bredenkamp test has lower accuracy than Hildebrand and Kubinger test to detect row, column and interaction effect.

4. APPLYING THE TEST STATISTIC OF BREDENKAMP, HILDEBRAND AND KUBINGER TEST

The secondary data involved the effect of brand factor and preparation method factor in Chinese herb lead content (Ratnapani, 2004). The aim of her research is to know the interaction effect between preparation method factor and brand factor to lead level of Chinese medicinal herbs and also the main effect of both factors. The sample size of the research is small, only 2 (two) for each treatment combination because of the high cost. The Chinese medicinal herbs are more expensive than other herbs and lead level measurement by Spektrofotometri Serapan Atom (SSA) also need high cost. The information about population distribution is not surely known although normality test had been done. The result of normality test could be doubtful because of the small sample size (Conover, 1980). Therefore the Bredenkamp, Hildebrand and Kubinger test was apropriate to be applied to this case and the analysis result based on Matlab output listed briefly in Table 4.1 below (**= significant on 5%*):

Table 4.1. The Analysis Result of Bredenkamp, Hildebrand and Kubinger Test

Value	Bredenkamp	Hildebrand	Kubinger
Test statistics on row effect (preparation) Chi Square table	0.641026 3.84146	0.025641 3.84146	0.025641 3.84146
Test statistics on c olumn effect (brand) Chi Square table	7.73077* 5.99146	2.000000 5.99146	1.88462 5.99146
Test statistics on interaction effect Chi Square table	1.16667 5.99146	0.666667 5.99146	0.782051 5.99146

5. CONCLUSIONS

The research results were (1) the relevant tests with two-way ANOVA on nonparametric statistics were Bredenkamp, Hildebrand and Kubinger to know the difference of row factor, column factor, and an interaction between the row and column factor. The formulas in Bredenkamp, Hildebrand and Kubinger tests are derived from two-way ANOVA formulas, especially factorial experimental design. Whereas Friedman test is derived from the formula of Randomized Block Design (2) By Bredenkamp method, the interaction effect and preparation method was not significant but brand factor was significant and by Hildebrand and Kubinger method, the interaction effect, brand factor and preparation method was not significant on alpha five percent.

REFERENCES

Bain, L.J., Engelhardt, M. 1991. *Introduction to Probability and Mathematical Statistics*, 2nd Edition, Duxbury Press, California.

Conover, W.J. 1980. *Practical Nonparametric Statistics*, 2nd Edition, John Wiley&Sons, Inc., Canada.

Djarwanto, Ps. 2003. *Statistik Nonparametrik Edisi 2003/2004*, BPFE, Yogyakarta.

Hühn, M., Léon, J. 1995. *Nonparametric Analysis of Cultivar Performance Trials: Experimental Results and Comparison of Different Procedures Based on Ranks*, Agronomy Journal, vol. 87, 627-632. URL: http://www.scilib.univ.kiev.ua/article.php?1091042 .1995.

Neter, J., Wasserman, W., Kutner, M.H. 1990. *Applied Linier Statistical Models: Regression, Analysis of Variance, and Experimental Design*, 3rd Edition, Rhicard D. Irwin, Inc., Boston.

Ratnapani, A.S.. 2004. *Penetapan Kadar Logam Timbal (Pb) dan Khromium (Cr) pada Jamu Cina Berbagai Merek dalam Kemasan Kaleng yang Beredar di Pasaran dengan Metode Spektrofotometer Serapan Atom*, UII Yogyakarta.

Siegel, S., Castellan, N.J. 1988. *Nonparametric Statistics for The Behavioral Science*, 2nd Edition, McGraw-Hill, Singapore.

A Family of Non-Parametric Tests for Decreasing Mean Time to Failure with Censored Data

Deemat C Mathew*
St. Thomas College Palai, Kottayam, India - deematcm@yahoo.com
Sudheesh K Kattumannil
Indian Statistical Institute, Chennai, India - skkattu@isichennai.res.in

Abstract

In this paper, we obtain an exact test for testing exponentiality against decreasing mean time to failure class alternatives. Asymptotic properties of the test statistic are studied. We also discussed how the right censored observations can be incorporated in the proposed testing method. We studied the efficiency loss due to censoring and calculated efficiency loss for censoring random variables.

Keywords: Exponential distribution; Mean time to failure; Pitman's asymptotic efficacy; Replacement model; Right censoring.

1. INTRODUCTION

The class which deals with failure with replacement distribution plays a key role in reliability theory as engineers can design appropriate maintenance policies for a particular task. Planned replacements are generally preferred to unscheduled maintenance to reduce in-service costs that are inherent in unexpected failures. In such cases, one strategy is to resort an age replacement policy in which an item is replaced either when it fails or at a fixed age t whichever is earlier. In this context, Barlow and Proschan (1965) introduced the concept of mean time to failure (MTTF) to describe the reliability characteristics of a repairable system and studied the monotonic behaviour of MTTF.

The study of decreasing mean time to failure (DMTTF) classes of life distributions has been received much attention in recent time; see Knopik (2006), Li and Xu (2008), Asha and Nair (2010), Kayid et al. (2013) and Kayid and Izadkhah (2014) and the references the rein. If the average waiting time between the consecutive failures is an important criterion in determining replacement policy, then the problem of testing exponentiality against DMTTF class has great importance. Li and Xu (2008) and Kayid et al. (2013) have initiated some work in this direction and their test statistics are computationally complex. Motivated by these recent developments, we propose a family of non-parametric tests for testing exponentiality against DMTTF class alternative.

Semi-parametric and non-parametric analysis of right censored observations are extensively studied in literature. Hollander and Proschan (1975) developed a test for exponentiality against DMRL alternative and Chen et al. (1983) extended their test to the case of randomly right censored data. Bergman and Klesfjo (1989) developed a family of test statistics for testing exponentiality against DMRL when the data is both complete and censored. Lim and Park (1993) generalized Ahmad (1992) test to accommodate the situation that the data is randomly censored and compared their test with Chen et al. (1983) test. In this study, we discuss how to incorporate right censored observation in the proposed testing method.

The rest of the paper is organized as follows. In Section 2, we propose a family non-parametric tests for testing exponentiality against DMTTF class. We derive the exact null distribution of the test statistic and then calculated the critical values for different sample sizes. The asymptotic properties of the test statistic are proved in Section 3. In Section 4, we discuss how to incorporates the right censored observations in our study. In Section 5, we conclude the study.

2. EXACT TEST

Let X be a non-negative random variable with absolutely continuous distribution function F (.). Suppose $\overline{F}(x)$ = P(X> x) denotes the survival function of X at x. Also let.

$$\mu = E(X) = \int_0^\infty \overline{F}(t)dt < \infty$$

Consider an age replacement policy in which a unit is replaced by a new one (whose lifetime distribution is the same as F) at actual failure or a fixed time t after installation, whichever occur first and let X[t] be the associated random variable of interest. The survival function of X[t] is given by (Barlow and Proschan, 1965)

$$S_t(x) = \sum_{n=0}^{\infty} \overline{F}^n(t)\overline{F}(x-nt)I_{[nt,(n+1)t]}(x), x > 0$$

where I (.) denotes the indicator function. The MTTF, denoted by M (t) is given by

$$M(t) = \frac{\int_0^t \overline{F}(x)dx}{F(t)}, \text{for all } t > 0$$

Definition 2.1 *A random variable X belongs to the DMTTF class if the function M (.) is non-increasing for all t > 0.*

Next, we develop a family of test for testing exponentiality against DMTTF class. We are interested to test the hypothesis

H_0: F is exponential vs H_1: F is DMTTF (and not exponential),

on the basis of the random sample $X_1, X_2, \ldots, X_n$; from F. A measure of deviation from exponentiality towards DMTTF class is given by

$$\Delta(F) = \frac{k+2}{k+1}\int_0^\infty x d\overline{F}^{k+2}(x) - \int_0^\infty x d\overline{F}^{k+1}(x) \tag{2}$$

The natural plug-in estimator for $\Delta(F)$ is given by

$$\hat{\Delta} = \frac{k+2}{k+1}\sum_{i=1}^{n} X_{(i)}\left[\left(1-\frac{i-1}{n}\right)^{k+2} - \left(1-\frac{i}{n}\right)^{k+2}\right] - \sum_{i=1}^{n} X_{(i)}\left[\left(1-\frac{i-1}{n}\right)^{k+1} - \left(1-\frac{i}{n}\right)^{k+1}\right],$$

where $X_{(i)}$, $i = 1,2,\ldots,n$ be the i-th order statistics based on the random sample $X_1, X_2, \ldots, X_n$; from F. To make the test scale invariant we consider the measure . $\Delta^*(F) = \frac{\Delta(F)}{\mu}$ Hence the test statistic is

$$\hat{\Delta}^* = \frac{\hat{\Delta}}{\overline{X}} \tag{3}$$

where is $\overline{X}$ the sample mean. We reject the null hypothesis H_0 against alternative hypothesis H_1 for large values of $\hat{\Delta}^*$.

The exact null distribution of $\hat{\Delta}^*$ is obtained in the following theorem.

Theorem 2.1 *Let X be a continuous non-negative random variable with* $\overline{F}(x) = e^{-\frac{x}{2}}$. *Let X1, X2,…,Xn* random sample from F. Then for fixed n

$$P(\widehat{\Delta}^* > x) = \sum_{i=1}^{n} \prod_{j=1, j\neq i}^{n} \left(\frac{d_{i,n} - x}{d_{i,n} - d_{j,n}}\right) I(x, d_{i,n}),$$

provided di;n ≠ dj;n when i ≠ j, where

$$I(x, y) = \begin{cases} 1 \; if \; x \leq y \\ 0 \; if \; x > y \end{cases}$$

and

$$d_{i,n} = \frac{1}{k+1}\left(\frac{n-1+1}{n}\right)^k \left[(k+2)\left(\frac{n-1+1}{n}\right) - (k+1)\right]$$

We can rewrite $\hat{\Delta}^*$ given in equation (3) as follows

$$\widehat{\Delta} = \sum_{i=1}^{n} X_{(i)}\left(1 - \frac{i-1}{n}\right)^{k+1}\left[\left(1 - \frac{i-1}{n}\right)\left(\frac{k+1}{k+2}\right) - 1\right] - \sum_{i=1}^{n} X_{(i)}\left(1 - \frac{i}{n}\right)^{k+1}\left[\left(1 - \frac{i}{n}\right)\left(\frac{k+1}{k+2}\right) - 1\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} d_{i,n} D_i$$

where $D_i = (n-i+1)(X_{(i)} - X_{(i-1)})$ with: $X_{(i)} = 0$. Hence $\widehat{\Delta}^* = \frac{\sum_{i=1}^{n} d_{i,n} D_i}{\sum_{i=1}^{n} D_i}$; where di;n's are as given in the theorem. Note that the exponential random variable with rate ½ is same as the χ^2 random variable with 2 degrees of freedom. Hence the result follows from Theorem 2.2 of Box (1954).

The critical values of the exact test for different values of *n* and *k* are tabulated in Table 1.

Table 1. Critical Values

	n	90% level	95% level	97.5% level	99% level
k = 0	2	0.9000	0.9500	0.9750	0.9900
	4	0.5783	0.6652	0.7342	0.8045
	10	0.3237	0.3854	0.4378	0.4975
	15	0.2522	0.3038	0.3481	0.3985
	20	0.2118	0.2571	0.2961	0.3409
	50	0.1237	0.1530	0.1783	0.2075
	100	0.0838	0.1045	0.1225	0.1435
k = 0.25	2	0.7116	0.7558	0.7779	0.7912
	4	0.4320	0.5078	0.5683	0.6289
	10	0.2264	0.2761	0.3191	0.3685
	15	0.1732	0.2135	0.2487	0.2895
	20	0.1441	0.1788	0.2090	0.2442
	50	0.0827	0.1040	0.1227	0.1446
	100	0.0558	0.0706	0.0835	0.0987
k = 0.5	2	0.5822	0.6274	0.6470	0.6588
	4	0.3432	0.4100	0.4629	0.5164
	10	0.1719	0.2131	0.2493	0.2910
	15	0.1302	0.1629	0.1917	0.2255
	20	0.1078	0.1355	0.1600	0.1887
	50	0.0614	0.0780	0.0926	0.1098
	100	0.0413	0.0527	0.0626	0.0744
k = 1	2	0.4375	0.4687	0.4844	0.4937
	4	0.2446	0.2973	0.3391	0.3815
	10	0.1156	0.1459	0.1730	0.2049
	15	0.0864	0.1097	0.1306	0.1557
	20	0.0711	0.0904	0.1079	0.1288
	50	0.0399	0.0511	0.0609	0.0728
	100	0.0267	0.0341	0.0408	0.0486

3. ASYMPTOTIC PROPERTIES OF THE TEST STATISTIC

Here first we derive the asymptotic distribution of the proposed test statistic. We also study the symptotic efficiency of the test.

Theorem 3.1 As $n \to \infty$,, the distribution of $\sqrt{n}(\widehat{\Delta} - \Delta(F))$ converges to Gaussian with mean zero and variance

$$\sigma^2 = 2\int_0^\infty \int_z^\infty \overline{F}(x)F(z)J(F_X(x))J(F_X(z))dxdz$$

where

$$J(u) = \frac{1}{k+1}\left[(k+2)^2(1-u)^{k+1} - (k+1)^2(1-u)^k\right].$$

Proof: Consider

$$\widehat{\Delta}' = \frac{1}{n} = \sum_{i=1}^{n} X_{(i)}J(F_n(x)),$$

With $J(u) = \frac{1}{\mathrm{k}+1}\left[(k+2)^2(1-u)^{k+1} - (k+1)^2(1-u)^k\right]$, , where $F_n(.)$ is the empirical distribution function. It is easy to see

$$\widehat{\Delta} = \widehat{\Delta}' + \sigma(1/\sqrt{n})$$

Note that $\hat{\Delta}$ and $\hat{\Delta}'$ have same asymptotic distribution. The asymptotic normality of $\hat{\Delta}'$ follows from Theorem 1 of Sandstrom (1987) by taking uniform weight in weighted empirical distribution function. Using Slutsky's theorem we have the following result.

Corollary 3.1 *As* $n \to \infty$, *the distribution of* $\sqrt{n}\left(\hat{\Delta}^* - \left(\frac{\Delta(F)}{\mu}\right)\right.$ *converges to Gaussian with mean zero and variance* $\frac{\sigma^2}{\mu^2}$.

Corollary 3.2 *Le X be continuous non-negative random variable with then as* $n \to \infty$, *the distribution of is Gaussian* $\sqrt{n}\widehat{\Delta}^*$ *with mean zero and variance* $\overline{F}(x) = e^{-\frac{x}{\lambda}}$

$$\sigma_0^2 = \frac{1}{(k+1)(2k+1)(2k+3)}$$

When k = 0, the variance in Corollary 3.2 reduces to 1/3. In case of asymptotic test, for large values of n, we reject the null hypothesis H_0 in favour of the alternative hypothesis H_1, if

$$\frac{\sqrt{n}\widehat{\Delta}^*}{\sigma_0} > Z_\alpha,$$

where Z_α is the upper α-percentile of $N(0,1)$.

The Pitman efficiency is the most frequently used index to make a quantitative comparison of two distinct asymptotic tests for a certain statistical hypothesis. The Pitman's asymptotic efficacy (PAE) is defined as

$$PAE(\Delta^*(F)) = \frac{\left|\frac{d}{d_\lambda}\Delta^*(F)_{\lambda\to\lambda_0}\right|}{\sigma_0} \tag{4}$$

where λ_0 is the value of λ under H0. We calculate the PAE value for two commonly used alternatives which are the members of DMTTF class

(i) the Weibull family: $\overline{F}(x) = e^{-x^{\lambda}}$ for $\lambda > 1, x \geq 0$
(ii) the linear failure rate family: $\overline{F}(x) = e^{(-x-\frac{\lambda}{2}x^2)}$ for $\lambda > 0\,, x \geq 0$
Exponential distribution is the special case of Weibull distribution when λ = 1 and the linear failure rate family reduces to exponential when λ = 0. With some algebraic calculations we can show that the efficacy values for these families are

1. Weibull family:

$$PAE = \frac{\sqrt{(2k+1)(2k+3)}}{\sqrt{(k+1)}}\log\left(\frac{k+2}{k+1}\right)$$

2. Linear failure rate family:

$$PAE = \frac{\sqrt{(2k+1)(2k+3)}}{(k+1)(k+2)\sqrt{(k+1)}}$$

The maximum values of PAE are located at approximately at k = 0.25 and 0 for the Weibull and linear failure rate families respectively.

4. CASE WITH CENSORED OBSERVATIONS

In this section we extend the proposed testing method to accommodate right censored observations, which is common in lifetime data analysis.

Suppose we have randomly right-censored observations such that the censoring times are independent of the lifetimes. Under this set up the observations are n independent and identical copies of (Y, δ), with $Y = min(X, C)$, where C is the censoring time and $\delta = I(X \leq C)$. Now we need to address the testing problem mentioned in Section 2 based on n independent and identical observations $\{(Y_i, \delta_i), 1 \leq i \leq n\}$. Observe that $\delta_i = 1$ means ith object is not censored, whereas $\delta_i = 0$ means that i^{th} object is censored by C ,on the right. The Kaplan-Meier estimator of the survival function is given by

$$\overline{F}_n = \prod_{i:Y_{(i)} \leq t}\left(\frac{n-1}{n-i+1}\right)^{\delta_{(i)}}$$

for $t \in [0, Y_{(i)}]$ where $Y_{(i)}$ is the ordered Y 's. Replacing the survival function in Δ using $\overline{F}_n$ we obtain the test statistic in the right censored case as

$$\widehat{\Delta}_c = \sum_{i=1}^{n}\left\{\frac{k+2}{k+1}\prod_{j=1}^{i-1}\left[(n-j)(n-j+1)^{-1}\right]^{(k+2)\delta_{(i)}} - \prod_{j=1}^{i-1}\left[(n-j)(n-j+1)^{-1}\right]^{(k+1)\delta_{(i)}}\right\}(Y_{(i)} - Y_{(i-1)})$$

We consider the scale invariant test statistic

$$\widehat{\Delta}_c^* = \frac{\widehat{\Delta}_c}{\overline{Y}_c},$$

Where $\overline{Y}_c$ a consistent estimator of μ given by $\overline{Y}_c = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\delta_i}{\widehat{K}_c(Y_i)}$ (Zhao and Tsiatis, 2000) and $\widehat{K}_c$ is the Kaplan-Meier estimator of $\overline{G}_c$, the survival function of the censoring variable C. Hence in right censoring situation, the test procedure is to reject H_0 in favour of H_1 for large values of $\hat{\Delta}_c^*$.

Next we obtain the limiting distribution of the test statistic. To establish the asymptotic normality of the statistic $\hat{\Delta}_c^*$ we assume the following conditions on the distributions of X and C.

1. For some $\beta \in (0,1)$,

$$\int_0^{\infty}\overline{F}^{\beta}(x)dx < \infty \text{ and } \int_0^{\infty}\left[\overline{F}^{\beta}(x)\overline{G}(x)\right]^{-1}dF(x)dx < \infty \qquad (5)$$

2. $n^{1/2}\int_{M_n}^{\infty}\overline{F}(x)dx$ converges in probability to 0, where $M_n = \max(Y_1,\dots Y_n)$.

Theorem 4.1 Under the above assumptions, as $n \to \infty$, *the distribution of* $\sqrt{n}(\overline{\Delta}_c - \Delta(F))$ *is Gaussian with mean zero and variance* σ_{1c}^2, *where* σ_{1c}^2 *is given by*

$$\sigma_{1c}^2 = \int_0^\infty \int_0^\infty \overline{F}(x)\overline{F}(y)JF(x))J(F(y)) \int_0^{\min(x,y)} [\overline{H}(u)\overline{F}(u)]^{-1} dF(u)dxdy\,.$$

Proof: The result follows from Theorem 1 of Joe and Proschan (1982).

Corollary 4.1 *Under the assumptions of Theorem 4.1, the distribution of* $\sqrt{n}(\widehat{\Delta}_c^* - \Delta^*)$ as $n \to \infty$, is Gaussian with mean zero and variance σ_c^2, where

$$\sigma_c^2 = \frac{\sigma_{1c}^2}{\mu^2}$$

Proof: Note that $\overline{Y}_c$ is a consistent estimator for μ. Hence the result follows from Theorem 4.1 by applying Slutsky's theorem.

Corollary 4.2 *Let X be continuous non-negative random variable with* $\overline{F}(x) = e^{-x}$. *Under the assumptions as* $n \to \infty$, *the distribution of* $\sqrt{n}\widehat{\Delta}^*$ *is Gaussian with mean zero and variance* σ_{c0}^2, *where* σ_{c0}^2

$$\sigma_{c0}^2 = \int_0^\infty \left(\overline{G}(u)\right)^{-1} \frac{e^u}{(k+1)^2}\left[(k+2)e^{-(k+2)u} - (k+1)e^{-(k+1)u}\right]du \tag{6}$$

We can also express σ_{c0}^2 as

$$\sigma_{c0}^2 = \int_0^\infty \left(\overline{G}(u)\right)^{-1} \frac{1}{(k+1)^2}\left[(k+2)x^{(k+1)} - (k+1)x^k\right]dx \tag{7}$$

If all the observations are uncensored, $\overline{G}(u) = 1 \forall u$ and we obtain

$$\sigma_{c0}^2 = \frac{1}{(k+1)(2k+1)(2k+3)}$$

which is same as the asymptotic null variance for the complete data. An estimator $\widehat{\sigma}_{c0}^2$ for σ_{c0}^2 is obtained by replacing with $\overline{G}(u)$ its Kaplan-Meier estimator. Hence in case of right censoring we reject the null hypothesis in favour of H_1, if

$$\frac{\sqrt{n}\Delta_c^*}{\widehat{\sigma}_{c0}} > Z_\alpha\,.$$

Next we study the efficiency loss due to censoring by computing the efficacy value of our test based on $\hat{\Delta}_c^*$ for uncensored model and the efficacy value of the test based on $\hat{\Delta}_c^*$ for censored model. As both these tests have same asymptotic mean, the Pitman's asymptotic relative efficiency (ARE) of the test based on $\hat{\Delta}_c^*$ with respect to the test based on $\hat{\Delta}^*$ is given by

$$e = ARE(\Delta_c^*, \Delta^*) = \frac{\sigma_0^2}{\sigma_{c0}^2}$$

The quantity (1–e) can be taken as a measure of the efficiency loss (Lim and Park, 1993) due to censoring. As illustration we calculate the ARE value when the censoring variable C has different distributions. The ARE value when C is exponential with survival function $F(x) = e^{-\frac{x}{\lambda}}$ is given in Table 2. From Table 2 it is clear that as λ decreases, the efficiency loss decreases. We also calculate ARE value when the censoring variable C has logistic distribution with distribution function $F(x) = \frac{1}{1+e^{-\frac{x}{\lambda}}}$ and the corresponding values are tabulated in Table 3. In this cases also the efficiency loss decreases as λ decreases.

Table 2. Asymptotic relative efficiency when censoring random variable is exponential

λ	1/2	1/3	1/4	1/10	1/20	1/50	1/100
k = 0	0.357	0.556	0.664	0.865	0.933	0.973	0.987
k = 0.25	0.544	0.697	0.774	0.911	0.956	0.982	0.991
k = 0.5	0.656	0.775	0.833	0.935	0.968	0.987	0.994
k = 1	0.778	0.856	0.893	0.958	0.979	0.992	0.996

5. CONCLUSION

In the present study, we obtained an exact test for testing exponentiality against decreasing mean time to failure class alternatives. Asymptotic properties of the test statistic are studied and showed the test statistics has asymptotic Gaussian distribution. PAE is calculated for different alternatives to study the efficiency of test. It is calculated that maximum values of PAE are located at approximately at k = 0.25 and 0 for the Weibull and linear failure rate families respectively. We also discussed how the right censored observations can be incorporated in the proposed testing method. We studied the efficiency loss due to censoring and calculated efficiency loss for censoring random variables.

Table 3. Asymptotic relative efficiency when censoring random variable is logistic

λ	1/2	1/3	1/4	1/10	1/20	1/50	1/100
k = 0	0.526	0.714	0.798	0.928	0.965	0.986	0.993
k = 0.25	0.705	0.821	0.873	0.954	0.978	0.991	0.996
k = 0.5	0.792	0.873	0.909	0.966	0.984	0.994	0.997
k = 1	0.875	0.922	0.944	0.979	0.989	0.996	0.998

REFERENCES

Ahmad, I.A. (1992). A new test for mean residual life time, Biometrika, 79, 416-419.

Asha, G. and Nair, N.U. (2010). Reliability properties of mean time to failure in age replacement models, International Journal of Reliability, Quality and Safety Engineering, 17, 15-26.

Barlow, R. E. and Proschan, F. (1965). Mathematical Theory of Reliability, Wiley, New York.

Bergman, B. and Klesfjo, B. (1989). A family of test statistics for detecting monotonic mean residual life, Journal of Statistical Planning and Inferences, 21, 161-178.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification.Annals of Mathematical Statistics, 25, 290-302.

Chen, Y.Y., Hollander, M. and Langberg, N.A. (1983). Tests for monotone mean residual life, using randomly censored data, Biometrics, 39, 119-127.

Hollander, M., Proschan, F., (1975). Tests for mean residual life, Biometrika, 62, 585-593.

Joe, H. and Proschan, F. (1982). Asymptotic normality of L-statistics with randomly censored data, Technical report, Florida University. M613.

Kayid, M., Ahmad, I. A., Izadkhah S. and Abouammoh, A. M. (2013). Further results involving the mean time to failure order and the decreasing mean time to failure class, IEEE Transactions on Reliability, 62, 670-678.

Kayid, M. and Izadkhah S. (2014). Mean inactivity time function, associated orderings and classes of life distributions, IEEE Transactions on Reliability, 63, 593-602.

Knopik, L. (2006). Characterization of a class of lifetime distributions, Control and Cybernetics, 35, 407-414. Li, X. and Xu, M. (2008). Reversed hazard rate order of equilibrium distributions and a related ageing notion, Statistical Papers, 49,749-767.

Lim, J.H. and Park, D.H. (1993). Test for DMRL using censored data, Journal of Nonparametric statistics, 3, 167- 173.

Sandstrom, A. (1987). Asymptotic normality of linear functions of concomitants of order statistics, Metrika, 34, 129-142.

Zhao, H. and Tsiatis, A. A. (2000). Estimating mean quality adjusted lifetime with censored data. Sankhya- Series B, 62, 175-188.

CPS10: Demography & Social Welfare Statistics (1)

Empirical Study of Unemployment Disparities and Labor Market Structures at Sub National Level in Indonesia using Spatial Panel Data Analysis, 2004-2014

Putu Wira Wirbuana
Statistics Indonesia, Jakarta, Indonesia – putuwiraw@bps.go.id
Chindy Saktias Pratiwi
Statistics Indonesia, Jakarta, Indonesia – chindy.pratiwi@bps.go.id

Abstract

The eighth goal of Sustainable Development Goals is to promote decent work and economic growth. This indicates that numerous problems related to labor force such as unemployment still become issues. Furthermore, analysis of labor market mostly focused at the national level. However, studies in some countries showed that there were persistent regional disparities in employment performance (OECD, 2005). Indonesia consists of provinces (sub-national level) with various factors endowments that make their economic structures different. It encourages different speed of development in each province. There are many researches about unemployment in Indonesia but the one that includes spatial effect is still limited. The aim of this study is to see the existence of unemployment disparities among provinces in Indonesia related to spatial effect and its changes during 2004-2014. Furthermore, this study develops a model to investigate the relationship between labor structures and unemployment disparities. The empirical study focused on the case of Indonesia's provinces from 2004 to 2014 with 30 provinces. Using Moran's I statistics, this study shows there were spatial autocorrelation of unemployment disparity between provinces in Indonesia especially for the period 2004-2005 and 2008-2012. Moreover Spatial Lag Fixed Effect Model shows that the spatial dependence is exist but it does not directly affect the unemployment disparity in each provinces. However, the spatial specific effects from this estimation model significantly affect the unemployment disparity. Besides, participation rate, share of employment in agricultural and manufacturing sector, and ratio of female labor over male labor also significant affect unemployment disparity.

Keywords: Unemployment disparity; Regional unemployment; Moran's I; Spatial Lag Fixed Effect Model.

1. INTRODUCTION

The eighth goal of Sustainable Development Goals is to promote decent work and economic growth. This indicates that numerous problems related to labor force and economic development still become issues. One of highlighted socioeconomic issues in many countries concerns on labor force particularly about unemployment, including Indonesia.

The labor policy in Indonesia focused on decreasing the unemployment rate. Many policies have been done in order to lower the unemployment rate. However that policies did not optimally enough to solve unemployment problems.

In particular, the analysis of labor market focused at the national level. However, in study at some countries, there are persistent regional disparities in employment performance. There are countries where labor shortages in certain regions coexist with continuously high unemployment in other regions. It is therefore important to assess the extent to which such disparities persist (OECD, 2005). The presence of geographical disparities in unemployment becomes a central issue in economics. Although regional unemployment rates broadly move in line with national rate, differences across regions remain over longer periods of time (Lolos and Papapetrou, 2011).

If it is connected with Indonesia's condition, although the unemployment rate at national level was in line with its target-Indonesia's medium term development plan for 2010-2014 targeted to reduce unemployment between 5 and 6 per cent-in 2014 at about 5.94 percent. This achievement is not followed by the unemployment rate in provincial level. Through geographical analysis, these rates may be different between provinces because of the characteristic of Indonesia as archipelago country. Indonesia consists of provinces with different economic structures according to the various factors endowments possessed. It lead regional disparities in economic performance because of the differences in the speed of economic growth among provinces (Yuniasih, 2013).

For the labor market itself, slightly less than 60 per cent of the total Indonesian labor force is concentrated in Java and Bali Island, Sumatera and Java Island alone absorb 78.9 percent of labor force, which is similar to the pattern of economic activities concentration (ILO, 2012). Furthermore, unemployment rate data of Statistics Indonesia shows that the unemployment rate in each province also diverse. Unemployment rates ranged from 3.1 percent in Bali to 13.7 percent in Banten in 2010. The range is still high in 2014 with the lowest unemployment rate was Bali (central Indonesia) at 1.90 per cent and the highest was Maluku (eastern Indonesia) at 10.51 per cent. Indonesia exhibits persistent regional inequalities in both its economic and employment outcomes. Inequalities in regional development have been linked to factors such as population density, resource endowment and geographic isolation. In general employment and economic outcomes tend to be worse in the eastern parts of Indonesia (ILO, 2013).

Many researches about unemployment had been developed in Indonesia. However, the study that includes spatial effect is still limited. This study focuses to assess the existence of unemployment disparities among provinces (sub-national level) in Indonesia, related to spatial effect and its changes during 2004-2014. Furthermore, this study develops a model to investigate the relationship between labor structures and unemployment disparities from 2004 to 2014.

2. LITERATURE REVIEW

Unemployment disparities are interpreted as a result of limited interregional labor mobility or of differences in the characteristics of the regional labor markets (Lolos and Papapetrou, 2011).

Cracolici, Cuffaro and Nijkamp (2007) assessed the spatial structure of regional unemployment disparities of Italian provinces (103 provinces) for the years 1998 and 2003. The unemployment disparity is calculated by using the differences between the unemployment rate in each province and the nation-wide unemployment. The proxies for provincial economic structure variables are share of employment in three sectors, which are manufacturing, agriculture, and service sector. While the proxies for mobility equilibrium variable is number of vacant (non-occupied) houses over the total numbers of available houses. Female labor force over total females at working age, male labor force over males at working age, population of 15-19 years old over total population, population over 65 years over total population, net migration balance are used as proxies for demographic equilibrium variables. The last variable is population density that is used as a proxy for consumer and producer amenities. The spatial autocorrelation is measured using a spatial weight matrix which is a contiguity matrix. This study uses the robust LM test to detect the spatial effects and OLS to estimate the model. The results suggest that there is a clear explanation of unemployment differentials and a significant degree of spatial dependence among labor markets at the provincial level in Italy. Provinces marked by high unemployment, as well as those characterized by low unemployment, tend to be spatially clustered, demonstrating the presence of unemployment 'persistency' in space and time regimes.

Other empirical studies on spatial dependence in regional unemployment were done by Mitchell and Bill (2004), assessed the regional unemployment disparities in Australia using cross sectional data with the control variables population density, labor force participation rate, human capital variables, industrial composition variable of employment, industrial composition variable, and other indicators including proportion using internet; Lottmann (2012) explained the regional unemployment differences in Germany using spatial panel data using 24 explanatory variables for the period from 1999 until 2007 and showed that German regional unemployment is of disequilibrium nature, which justifies political interventions. Furthermore, in particular Lolos and Papapetrou (2011) assessed unemployment disparities in Greek regions from 1981 to 2008 using GMM; Rios (2014) investigated the evolution of the geographical distribution of unemployment rates in European regions between 2000 and 2011 using panel data.

3. METHODS

Data

The empirical study focused on the case of Indonesia's provinces (sub-national level). Data that is obtained in this study is from Labor Force Survey which is conducted by Statistics Indonesia. The period study is from 2004 to 2014. Indonesia is currently divided into 34 provinces. However, because of the boundaries of some administrative provinces were changed, in this study, only the provinces that already existed in 2004 (30 provinces) were considered for the analysis.

The variables used in this study contain unemployment disparity as dependent variables and labor structures (participation rate, labor supplies that is represented by share of employment in the agricultural sector and manufacturing sector, and employment by gender) as independent variables.

Moran's I

Moran's I gives a formal indication of the degree of linear association between a vector of observed values *ud* and a weighted average of the neighbouring values, or spatial lag, W_{UD}. Queen contiguity is applied which method is started with a matrix W filled with zeros, then consider entities that share a common side or vertex with province of interest define W_{ij}=1.

The linear association between *ud* and W_{UD} underlines the spesification of spatial autoregressive processes, which are typically used to expres the generating mechanism behind the spatial dependence. Moran's I statistics can be expressed in matrix notation as:

$$I=(N/S_0)ud'W_{UD}/ud'ud \tag{1}$$

Where N stand for the number of province (30), S_0 is the sum of all elements in the spatial weights matrix ($S_0 = \Sigma_i \Sigma_j w_{ij}$), u_d are *UD* in deviations from the mean, and W_{UD} is the associated spatial lag.

The empirical model

This study uses spatial panel data model. In order to assess the spatial dependence, this study uses the starting model:

$$UD_{it} = \beta_0+\beta_1 PR_{it}+\beta_2 EAGR_{it}+\beta_3 EIND_{it}+\beta_4 FM_{it}+\varepsilon_{it} \tag{2}$$

for t =2004, 2005, …, 2014 and i is each provinces in Indonesia

The dependent variable UD_{it} is an unemployment difference, the difference between unemployment rate in each province in year t and the national unemployment rate in the same period. While the explanatory variables comprise labor demand that is represented by PR_{it} which is participation rate-the ratio of labor force and working age population- of province i at year t; labor supplies that is represented by $EAGR_{it}$ and $EIND_{it}$, which are share of employment in agricultural sector ($EAGR_{it}$) and in manufacturing sector ($EIND_{it}$) over provincial employment of province i at year t; and demographic equilibrium that is represented by FM_{it} is the ratio of female labor to male labor of province i at year t.

Anselin (1993) explained that spatial dependence may incorporate in two distinct way: as a spatially lagged dependent variable (*spatial lag* model-SAR) or a spatially autoregressive process in the error term (*spatial error* model-SEM). Elhorst (2009) explained that in SAR model, the dependent variable depends on the dependent variable observed in neighbour units and on a set of observed local characteristics. The SAR model can be formulated as:

$$UD_{it}=\delta\Sigma_{j=1}^{N} w_{ij}\, UD_{jt} + \boldsymbol{x_{it}\beta}+\mu_i+\varepsilon_{it} \tag{3}$$

where δ is called the spatial autoregressive coefficient and w_{ij} is an element of a spatial weights matrix $\boldsymbol{W}$ describing the spatial arrangement of the units in the sample. It is assumed that $\boldsymbol{W}$ is a pre-specified non-negative matrix of order N^2. Anselin et al (2006) in Elhorst (2009) stated that the spatial lag model is the formal specification of the equilibrium outcome of a spatial or social interaction process. Therefore, it can be stated that the value of the unemployment disparity for one province is determined by the unemployment disparity in neighboring provinces.

On the other hand, according to Elhorst (2009), SEM posits that the dependent variable depends on a set of observed local characteristics and that the error terms are correlated across space. The SEM model can be formulated as:

$$UD_{it}=\boldsymbol{x}_{it}\boldsymbol{\beta}+\mu_i+\phi_{it}, \tag{4a}$$

$$\phi_{it}=\rho\Sigma_{j=1}^{N} w_{ij}\,\phi_{it}+\varepsilon_{it} \tag{4b}$$

where ϕ_{it} reflects the spatially autocorrelated error term and ρ is the spatial autocorrelation coefficient. Elhorst (2003) mentioned that there were four panel data models commonly used in applied research extended to include spatial error autocorrelation or a spatially lagged dependent variable: fixed effects, random effects, fixed coefficients, and random coefficients models. The simple pooled linear regression model with spatial specific effects is considered but without spatial interaction effects is:

$$UD_{it}=\boldsymbol{x}_{it}\boldsymbol{\beta}+\mu_i+\varepsilon_{it} \tag{5}$$

μ_i denotes a spatial specific effect. The spatial specific effects are needed to control for all space-specific time-invariant variables whose omission could bias the estimates in a typical cross-sectional study. According to Elhorst (2009) these spatial specific effects may be treated as fixed effects or as random effects. In the fixed effects model, a dummy variable is introduced for each spatial unit, while in the random effects model, μ_i is treated as a random variable that is independently and identically distributed with zero mean and variance σ_μ^2. Furthermore, the random variables μ_i and ε_{it} are assumed independent of each other.

In order to get the best estimation model for spatial panel data, Lagrange Multiplier (LM) test is used for spatially lagged dependent variable and for spatial error correlation. If these two LM tests are rejected, then Robust LM test is needed.

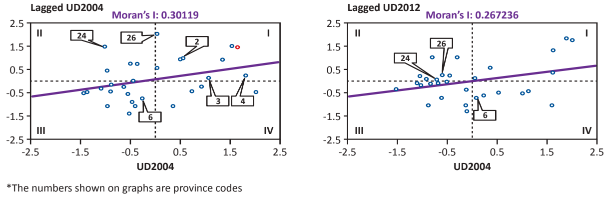
After determining SAR or SEM that is used, the next step is determining whether fixed effect model or random effect model that is used. This step can be done through Hausman test. The statistics of Hausman test is:

$$W=(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})'[Var(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE})]^{-1}(\hat{\boldsymbol{\beta}}_{FE}-\hat{\boldsymbol{\beta}}_{RE}) \tag{6}$$

This statistics is chi-squared distributed with degree of freedom is the numbers of independent variables, which is 4. The null hypothesis of Hausman test is that random effect is better than fixed effect as the model.

4. RESULTS

The result of Moran's I shows that in 2004-2005 and 2008-2012 give positive value and significant at alpha 10 percent. Those were indicated the existence of spatial autocorrelations on unemployment disparities between provinces in Indonesia. In order to know the pattern of these spatial dependences, it can be analyzed through Moran scatter plot. For analysis purposes, Moran scatterplots in 2004 and 2012 are displayed in order to see the shifting of pattern between 2004 and 2012.



*The numbers shown on graphs are province codes

Figure 1. Moran Scatterplot of Regional Unemployment Disparity in: (a) 2004;(b) 2012

Figure 1(a) shows Moran scatterplot of regional unemployment disparity in 2004. According to Figure 1(a), positive spatial dependence is described by the first quadrant which had higher provincial unemployment rate than the national unemployment rate surrounded by similar unemployment rates of neighboring provinces. The province with highest positive unemployment disparity was Riau[4]. Riau[4] had 5.39 percent in unemployment disparity and its neighboring provinces, Sumatera Utara[2] and Sumatera Barat[3] also had similar value (positive and high) of unemployment disparities, which were 1.22 percent and 2.88 percent.

After nine years, from 2004 to 2012, there were changes in Moran scatterplot composition. According to Moran scatterplot in Figure 1(b), twelve provinces shifted from the original quadrant in 2004. Some of them were Sumatera Selatan[6], Sulawesi Tengah[24] and Sulawesi Tenggara[26]. Sumatera Selatan[6] shifted from third quadrant in 2004 to forth quadrant in 2012. In 2004, unemployment disparity of Sumatera Selatan[6] was negative, and so were its neighboring provinces. In 2012, Sumatera Selatan experienced[6] a lower unemployment disparity, from -1.49 percent in 2004 then decreased at -0.48 percent in 2012. While for the neighboring provinces around Sumatera Selatan[6], they also experienced lower unemployment disparity, but the value of unemployment disparity were higher than Sumatera Selatan[6]. This condition made Sumatera Selatan[6] shifted from third quadrant in 2004 to fourth quadrant in 2012.

The empirical investigation is started by assessing the existence of spatial dependence among provinces then examining the relationship between labor structures and unemployment disparities. According to Table 1, both LM (spatial lag) and LM (spatial error) are significant in 10% significance level. These indicate that there were spatial dependence for the unemployment disparity in 2004-2014. While the result of robust LM showed high *p-value* that is insignificant. However, from the result, the *p-value* of robust LM (lag) is higher than LM (error), therefore spatial lag model (SAR) is used as the estimation model. These result showed that each province is not independent to the others.

Table 1. Result of LM test and Robust LM test

LM test	Statistics	$\chi^2_{(0.05,1)}$	p-value
LM-SAR	3.223259	3.841459	0.07259898
LM-SEM	3.236379	3.841459	0.07201964
Robust LM-SAR	0.003966529	3.841459	0.9497821
Robust LM-SEM	0.017087074	3.841459	0.8959988

In spatial panel data analysis, there are two types of panel data model which are fixed effect and random effect model. In order to assess which model is the best, Hausman test statistics is calculated. Significant results from hausman test means that fixed effect model is better than random effect model. Therefore, Spatial Lag Fixed Effect Model is used in this study. The estimation model is shown by following equation.

$$y_{it} = 0.0223\,\Sigma_{j=1}^{n} w_{ij}\,y_{it} + 0.1581^{***}\,PR_{it} - 0.126487^{***}\,EAGR_{it} - 0.215976^{***}\,EIND_{it} - 13.4000^{***}\,FM_{it} + \hat{\mu}_{i}^{**} \qquad (7)$$

R^2=87.68%

*,**,***: significant at alpha 10%, 5%, 1%

According to equation (7), the spatial autoregressive coefficient is ignored. Even though the spatial dependence is exist according to the result of LM test, the unemployment disparity of a province is not directly affected by unemployment disparity in the neighboring provinces. LeSage (1999) explained that spatial dependence can exist because laborers are mobile and can cross province to find employment in neighboring province where the unemployment rates measured on the basis of where people live. However, in Indonesia, this condition only valid in Java Island, especially in three regions which are DKI Jakarta, Banten, and Jawa Barat. DKI Jakarta is the capital city of Indonesia and all the economic activities is centralized there. People who live in Banten and Jawa Barat seek for jobs to Jakarta. The government also provides transportation that connects those three provinces so that worker can easily cross province and work in Jakarta. Statistics Indonesia (2014) reported that the result of Jabodetabek Commuter Survey[1] shows that 14.09 percent of 9.2 million people in DKI Jakarta is commuting people with 85.47 percent of it is workers. However, this condition is different with other provinces, especially provinces outside Java Island and provinces with archipelago territory. The transportation is still limited and spend much time to cross the province. It makes those provinces less mobility in short

1 - Jabodetabek is abbreviation of "Jakarta-Bogor-Depok-Tangerang-Bekasi" (cities in DKI Jakarta, Jawa Barat, and Banten provinces)

term. People who cross to other provinces usually choose to stay in destination province in long term instead of commuting. ILO (2012) reported that slightly less than 60 per cent of total per cent of the total Indonesian labor force is concentrated in Java Island. This makes people outside Java prefer to move to Java, instead of their neighboring provinces.

Besides, from equation (7), all the independent variables capturing labor structures statistically significant affect the unemployment disparities at 10 percent significance level. The supply side effect on unemployment disparity is captured by coefficient of participation rate, *PR*, which has positive relationship with unemployment disparity. Every 1 percent increasing in participation rate will lead the unemployment disparity by 0.16 percent. This effect is in line with study of unemployment disparity in Greek region by Lolos and Papapetrou (2003). The higher labor supply will leads to higher unemployment. Furthermore, Layard (1997) explained that the effect of participation rate is expected to be positive, if the participation rate increases, the number of unemployment also increase.

The demand side effects on unemployment disparity are highlighted by the coefficient of *EAGR* and *EIND*. Both coefficients have negative relationship with unemployment disparities. The increasing of employment in agricultural sector by 1 percent will decrease the unemployment disparity by 0.13 percent, while the increasing of employment in manufacturing sector by 1 percent will decrease the unemployment disparity by 0.22 percent. This result shows that labor demand in manufacturing sector effect more in decreasing the unemployment disparity than labor demand in agricultural sector. The unemployment in Indonesia prefers to work in manufacturing sector rather than in agricultural sector. Supriyanti (2010) explained that in Indonesia the agricultural sector is less preferred by educated people and young society. Agricultural sector gives lower wage than other sectors so that this sector is not too promising for their future.

Another finding in this study is the demographic effect on unemployment disparity that represented by *FM*, the ratio of female labor on male labor. Coefficient of *FM* is negative, it means that in every shifting in ratio of female labor forced on male labor forced can decrease the unemployment disparity by 13.4 per cent. In fact, the *FM* value in this study is less than one. Therefore, the decreasing of unemployment disparity will be higher, almost 13.4 per cent, if the value of *FM* is approximately 1. It can be concluded that female participation has stronger effect than male participation on decreasing the unemployment disparity.

The estimation of Spatial Lag Fixed Effect Model also produces the spatial specific effects coefficient that owned by each province. These spatial specific effects represent the omitted variables outside the model that impact unemployment disparity in each province and cause a specific pattern of unemployment disparity for each province. The spatial specific effects indicate the unemployment disparities among provinces that are caused by diversity in local characteristics such as the cultural and social characteristics, local political issues, economic condition, and decentralization policy.

5. CONCLUSIONS

The aim of this study is to examine the existence of spatial dependence of unemployment disparities among provinces in Indonesia during 2004-2014 and its relationship with labor structures. Using Moran's I statistics, it can be concluded that there were spatial autocorrelation of unemployment disparities between provinces in Indonesia especially for the period 2004-2005 and 2008-2012. Furthermore, there were different patterns of unemployment disparities structures during the period 2004-2014.

Other findings in this study come from the empirical analysis of Spatial Lag Fixed Effect Model. The spatial dependence is exist but it does not directly affect the unemployment disparities in each provinces. From labor participation rate, it shows that the increasing in labor participation rates can lead the unemployment disparities. The increasing of participation rate will increase labor supply. The expected effect of this condition is that it will be able to decrease unemployment disparity. However, according to this study, in Indonesia the increasing of labor supply is not followed by optimal labor absorption. Therefore, policy maker should provide more job opportunities effectively that are distributed evenly among provinces and encourage people to create their own job (as entrepreneur).

From the labor demand side, this study shows that the increasing labor demand in both agricultural and manufacturing sectors will decrease the unemployment disparity. Furthermore, the labor demand in manufacturing sector has more effect in decreasing the unemployment disparity than labor demand in agricultural sector. Therefore, improvement in job opportunity for both sectors are crucial and labor absorption of those sectors are also important.

While for the demographic effect on unemployment disparity shows that the increasing in ratio of female labor forced on male labor forced can decrease the unemployment disparity. It means, the participation of female in labor market is important to reduce the unemployment disparity. In Indonesia, the government already gives attention on protecting female labor to participate in labor market by applying Indonesian Labor Law, Act 13 of 2003. However, the participation rate of female labor is still low in Indonesia. Even though there was increasing on employment-to-population ratio of women since 2006, the gender gaps between male and female labor remain large (ILO, 2011). Therefore, it is important for policy makers to enlarge public awareness raising campaigns for female labor.

Another finding in this study is the existence of spatial specific effects that give a positive impact on unemployment disparity. These spatial specific effects are represented by local characteristics such as the cultural and social characteristics, local political issues, economic condition, and decentralization policy.

REFERENCES

Anselin, Luc(EDT). (2003). A Companion to Theoretical Econometrics, Edited by Badi H. Baltagi p 310-330. Oxford: Blackwell Publishing Ltd.

Anselin, Luc. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association, Spatial analytical perspectives on GIS, 111, 111-125.

Baltagi, Badi H. (2005). Econometric Analysis of Panel Data (third ed.). Chichester: John Wiley & Sons Ltd

Cracolici, Maria Francesca; Cuffaro, Miranda; Nijkamp, Peter. (2007). Geographical Distribution of Unemployment: An Analysis of Provincial Differences in Italy, Tinbergen Institute Discussion Paper, No. 07-065/3.

Elhorst, Paul J. (2009). Spatial Panel Data Models. Handbook of Applied Spatial Analysis (Fischer, Manfred M and Arthur Getis) p.377-407. Heidelberg.

Gujarati, Damodar N. (2004). Basic Econometrics [Fourth Edition. McGraw Hill, India.

ILO. (2012). Labor and social trends in Indonesia 2011: Promoting job-rich growth in provinces/International Labor Offi ce. Jakarta: ILO.

ILO. (2013). Labor and social trends in Indonesia 2012: Working for a sustainable and equitable economy/ International Labor Offi ce – Jakarta:ILO.

James P. LeSage. (1999).The Theory and Practice of Spatial Econometrics. Department of Economics University of Toledo.

J. Paul Elhorst. (2003). Specification and Estimation of Spatial Panel Data Models, International Regional Science Review 2003 26: 244. Retrieved from irx.sagepub.com

Lolos, Sarantis and Papapetrou , Evangelia. (2011). Unemployment disparities and persistenceAssessing the evidence from Greek regions, 1981-2008 . Grecce.

Lottmann, Franziska. (2012). Explaining Regional Unemployment Differences In Germany: A Spatial Panel Data Analysis, SFB 649 Discussion Paper 2012-026. Berlin.

Mussida, Chiara and Pastore, Francesco. (2015). Geographical Labor Market Imbalances: Recent Explanations and Cures. Springer-Verlag Berlin Heidelberg.

Mitchell, William and Bill, Anthea. (2004). Spatial dependence in regional unemployment in Australia, Centre of Full Employment and Equity, Working Paper No. 04-1 November 2004. Australia.

OECD. (2005). OECD Employment Outlook. Retrieved from http://www.keepeek.com/Digital-Asset-Management/oecd/employment/oecd-employment-outlook-2005_empl_outlook-2005-en#page73.

Statistics Indonesia of DKI Jakarta Province. (2015). Berita Resmi Statistik BPS Provinsi DKI Jakarta No.12/02/31/Th.XVII, 16 Februari 2015.

Supriyati. (2010). The Dynamic of Agricultural Employment Economy: Problems and Policy Strategy, Agricultural Policy Analysis Volume 8 No.1 March 2010. Jakarta.

Yuniasih, Aisyah Fitri. (2013). Disparitas, Konvergensi, Dan Determinan Produktivitas Tenaga Kerja Regional di Indonesia. Bogor: Sekolah Pascasarjana Institut Pertanian Bogor.

Under-Five Mortality in India: Effects of Neighbourhood Contexts with an Application of Multilevel Cox Proportional Hazard Model

Awdhesh Yadav*
Research Fellow, Public Health Foundation of India (PHFI), Delhi, India
Email: awdhesh.stat@gmail.com

Abstract

Although consideration work has been done to understand the effect of individual level factors on under-five mortality, less is known about the community (neighbourhood) characteristics affect health outcomes for children, even though they have a prominent role in theoretical model. This study address important issues in under-five mortality in India. The objective of this paper is to determine the important of community, mother and individual level effect on under-five mortality in India. Using data from the latest round of Demographic Health Survey (DHS)-2005-06, known as National Family Health Survey (NFHS) in India, multilevel cox proportional hazard analysis was performed on a nationally representative sample of 51, 555 children nested within 36,850 mothers who were also nested within 3842 communities. Hazard ratio (HR) with 95% confidence interval (CI) were used to express measure of association among the characteristics. Variance partition coefficient (VPC) and Wald statistics were used to express measures of variation. The results indicate that pattern of under-five mortality were clustered within mothers and communities. The community level variables like region, place of residence, community poverty level, community education level, ethnic fractionalization index were significantly determine under-five mortality in India. The risk of under-five deaths were significantly higher for children residing in North (HR: 1.34; 95% CI: 1.15-1.60), East (HR: 1.84; 95% CI: 1.59-2.13) and West regions (HR: 1.45; 95% CI: 1.23-1.70) compared to South region. In addition, the proportion of women in community completing secondary school (HR: 0.58; 95% CI: 0.51-0.66) were significantly more likely to increase the child survival. The mother level variables like maternal education, BMI, mother age at birth and breast-feeding were significantly determine under-five mortality. The results suggest to address the contextual level factors to address under-five mortality in India. The findings also suggest the need to focus on community-level intervention aim at improving the socioeconomic conditions of mothers, especially disadvantage regions such as North, East and West.

Keywords: Under-five mortality, Multilevel, Community, India.

Grandparents' Co-residence and Grandchildren's Weight Outcome in China

Qinying He *
College of Economics and Management, South China Agricultural University, Guangzhou, 510642, China –heqy83@scau.edu.cn
Xun Li
College of Economics and Management, Wuhan University, Wuhan, 430071, China–li.xun@whu.edu.cn
Rui Wang
Department of Economics, Tulane University, New Orleans, LA, 70130,USA – rwang4@tulane.edu

Abstract

Childhood obesity in China has been increasingly cited as a major public health issue in recent decades. Although grandparents are more likely to spoil grandchildren by feeding them more food and limiting their physical activity, the effect of grandparents on grandchildren's weight outcome is under-analyzed. Using data from the China Health and Nutrition Survey (CHNS), this study examines the effect of grandparents' co-residence on childhood weight outcome with a sample of 2 to 13 year old children and adolescents in China. Instrumental variable estimation is used to address the endogeneity of grandparents' co-residence. After controlling for a rich set of covariates and unobserved individual heterogeneity, we show that co-residence effect is significantly positive. It is stronger in rural areas and co-residence affects grandchild's weight outcome through changes in dietary patterns and physical activities. Robustness check including an ordered logit model with BMI category as well as estimation with additional data validates our findings.

JEL: I12 I14

Keywords: Obesity; Co-residence of grandparent; Nutrition intake; Physical activity.

1. INTRODUCTION

Childhood overweight and obesity in China has been increasingly cited as a major public health issue in recent years. The prevalence of childhood obesity has increased steadily with the rapid rise in living standards (Zhang et al., 2016). Between 1985 and 2000, the obesity rate among children aged 7-18 years increased fourfold according to the National surveys on the health of school children (China Daily, 2008). Additional, according to a report from China's Working Group on obesity in 2004, 21.7% of 2- to 18-year-old were obese (Shan et al., 2010). These numbers seem to be less serious and shocking compared to the statistics in some developed countries such as United States. However, obesity deserves added attention given the huge population base in China and the fact that childhood obesity is associated with depression, low cognitive ability, and health problems. Even more, childhood health is widely regarded as an important type of human capital in both developing and developed countries, which has significant implications for educational performance and long-term labor market outcomes in future years.

A substantial volume of previous work has focused on possible causes of rapid increase in childhood and adolescent obesity in China including health-related behavior such as physical activity (Li et al., 2010), lack of sleep (Meng, 2012), sugar-sweetened beverage intake (Shang, 2012), as well as social causes such as family income, dietary culture and population policy (e.g., Han and Wen, 2014; Zhang et al. 2015; Yang, 2007; Zhang, Xu and Liu, 2016). However, the effect of grandparents' co-residence on childhood weight outcome is still under-analyzed.

In China, particularly in rural areas, it is a tradition that grandparents live with the children and being the main caregiver for their grandchildren. Their co-residence can affect childhood weight outcomes for the following reasons. First, grandparents, especially in rural area, often consider overweight baby as a healthy baby and a relatively thin baby as a sign of malnutrition, due to the lack of health-related (particularly nutrition-related) knowledge. Therefore, they tend to over-feed their grandchildren. Second, most of the Chinese grandparents have experienced underweight, under-nutrition, food shortage, physical hardship and deprivation in their early lives (Li et al., 2015). They now play a central role in taking care of their grandchildren. They tend to share more food with the children than is necessary as an expression of their love. Third, after the implementation of the one-child policy (OCP), the child, particularly in the one-child family, is more likely to be spoiled with expensive food with high-sugar, high-fat and high-protein content and less exercise. For example, grandparents tend to protect their grandchildren from the risk of injury in some competitive sports, e.g., football games.

The goal of this article is to contribute to the literature through a thorough investigation of the effect of grandparents' co-residence on children's weight outcome. We also test the channels by examining the impact of grandparents' co-residence on child average daily nutrition intake and physical activity. This paper disentangles socioeconomic status by controlling for variables such as parental education and family income to explain a potential relationship between grandparents' presence and grandchildren's weight outcome. Our results show that the grandparents' co-residence increases the weight outcome of children and adolescents aged 2 to 13 years. Our results also reveal that such effect is stronger in rural areas and the effects work through the influence of dietary patterns and physical activities. Robustness checks are also implemented to validate that first, the grandparents' co-residency significantly increase the odd of changing from normal weight to overweight, as well as from overweight to obesity. Second, the children taken care of by grandparents are indeed more likely to have a higher weight outcome.

The rest of this paper is organized as follows. The literature review and background introduction is introduced in Section 2. Section 3 describes our data. Section 4 presents the empirical results as well as the robustness check. We conclude this article in Section 5 with a discussion and remarks.

2. LITERATURE REVIEW AND BACKGROUND INTRODUCTION

2.1 Literature review

The factors that affect the childhood weight outcome are widely analyzed in both China and many Western countries. Generally speaking, the factors can be divided into three groups. The first one is the children's demographic information such as age, gender and education. The second one is the health-related behavior, which can be nutrition intake, physical exercise, and even smoking. The third one is socioeconomic factors such as household income, size, and parents' working status. The first two groups are widely analyzed and the effects in China and Western countries are found similar. For example, Li et al. (2010) believe a 20-min-per-day vigorous physical activity could effectively prevent childhood obesity, and that insufficient physical activity could lead to childhood obesity. Shang (2012) finds a positive correlation between childhood obesity and sugar-sweetened beverage intake. Similarly, Harris et al. (2009) document the importance of physical exercise and balanced of nutrition intake to the U.S. children. The findings regarding the first two groups are considered common knowledge. However, regarding with some socioeconomic factors, the findings in China are different from that in Western countries. For example, Han and Wen (2014) document a positive correlation between family income and childhood obesity using the data from Shanghai. While most papers studying Western countries discover a negative correlation between childhood obesity and low socioeconomic status. In addition, as an emerging economy with deep-rooted Eastern traditions, some factors that affect the childhood outcome are unique in China. For example, in the late 1970s, the Chinese government introduced the controversial one-child policy (OCP)[1], which dramatically reduced the birth rate. The policy is found to alter mothers' care-taking behaviors and their child's dietary habits and nutritional intake (Zhang et al., 2016).

2.2. Grandparents' co-residence and its transition

In China, especially in rural areas, sons are entitled to co-reside with their parents until family division occurs or until both of their parents pass away. Traditional family arrangements are beneficial in Chinese

1 In order to promote modernization by slowing rapid growth of China's population, the Chinese government enforced a population control policy in 1979 which required one family to have no more than two children.

society as they represent the fulfillment of a cultural ideal. However, this culture of co-residence is changed due to several reasons. First, the household income increases after the Reform and Open. Many adult children can afford their own houses or apartment, and at the same time, the gradually developed social pension program support resulted in more and more elderly people not living with their children. Second, after the Reform and Open, more and more young people migrate from rural areas to big cities, which separated many adult children from their aging parents and imposed significant challenges on traditional patterns of familial support for rural older people. Third, the OCP reduced the numbers of the children sharply. More and more parents live alone until they cannot take care of themselves.

The public childcare service is not well developed in China, especially for children who are under 3 in China. The main solution is that grandparents or employed babysitter/nanny (Ayi in Chinese) take care of grandchildren. However, employing an Ayi gives a financial burden on the family. It also often happens that Ayi does not care the baby well. Therefore, the grandparents are the main child carers without employing an Ayi, or the supervisor of Ayi in the caring of the grandchildren in the multi-generational family.

2.3. Two possible channels

Given the framework developed by Cutler et al. (2003) on weight gain through the flow of calories, grandparents' co-residence can impact the grandchildren's weight outcome through two possible channels. First, due to the lack of health-related (particularly nutrition-related) knowledge, they tend to over-feed their grandchildren. Most of the Chinese grandparents have experienced underweight, under-nutrition, food shortage, physical hardship and deprivation in their early lives (Li et al., 2015). They now play a central role in taking care of their grandchildren and tend to over-feed them. Second, the grandparent may decrease the grandchild's physical activity because they are more risk-averse. For example, it is commonly seen that the grandparents hold kids in their arms and would not like them to walk too much. The grandparents generally regard the physical activity such as football game is full of risk. Therefore, in their presence, the level of physical activity could be reduced.

3. METHODOLOGY AND DATA

To estimate the effect of grandparents' co-residence on child weight status, we first specify the following model:

$$y_{it} = x'_{it}\beta + z_{it}\delta + u_{it},$$

where y_{it} is the weight outcome (in kg) for child i in year t, x_{it} denotes the vector of child's demographic information (e.g., height, age and gender), and family information (e.g., parents' education level, household income and household size), and z_{it} is the status of grandparents' co-residence. z_{it}=1 if the grandparents live together with their children at time t for the household that child i belongs to; otherwise, z_{it}=0.

Status of grandparents' co-residence is potentially endogenous. It is possible that the parents who live with adult children and take care of the grandchildren are comparatively healthy than those who do not. So their grandchildren may get better genes from them and thus have a higher weight. To address the possible endogeneity problem which would lead to a biased estimation, we exploit the proportion of households with grandparents' co-residence in community except the household under concern as an instrument[2], which can be reasonable for two reasons. First, the residents usually share the same cultural ideal within a certain region (e.g. community in our case). So the proportion of households with grandparents' co-residence in a given community should closely correlate with grandparents' co-residence of household under concern. Second, the proportion is only plausibly related to the co-residence status while it is not correlated with the weight outcome directly.

The data used in this paper are from China Health and Nutrition Survey (CHNS), an ongoing international collaborative project between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute of Nutrition and Food Safety at the Chinese Center for Disease Control and Prevention. The survey population covers individuals in about 4,400-7,200

2 We limited our sample to the children living at in the community with at least two households.

households from communities with diverse socioeconomic backgrounds in 9-12 provinces and major cities of China. Data from Liaoning, Heilongjiang, Shandong, Henan, Hubei, Hunan, Guangxi, Guizhou, Jiangsu, Chongqing, Beijing, and Shanghai[3] were collected in nine waves during 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009 and 2011. These CHNS data provide detailed economic, demographic and health information. The Chinese provinces and mega cities vary substantially in geography as well as economic development. The CHNS data cover years during the era of rapid economic transition that witnessed considerable children's weight. The survey's multistage random cluster sampling method, which is based on different income levels (high, medium, and low) and weighted sampling, entails the following steps: after randomly selecting four counties and two cities with each province, the CHNS randomly identifies villages and towns in each county and urban and suburban regions in each city. It then selects 20 households from each of these communities. For the purpose of this paper, we use the subsample of CHNS that was collected for the primary school children aged 2-13 at the time of the survey[4]. We do not include older children because puberty starts about age 13, and growth spurts during puberty thereafter can make anthropometric measures less reliable. Another reason is that children older than 13 years of age are likely to have more control over their own food choices and their health are less affected by household composition. In the empirical analysis, all the data are pooled together to identify the effect of grandparents' co-residence.

Importantly for this paper, the CHNS records anthropometrical measurements, height and weight, for each individual. The household income is constructed using information on wages, gardening income, farm/livestock/poultry/fishing income, business income and government support. The family income is deflated to Yuan of year 2011, using the price index provided by the CHNS project team, which adjusts for province-level regional differences in the cost of living and rural-urban differences within each province.

Table 1[5] summarizes the key variables used in this article, including demographic characteristics of the children, father and mother, nutrition intake and physical activity of the children, etc. Parents in the families with co-residence of grandparents have higher education levels and family income. The children in this type of families are also younger, which is consistent with common sense that younger children need grandparents' care and grandparents choose to co-reside with them. Without controlling for other variables, the weight and nutrition intake of children in the families with co-residence of grandparents are significantly lower because they are younger. Child average daily nutrition intake includes two measures: fat intake (in grams) and protein intake (in grams).

4. ESTIMATION RESULTS

Table 2 represents the estimation results for the effect of grandparents' co-residence on childhood weight status with two model specifications. Compared to the first one, the second includes a dummy variable of grandparents' co-residence and its interaction term with the dummy of rural areas. The results show that the parental education level and family income significantly increase the childhood weight outcome. A higher family income may offer more nutritious foods and healthy inputs into a child's growth. As the family income increases, it would offer high-energy and high-fat food to the children. In addition, the child in a family with larger size tends to have a lower weight outcome, which is consistent with the quality-quantity (Q-Q) trade-off theory (e.g., Rosenzweig & Wolpin, 1980; Hanushek, 1992; Parish and Willis, 1993). For both specifications, it is clear that grandparents' co-residence has a significant positive effect on the weight outcome of the children. Compared to the urban area, the effect of grandparents' co-residence is smaller in rural area. Keeping other variables fixed, on average, grandparents' co-residence is estimated to increase the weight about 0.416 kg for child in urban area and only 0.073 kg for child in rural area.

3 Heilongjiang only joined the CHNS in 1997. And three mega cities (Chongqing, Beijing, and Shanghai) were added in the survey of 2011 wave.

4 We do not include the children from 1989 in our sample because the questionnaires and sampling in wave 1989 was substantially different from those used in subsequent waves and the health and nutritional data was collected only from preschoolers and adults aged 20-45 in 1989.

5 Due to the limitation of space, we do not include the tables in this manuscript. The tables can be requested from the authors.

To address the potential endogeneity problem of grandparent's co-residence, we apply the instrument variable (IV) in two steps since the co-residence is binary variable. In the first step, a Probit model is used to predict the probability of the grandparents' co-residence in a family. After variables such as child's age, gender, and maximum years of parents' education are controlled for, the instrumental variable, percentage, is significantly positive. It indicates that a family is more likely to co-reside with grandparents if it is in a community with a higher proportion of co-residence, if the child is a boy, or if it is a big family. In contrast, grandparents are unlikely to live with adult parents if the grandchildren have grown up.

Consistent with the OLS regression results, the IV estimation results suggest that the grandparents' co-residence significantly increases the child weight by 0.548 kg. When we consider there is a difference between urban and rural areas, the estimation results show that child in an urban area would be about 0.824 kg heavier if he/she had at least a grandparent's co-reside with the family. The effect will be weaker for child in rural area, only about 0.3 kg heavier. Compared to the estimation results in Table 2, the IV estimation results show there exist stronger effect of grandparent's co-residence on the childhood weight outcome[6].

Table 4 illustrates the effect of grandparents' co-residence on dietary pattern. After controlling variables such as gender, age, mother's years of education, household income per capita, we find that the grandparent's co-residence significantly increases the fat intake and protein intake of the children. The effect is much lower in a rural area than that in an urban area. One possible reason may be that the average income and food supply is relative lower in rural area.

Table 5 presents the effect of grandparent's co-residence on children's weight outcome through changing the children's physical activity. This measure of physical activity is the indicator of whether a child usually does physical exercises or not. Probit regression estimates the association between grandparent's co-residence and a child's physical activity. As CHNS data on the physical activity are for two difference age groups, children below or above 6 years, our analysis uses these two different age groups. From Table 5, we find that grandparent's co-residence has an insignificant positive effect on non-rural children, but a significantly lower effect on rural children.

In order to better understand whether grandparent's co-residence increases the odds for being overweight and obesity, following Li and Lopez (2016) we also estimated an ordered logit model with body weight categories. Table 6 illustrates the estimation results of the ordered logit model. Based on the standard of WHO, in the first model specification, the dependent variable reflecting weight categories is defined by -3 for grade 3 thinness, -2 for grade 2 thinness, -1 for grade 1 thinness, 0 for normal weight, 1 for overweight, and 2 for obese,. In the second model, the dependent variable reflecting weight categories is defined by 1 for grade 3, grade 2, and grade 1 thinness, 2 for normal weight, 3 for overweight, and 4 for obese. The results for these two model specifications are quite similar. We find that grandparent's co-residence significantly increases the likelihood of children from being overweight when compared to normal, and from being overweight to obesity. The results also indicate that the effects in the non-rural area are higher than in the rural area. One attractive feature of the ordered logit model is that one can use them to predict the probabilities of each weight status that an individual belongs to. For example, on average, the probability that a child is of normal weight is 0.68, and the probabilities that a child is overweight or obese are 0.08 and 0.03, respectively. Similarly, one can obtain counterfactual results of the marginal probability of an individual changing weight status due to changes in the co-residence status of grandparents.

Another concern is that the grandparents may just live with the adult children, and not take part in taking care of the grandchildren. To address this concern, we take advantage of the CHNS data, which records information of the location of care for children under 6. It can be at parents' home, or at grandparents' home. We analyze the weight outcome if the children are taken care of only in grandparents' home, with the underlying assumption that the children taken care by grandparents are only affected by their grandparents. The results are given in Table 7. Column (1) shows that the impact is 0.17 and significant if the child is taken care of only by grandparents, indicating that the children taken care of by grandparents are indeed more likely to have a higher weight outcome. Column (2) and Column (3) illustrate the

6 An over-identification test cannot be implemented since the number of instrumental variables equals to the number of endogenous variables (i.e. one instrumental variable).

effects from the paternal grandparents and maternal grandparents. In China, gender discrimination still exists. Therefore, many older people regard a married daughter to belong to her husband's family. She will be a guest when she comes back to visit her parents. Similarly, their daughter's children are also no longer in this family tree because the family name of children follows that of their father. As a result, in many provinces, particularly in rural area, paternal grandparents love the children much more than the maternal grandparents. Our results are also consistent with this practice. We can see that the children living with paternal grandparents are likely to get higher weight outcome, although the estimates are insignificant.

5. CONCLUDING REMARKS

This article estimates the effect of grandparents' co-residence on the children's weight outcomes, and it reveals the channels by which the weight outcome is affected. Using data from the China Health and Nutrition Survey (CHNS), this study analyzes the effect of grandparents' co-residence on childhood weight outcome with a sample of 2- to 18-year-old children and adolescents in China. The instrumental variables estimation is used to address the endogeneity of grandparents' co-residence.

The empirical analysis validates that the grandparents' co-residence has a significant positive on children's weight outcome. The effects are stronger in rural areas than in non-rural areas. This article also provides evidence that the grandparents' co-residence impacts children's weight outcome mainly through nutrition overfeeding such as fat and protein. The co-residence impacts the level of physical activity significantly only in rural areas. The robustness checks validate our conclusions and strengthen our arguments. First, the grandparents' co-residency significantly increases the odds of changing from normal weight to overweight, as well as from overweight to obesity. Second, the children taken care of only by grandparents are indeed more likely to have a higher weight outcome.

These findings extend the existing literature on childhood obesity in developing countries, such as Zhang et al (2016). Beyond the demographic characteristics and socioeconomic factors, this article provides new evidence that grandparents' co-residence is another important factor that matters. Further results also validate the channels through which it exerts its effect. Additionally, this article provides policy implications that future preventive intervention should include strategies that target grandparents, e.g., provide some training program to increase their knowledge of nutrition and health.

REFERENCES

Brauw, A. D., & Mu, R. (2011). Migration and the overweight and underweight status of children in rural china. *Food Policy, 36*(1), 88-100.

China Daily, (2008), "Childhood obesity a heavy burden for parents". Available at http://www.chinadaily.com.cn/china/2008-12/03/content_7265415.htm

Cutler, D. M., & Shapiro, J. M. (2003). Why have americans become more obese?. *Journal of Economic Perspectives, 17*(3), 93-118.

Goode, A., Mavromaras, K., & Zhu, R. (2014). Family income and child health in china ☆. *China Economic Review, 29*(C), 152-165.

Han, W., & Wen, J. (2014). The survey of China children's physical & psychological health, 2014 Shanghai. *East China Normal University & New York University Shanghai Report.*

Hanushek, E. A. (1992). The tradeoff between child quantity and child quality. *Journal of Political Economy, 100*(1), 84-117.

Harris, J., Bargh, J.A. and Brownell, K. D. (2009). Priming Effects of Television Food Advertising on Eating Behavior. *Health Psychology: Official Journal of The Division of Health Psychology*, 28(4): 404-413.

Li, B., Adab, P., & Cheng, K. K. (2015). The role of grandparents in childhood obesity in China-evidence from a mixed methods study. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 1.

Li, X., & Lopez, R. A. (2016). Food environment and weight outcomes: a stochastic frontier approach. *Applied Economics*, 1-12.

Li, Y., Hu, X., Schouten, E. G., Liu, A., Du, S., Li, L., et al. (2010). Report on childhood obesity in China: Effects and sustainability of physical activity intervention on body composition of Chinese youth. *Biomedical and Environmental Sciences*, 23, 180–187.

Luo, J., & Hu, F. B. (2002). Time trends of obesity in pre-school children in China from 1989 to 1997. *International Journal of Obesity*, 26(4), 553-558.

McKinney, C.O.(2015). Early Childhood Overweight and Obesity in Multigenerational Households. *GrandFamilies: The Contemporary Journal of Research, Practice and Policy*, *2*(1), 2.

Meng, Li, Ping, et al. (2012). Report on Childhood Obesity in China (10): Association of Sleep Duration with Obesity. *Biomedical and Environmental Sciences*, 25(2), 133-140.

Parish, W. L., & Willis, R. J. (1993). Daughters, education, and family budgets taiwan experiences. *Journal of Human Resources, 28*(4), 863-898.

Rosenzweig, M. R., & Wolpin, K. I. (1980). Testing the quantity-quality fertility model: the use of twins as a natural experiment. *Econometrica Journal of the Econometric Society, 48*(1), 227-240.

Shan, X.Y., Xi, B., Cheng, H., Hou, D.Q., Wang, Y.F., and Mi, J. (2010), "Prevalence and behavioral risk factors of overweight and obesity among children aged 2–18 in Beijing, China", *International Journal of Pediatric Obesity*, 5(5): 383-389.

Shang, Xian, Wen, et al. (2012). Report on Childhood Obesity in China (9): Sugar-sweetened Beverages Consumption and Obesity. *Biomedical and Environmental Sciences*, 25(2), 125-132.

Yang, J. (2007). China's one-child policy and overweight children in the 1990s. *Social Science & Medicine*, 64(10), 2043–2057

Zhang, J., Wang, H., Wang, Y., Xue, H., Wang, Z., Du, W., and Zhang, B. (2015). Dietary patterns and their associations with childhood obesity in China. *British Journal of Nutrition*, 113(12), 1978–1984.

Zhang, J., Xu, P., & Liu, F. (2016). One-child policy and childhood obesity. *China Economic Review.* Forthcoming.

Modelling Work Status of Housewife Using Multilevel Model For Binary Response

Wahyu Wibowo*
Department of Business Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – wahyu_w@statistika.its.ac.id
R. Mohamad Atok
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – radenatok@gmail.com
Fanny Ayu Octaviana
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – fannyayuoct@gmail.com

Abstract

The lack of revenue generated by husband as head of household and breadwinner make most women opt to work in order to meet needs of the family. In this study, multilevel modelling will be conducted on work status of women based on the latest education, age and household expenditure. This study aims to determine the relationship between work status of women based on variables at each level and an appropriate model of research on work status of women. The data used is secondary data obtained from the Indonesia Family Life Survey (IFLS) in 2014, i.e. the total of data is 4431 women who have ever been married or married status and as many as 99 data is the average salary of women workes that corresponds to data in the sub-district. Multilevel model used in this study is a 2 levels regression model for binary response. The variables of this research is work status of women as the response variable (Y), with two categories: 0 to women did not work and 1 to women work, and the latest education, age and household expenditure as a predictor variable (X) at level 1 (individual) and the average salary of women workers as a predictor variable (Z) at level 2 (sub-district). The first step in analysing the data with multilevel modelling is modelling with binary logistic regression at level 1 (individual), then do the model with 2 levels regression model for binary response to determine the effects of their level 2 (sub-district). From the analysis, it can be concluded that the best model for work status of housewife case in Indonesia are using a 2 levels regression model for binary response with the latest education and household expenditure variables that influence work status of housewife because the generated of deviance value is most small compared to other models. That is, the housewife's decision to participate in the work world is also influenced by the area of residence (level 2/sub-district).

Keywords: Work status of housewife; Binary logistic regression; Multilevel model.

1. INTRODUCTION

According to the data of National Employment Survey (Sakernas) in 2013, the number of women who work as laborers/employees in Indonesia from 1980 reached 32.65 percent tend to growing up to 52.36 percent in 2013. The data also supported by increasing population in Indonesia from year to year. Several reasons or factors that cause women prefer to work, such as family income, especially if the revenue generated by husband of a relatively small, and various needs such as education, skills, and need to self development. Linandar (2009) stated that women may also need to show that they capable of achievement in public life, and to obtain status or more power in family life.

In various cases, such as social and educational, often encountered the population data a have hierarchical structure. Data structure is usually derived from stratified or tiered (hierarchy) population and in groups (clusters). According to Hox (2002), the data have a hierarchical structure, there is actually an effect of the hierarchy level to the observations unit. Furthermore, Goldstein (1995) suggested that development of ordinary regression to overcome the issues from a hierarchical data structure that is Multilevel Modeling analysis. Multilevel model is part of Generalized Linear Mixed Model (GLMM), that is the response variable influenced by fixed and random factors.Levels of a hierarchical structure in multilevel model is defined as a level. The simplest multilevel model is a model with two level, i.e there are only two level on a hierarchical data structure. The lowest level is individual called level 1 and the higher level is District called level 2.

Research of GLMM application have been done previously. Fong, Rue, and Wakefield (2009) used Bayesian inference for generalized linear mixed models and applied in longitudinal data with integrated nested Laplace approach. Khan and Shaw (2011) applied Multilevel logistic regression to analyze binary contraceptive prevalence data from health surveys in Bangladesh using 3-levels model with Penalized Quasi-Likelihood approach. Multilevel modeling for binary data was also applied on education data in the United States using 2-levels model as in Agresti (2007). This paper aims to study the relationship between work status of housewife based on the latest education, age and household expenditure using 2-levels model.

2. MODEL

If the response variable follows normal distribution with proportion parameter (π_{ij}), then link function used was logit ($\log\{\pi/(1-\pi)\}$) so that the model is called the logistic model.

In general, 2 levels model with binary response can be written as follows [6]

$$log\left[\frac{\pi_{ij}}{(1-\pi_{ij})}\right] = \beta_{0j} + \beta_1 X_i + \varepsilon_i \text{ (level 1 model)} \tag{1}$$

and

$$\beta_{0j} = \beta_0 + u_j \text{ (level 1 model)} \tag{2}$$

In general, the mathematical formula for random-intercept models with 2 levels for binary response as follows

$$log\left[\frac{\pi_{ij}}{(1-\pi_{ij})}\right] = \beta_0 + \beta_1 X_i + u_j + \varepsilon_i \tag{3}$$

Where u_j an error on level 2 follows normal distribution with average 0 and variance $\sigma_u^{\,2}$ and independent of ε_i.

In multilevel regression model, there is a method used to determine the suitability models, i.e Generalized Linear Mixed Model (GLMM) [6], where the response variable influenced by fixed and random factors. The general form of GLMM as follows

$$y = \boldsymbol{X\beta} + u + \varepsilon \tag{4}$$

Parameter estimation in logistic regression was conducted using Maximum Likelihood Estimation (MLE). If x_i and y_i is a pair of independent and dependent variables in the i-th observation and it is assumed that each pair of mutually independent observations with a pair of other observations, i = 1, 2, ..., n then the probability function for each pair as follows

$$f(Y_i) = \pi(\boldsymbol{X}_i)^{Y_i}(1-\pi(\boldsymbol{X}_i))^{1-Y_i};\ \text{Yi} = 0, 1 \tag{5}$$

In a multilevel model, the numerical integration procedure used is Gauss-Hermite Quadrature procedure to calculate the integral numerically by following equation.

$$\int_{-\infty}^{\infty} (g(v))\ exp(-v^2)\ dv \approx \sum_{k=1}^{m} w_k g(v_k) \tag{6}$$

where :

w_k : Weight quadrature

v_k : Point quadrature

The derived results did not close form because the value of β is a scalar, so to find the estimated value of β used numerical iteration. Maximum Likelihood Estimation methods require a good initial values for parameters. The parameters for non-linear models estimated using Generalized Least Square (GLS) which is used as the initial value of regression coefficient [3]. The next iteration calculation using Newton Raphson Iteration.

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - H^{-1}(\boldsymbol{\beta}_m)\boldsymbol{g}(\boldsymbol{\beta}_m) \tag{7}$$

Iterations stop if

$$\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\| < \varepsilon$$

ε is a very small positive number.

3. DATA

The data used is secondary data obtained from Indonesia Family Life Survey (IFLS) in 2014, which is the total of data is 4431 women who have never married or married status and as many as 99 data is the average salary of women labourers that corresponds to data in the District.

Tabel 1. Research Variable

Variable	Indicator	Information	Scale
Y	Work Status of Housewife	0: Not work 1: Work	Nominal
Level 1 (Individual)			
X_1	The Latest Education	0 : No school 1 : Elementary school 2 : Junior high school 3 : Senior high school 4 : College	Ordinal
X_2	Age (year)	-	Ratio
X_3	Household Expenditure (Rp)	-	Ratio
Level 2 (District)			
Z	Average Salary Of Women Laborers (Rp/Months)	-	Ratio

4. RESULT

The method used to determine the suitability models in multilevel regression model is Generalized Linear Mixed Model (GLMM) as in Agresti (2007). In general, 2 levels regression model with random-intercept as follows

$$Y_{ij} = \boldsymbol{\beta X_{ij}} + u_j + \varepsilon_i \tag{8}$$

Where u_j an error on level 2 follows normal distribution with average 0 and variance σ_u^2 and independent of ε_i.

If Y variable follows Bernoulli distribution for each single observation, so probability function for each observation is given as follows.

$$f(Y_i) = (\pi(\mathbf{X}_i))^{Y_i}(1-\pi(\mathbf{X}_i))^{1-Y_i};\ Y_i = 0,1 \tag{9}$$

Suppose n samples taken randomly within a population ($Y_{ij}, X_{1ij}, X_{2ij}, ..., X_{kij}$; i = 1, 2, ..., n ; j = 1, 2, ..., m) where $Y_{ij} \sim B(1,\pi(\mathbf{X}_{ij}))$ is an independent and $\mathbf{X}_{ij} = [1\ X_{1ij}\ X_{2ij} \ldots X_{kij}]^T$. Next is making likelihood function of Y as follows.

$$L = f(Y_{1ij}, Y_{2ij}, \dots, Y_{nij})$$
$$L = f(Y_{1ij}), f(Y_{2ij}), \dots, f(Y_{nij})$$
$$L = \prod_{i=1}^{n} f(Y_{ij})$$
$$L = \prod_{i=1}^{n} \left(\pi(\mathbf{X}_i)\right)^{Y_i}\left(1-\pi(\mathbf{X}_i)\right)^{1-Y_i}$$
$$L = \prod_{i=1}^{n} \left(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_j\right)^{Y_{ij}}\left(1-(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_j)\right)^{1-Y_{ij}} \tag{10}$$

From equation (9), it is known that u_j is normal distribution with average 0 and variance σ_u^2, and $\boldsymbol{u}_j$ is random effect of 2 levels so that likelihood function to be integrated as follows.

$$L = \int_{-\infty}^{\infty} \prod_{i=1}^{n} f(Y_{ij}) f(\boldsymbol{u}_j) d\boldsymbol{u}_j$$
$$L = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} \left(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_j\right)^{Y_{ij}}\left(1-(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{u}_j)\right)^{1-Y_{ij}}\right] \frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left(-\frac{1}{2\sigma_u^2}\boldsymbol{u}_j^2\right) d\boldsymbol{u}_j$$
$$L = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} f\left(Y_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{u}_j\right)\right] \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}\boldsymbol{u}_j^2\right) d\boldsymbol{u}_j \tag{11}$$

Let $u = \sqrt{2\sigma_u^2}\, v = \theta v$, where $v = \dfrac{\boldsymbol{u}}{\sqrt{2\sigma_u^2}}$ and given $g(v) = \prod_{i=1}^{n} f(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{\beta}, \boldsymbol{u}_j)$ in order to obtain likelihood function as follows

$$\mathrm{L} = \int_{-\infty}^{\infty} (g(v)) \frac{1}{\theta\sqrt{\pi}}\, exp(-v^2)\, \theta dv$$
$$\mathrm{L} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (g(v))\ exp(-v^2)\, dv \tag{12}$$

From equation (12), can be solved using numerical methods are integral approach quadratic Gauss-Hermite Quadrature as follows.

$$\int_{-\infty}^{\infty} (g(v))\, exp(-v^2)\, dv \approx \sum_{k=1}^{m} w_k g(v_k)$$

where :

w_k : Weight quadrature

v_k : Point quadrature

The next step is to maximize *ln* L with lowering *ln* function of likelihood function of β and the result is zero.

The derived results did not close form because the value of β is a scalar, so that initial value of β searched by using the formula GLS (Generalized Least Square) as follows.

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} \tag{13}$$

Furthermore, the value of β searched using numerical iteration namely Newton Raphson iteration with the following formula.

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m - H^{-1}(\boldsymbol{\beta}_m)\boldsymbol{g}(\boldsymbol{\beta}_m) \qquad (14)$$

Iterations stop if

$$\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\| < \varepsilon$$

ε is a very small positive number.

Multilevel regression analyzes were used in this study is 2-levels regression analysis. 2-level regression analysis with binary response, namely with random intercept.

A. 2-Level Regression Analysis With Binary Response Without Average Salary Of Women Laborers Variable

Hypothesis :

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0 \quad ; i = 1, 2, 3$

Statistic test :

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Significance level : $\alpha = 0{,}05$

Critical areas : Reject H_0 if $Z > Z_{\alpha/2}$ or P-value $<\alpha$

Tabel 2. Parameter Estimation Test of Multilevel Regression Analysis Without Z

Variable	Estimation	Z	P-value
$X_1(0)$	0,102	0,675	0,500
$X_1(1)$	-0,076	-0,456	0,649
$X_1(2)$	-0,106	-0,621	0,534
$X_1(3)$	1,103	5,358	0,000
X_2	0,005	1,732	0,083 *
X_3	0,040	3,123	0,002
Constant	0,029	0,128	0,898

*not significant at α = 0.05

The value of $Z = 1.732 < Z_{\alpha/2} = 1.96$ or P-value = 0.083 > α = 0.05, thus reject H_0. That is, age variable (X_2) does not affect to work status of housewife. So, there should be 2-levels regression analysis without age variable (X_2).

Tabel 3. Parameter Estimation Test of Multilevel Regression Analysis Without Z and X_2

Variable	Estimation	Z	P-value
$X_1(0)$	0.033	0.230	0.818
$X_1(1)$	-0.196	-1.289	0.197
$X_1(2)$	-0.235	-1.530	0.126
$X_1(3)$	0.974	5.077	0.000
X_3	0.040	3.169	0.002
Constant	0.330	2.281	0.023

The lastest education (X_1) and household expenditure variable (X_3) is influence to work status of housewife. Thus, 2 levels regression model to work status of housewife as follows.

$$ln\left(\frac{\pi(X_{ij})}{1-\pi(X_{ij})}\right)=0.330+0.033X_{1i}\,(0)-0.196X_{1i}\,(1)-0.235X_{1i}\,(2)+0.974X_{1i}\,(3)+0.040X_{3i}$$

The result of 2-levels regression analysis from work status of housewife obtain Deviance value is 5792.2.

B. 2-Level Regression Analysis Binary Response With Average Salary Of Women Laborers

Hypothesis :

$H_0 : \beta_{i,j} = 0$

$H_1 : \beta_{i,j} \neq 0$; i = 1, 2, 3 dan j = 1

Statistic test :

$$Z = \frac{\hat{\beta}_{i,j}}{SE(\hat{\beta}_{i,j})}$$

Significance level : α = 0,05

Critical areas : Reject H_0 if $Z > Z_{\alpha/2}$ or P-value $<\alpha$

Tabel 4. Parameter Estimation Test of Multilevel Regression Analysis With Z

Variable	Estimation	Z	P-value
$X_1(0)$	0.103	0.684	0.494
$X_1(1)$	-0.075	-0.447	0.655
$X_1(2)$	-0.104	-0.608	0.543
$X_1(3)$	1.107	5.380	0.000
X_2	0.005	1.766	0.077 *
X_3	0.040	3.120	0.002
Z	-0.032	-1.031	0.303 *
Constant	0.077	0.333	0.739

*not significant at α = 0.05

The value of Z = 1.766 or P-value = 0.077 for X_2 and $|Z|$ = 1.031 or P-value = 0.303 for Z variable, each $Z < Z_{\alpha/2}$ = 1.96 or P-value > α = 0.05, thus reject H_0. The lastest education (X_1) and household expenditure (X_3) variable is influence to work status of housewife. Thus, 2 levels regression model to work status of housewife as follows.

$$ln\left(\frac{\pi(X_{ij})}{1-\pi(X_{ij})}\right)=0.077+0.103X_{1i}\,(0)-0.075X_{1i}\,(1)-0.104X_{1i}\,(2)+1.107X_{1i}\,(3)+0{,}040X_{3i}$$

The result of 2 levels regression analysis from work status of housewife obtain Deviance value is 5788.

C. Selection of the Best Model

Selection of the best model is done by comparing the Deviance value of each model. The results can be seen as follows.

Tabel 5. The Deviance Value Of Each Model

Model	Nilai *Deviance*
Binary logistic regression	5879,9
2 levels regression with random-intercept (without Z variabel)	5792,2
2 levels regression with random-intercept (without variabel Z)	5788

2 levels regression model with random intercept (with variable Z) is the best model because the Deviance value produced the smallest among the values Deviance from other models, although the independent variable (Z) is not significant. The results of 2 levels regression model with binary response as follows.

- Level 1 Model

$$ln\left(\frac{\pi(X_{ij})}{1-\pi(X_{ij})}\right) = \beta_{0j} + \beta_1 X_{1i}(0)+\beta_1 X_{1i}(1)+\beta_1 X_{1i}(2)+\beta_1 X_{1i}(3)+\beta_3 X_{3i}$$

- Level 2 Model

$$\beta_{oj} = \beta_{00} + \beta_{01} Z_j$$

Thus, the combination model is obtained between level 1 and level 2 which is formed can be seen as follows.

$$ln\left(\frac{\pi(X_{ij})}{1-\pi(X_{ij})}\right) = (\beta_{00}+\beta_{01} Z_j)+\beta_1 X_{1i}(0)+\beta_1 X_{1i}(1)+\beta_1 X_{1i}(2)+\beta_1 X_{1i}(3)+\beta_3 X_{3i}$$

$$ln\left(\frac{\pi(X_{ij})}{1-\pi(X_{ij})}\right) = 0.077-0.032Z_j+0.103X_{1i}(0)-0.075X_{1i}(1)-0.104X_{1i}(2)+1.107X_{1i}(3)+0,040X_{3i}$$

Housewives who have no background school and housewives who have the past educational background of high school, as well as household expenditure has increased housewife decision to participate in the work.

5. CONCLUSIONS

The best model for work status of housewife is 2 levels regression model with random-intercept (with Z variable) because the Deviance value is the smallest, equal to 5788. That is, housewife's decision to participate in the world of work is also influenced by their area of residence (Level 2 / District).

Based on the analysis and conclusions, the suggestions are as follows.

1. Modeling work status housewife with a random coefficient models or models with interaction (Cross-interactive model) should be developed to overcome problem more complicated of multilevel case is with interaction between independent variables.
2. The research of 3 levels regression model is level 1 on individual level, level 2 at district level and level 3 at district/city level need to be developed to determine factors that are supposed to influence work status of housewife in the district/city level.

REFERENCES

Agresti, A. (2007). Categorical Data Analysis Second Edition. John Willey & Sons. New York.

Fong, Y., Rue, H. and Wakefield, J. (2009). Bayesian inference for generalized linear mixed models. Biostatistics. Published by Oxford University Press.

Goldstein, H. (1995). Multilevel Statistical Models 2nd Ed. London : E-Book of Arnold.

Guo, G. dan Zhao, H. (2000). Multilevel Modeling for Binary Data. Annual Review of Sociology, Vol 26, pp 441-462.

Hox, J. J. (2002). Multilevel Analysis : Techniques and Applications. London : Lawrence Erlbaum Associates Publishers.

Khan, M. H. R. and Shaw, J. E. H. (2011). Multilevel Logistic Regression Analysis Applied to Binary Contraceptive Prevalence Data. Journal of Data Science 9, pp 93-110.

Linandar, T. N. (2009). Factors Affecting Women's Careers (Case Study: Agency for Community Empowerment and Family Planning Bogor City). Essay. Bogor : Faculty of Economics and Management. Institut Pertanian Bogor.

This page intentionally left blank

CPS11:
Macrofinancial Statistics (1)

The Use of Financial Stress Index in measuring Indonesia's Financial System Stability

Arisyi Fariza Raz*
Bank Indonesia, Jakarta, Indonesia – arisyi_fr@bi.go.id
Ina Nurmalia Kurniati
Bank Indonesia, Jakarta, Indonesia – inanurmalia@bi.go.id

Abstract

This paper estimates Financial Stress Index (FSI) that consists of stock market, bond market and currency exchange market as its main components. The study also conducts a qualitative and quantitative analyses to check the robustness of the estimated index. Finally, the study applies Markov Switching Autoregressive and threshold regression model to determine the stress period and the normal period. The outcome of this study suggests that FSI can be used as a complementary indicator to measure financial stability.

Keywords: Financial system stability, Systemic risk, Macro-financial linkages.
JEL Classifications: G01, G10, G20, E44

1. INTRODUCTION

2008 Global Financial Crisis showed that systemic financial stress could lead a devastating financial crisis, which eventually affected the real economy and caused economic recession. To examine the importance of systemic risk, IMF-BIS-FSB (2009) attempts the define systemic risk as "a risk of disruption to financial services that is (i) caused by an impairment of all or parts of the financial system and (ii) has the potential to have serious negative consequences for the real economy."

Based on this definition, a number of literature tries to measure the systemic level of a financial system by constructing a financial stress index. The earliest attempt to create a financial stress indicator was made by Illing and Liu (2006) by aggregating the movements and volatilities of various components in the financial system, including capital market, money market and currency exchange market. This approach is later followed by more recent studies (for instance, see Cambón and Estévez, 2016; Duprey et al., 2015; Hollo et al., 2012). Slightly different approaches are carried out by Albulescu (2008) and Karanovic and Karanovic (2015), which have broader perspective in measuring financial stability (e.g. by taking into account global economic climate), and Archarya et al. (2010) and Adrian and Brunnermeier (2011), which focus on the systemic impact of individual financial institution instead of measuring the stability of the whole system.

Given the existing literature, this study attempts to create an FSI for Indonesian financial system. This FSI is expected to be an intuitive measurement for financial system stability with high-frequency data that can complement current financial data and other financial indicators with relatively lower frequency data. The rest of the study is structured as follows. Section 2 describes the methodology used to construct the FSI in Indonesia. Section 3 presents the calculated index, robustness check and threshold calculation to determine stable and stress periods. Section 4 concludes.

2. METHODOLOGY

Following Duprey et al. (2015), FSI consists of three main components, i.e. stock market index (*STX*), bond market index (*R*10) and currency exchange market index (*EER*). The calculation details of each index are presented in the following subsections. The calculation of FSI requires data with daily and monthly frequency. Daily Jakarta Composite Index data are required to calculate the stock market index and daily Indonesia 10-year government bond yield are used to calculate the bond market index. Meanwhile, monthly real effective exchange rate (REER) data are used to calculate the currency exchange market index. All data are gathered form the Bloomberg Terminal over the period January 2005-July 2016.

In order to control the price fluctuations, *STX* and *R*10 are adjusted using the consumer price index (CPI), as follows:

$$\begin{cases} rSTX_t = \frac{STX_t}{CPI_t} \\ rR10_t = R10_t - \frac{CPI_t - CPI_{t-261}}{CPI_{t-261}} \cdot 100 \end{cases} \tag{1}$$

2.1. Stock Market Index

Stock market stress index is captured by two variables: (i) the monthly volatility (*VSTX*) which is calculated based on monthly average return of stock market index in its logarithmic form; and (ii) the cumulated maximum loss (*CMAX*) that is calculated based on the maximum loss compared to the highest stock market return over two years. It is computed over a rolling window of two years. The equation is presented as follows:

$$\begin{cases} \ln STX_t = \log(rSTX_{t-i}) - \log(rSTX_{t-1-i}) \\ \ln \widetilde{STX}_t = \frac{\ln STX_t}{\sigma_{\ln STX_{t,t-2609}}} \\ VSTX_t = \frac{\sum_{i=0}^{19} |\ln STX_t|}{20} \\ CMAX_t = 1 - \frac{rSTX_t}{\max_{i=0}^{521}(rSTX_{t-1})} \end{cases} \tag{2}$$

2.2. Bond Market Index

Bond market stress index also follows similar approach. The first variable is monthly volatility (*VR*10) which is calculated based on monthly average return of 10-year government bond yield. The second variable is the cumulative difference of the maximum increase of the real government bond (*CMIN*) that is compared to the minimum yield based on two years rolling window data. The variables are calculated as follows:

$$\begin{cases} chR10_t = rR10_t - rR10_{t-1} \\ \widetilde{chR10}_t = \frac{chR10_t}{\sigma_{\ln chR10_{t,t-2609}}} \\ VR10_t = \frac{\sum_{i=0}^{19} |\widetilde{chR10}_t|}{20} \\ CDIFF_t = \frac{(100 + rR10_t)}{\min_{i=0}^{521}(100 + rR10_{t-1})} \end{cases} \tag{3}$$

2.3. Currency Exchange Market Index

Slightly different from the previous indices, currency exchange market index uses monthly data despite its similar calculation approach. The first variable is exchange rate volatility (*VEER*) that is calculated based on the absolute value of monthly REER growth. The second variable is the cumulative change (*CUMUL*) of the REER's absolute changing rate over six months. Both indices are calculated as follows:

$$\begin{cases} \ln EER_t = \log(rEER_{t-i}) - \log(rEER_{t-1-i}) \\ \ln \widetilde{EER}_t = \frac{\ln EER_t}{\sigma_{\ln EER_{t,t-119}}} \\ VEER_t = |\ln EER_t| \\ CUMUL_t = |rEER_t - rEER_{t-6}| \end{cases} \tag{4}$$

2.4. Aggregation of Individual Indices

After calculating individual indices, we standardize the data into a common unit. Following Hollo et al. (2012) and Duprey et al. (2015), we use the empirical cumulative density function (CDF) computed over an initial window of 10 years, which then expanded to take new data into account:

$$\hat{z}_t = F_n(z_t < z) = \begin{cases} \frac{r}{n} \text{ untuk } z_{[r]} < z_t < z_{[r+1]}, r = 1,2,\dots,n-1 \\ 1 \text{ untuk } z_t > z_{[n]} \end{cases} \tag{5}$$

where $\hat{z}_t \in \{VSTX, CMAX, VR10, CMIN, VEER, CUMUL\}$.

This empirical CDF $F_n(z_t < z)$ transforms each variable into percentiles, where the smallest (highest) values correspond to the lowest (highest) levels of stress. Following the standardization, each variable is aggregated using simple average, which yields the following indices:

$$\begin{cases} I_{STX} = \frac{\widehat{VSTX}+\widehat{CMAX}}{2} \\ I_{R10} = \frac{\widehat{VR10}+\widehat{CMIN}}{2} \\ I_{EER} = \frac{\widehat{VEER}+\widehat{CUMUL}}{2} \end{cases} \tag{6}$$

Then, these three indices are aggregated further to create a single indicator. Following Hollo et al. (2012), a vector of indices; $I_t = (I_{STX}, I_{R10}, I_{EER})$ are multiplied by vector weights[1]; $w_t = (w_{STX}, w_{R10}, w_{EER})$ using Hadamard matrix multiplication approach; (w ¤ I_t). Subsequently, following Hollo et al. (2012) and Duprey et al. (2015), the aggregation is carried out by calculating the time-varying cross-correlations ρ_{ijt} between the indices i and j at period t. To calculate the time-varying cross-correlations, we use the Exponentially Weighted Moving Average (EWMA) with smoothing parameter $\lambda = 0.94$[2], thus yielding the following equations:

$$\begin{aligned} \sigma_{ij,t} &= \lambda\sigma_{ij,t-1} + (1-\lambda)\bar{s}_{i,t}\bar{s}_{j,t} \\ \sigma_{i,t}^2 &= \lambda\sigma_{i,t-1}^2 + (1-\lambda)\bar{s}_{i,t}^2 \\ \rho_{ij,t} &= \frac{\sigma_{ij,t}}{\sigma_{i,t}\sigma_{j,t}} \end{aligned} \tag{7}$$

where, $i, j = \{STX, R10, EER\}, i \neq j$. $\sigma_{ij,t}$ is covariance, $\sigma_{i,t}^2$ is volatility and $\bar{s}_{i,t} = I_{i,t} - 0.5$ is the demeaned indices from its theoretical mean (see Hollo et al., 2012; Duprey et al., 2015).

ρ_{ijt} is arranged to create a matrix C_t with a time-varying dimension of 3×3, consisting the cross-correlations of the indices:

$$C_t = \begin{bmatrix} 1 & \rho_{STX,R10,t} & \rho_{STX,EER,t} \\ \rho_{STX,R10,t} & 1 & \rho_{R10,EER,t} \\ \rho_{STX,EER,t} & \rho_{R10,EER,t} & 1 \end{bmatrix} \tag{8}$$

Then, matrix C_t is multiplied by the weighted matrices of each financial market index, thus as follows:

$$FSI_t = (w \circ I_t) \cdot C_t \cdot (w \circ I_t)' \tag{9}$$

Where $(w \circ I_t)$ is operated using Hadamard matrix multiplication and $(w \circ I_t)'$ is the transposed matrix $(w \circ I_t)$.

3. RESULTS

3.1. Calculated FSI

The calculated indices are exhibited in Figure 1. The standardization of the indices into a common unit limits the range movement of the indices between 0 and 1; (0,1], where 0 represents low financial stress and 1 represents high financial stress. Figure 1 shows that, despite the fluctuating movements, there were periods when all of the three indices were moving in the same direction such as during the 2008 Global Financial Crisis, 2013 The Fed's Tapering Tantrum and 2015 The FFR Normalization.

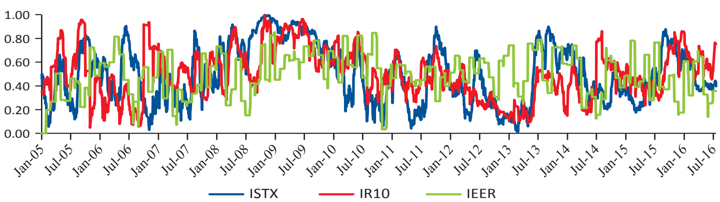


Figure 1. FSI's Individual Indices

1 For simplification, we use equal-weight for each indicator. Similar approach is also carried out by Hollo et al. (2012).
2 Based on a widely-accepted figure, for instance see Hollo et al. (2012).

Figure 3 shows the time-varying cross correlation of the three indices, which represent stock market, bond market and currency exchange market. In general, it supports the previous figure by showing high cross-correlations during the 2008 Global Financial Crisis.

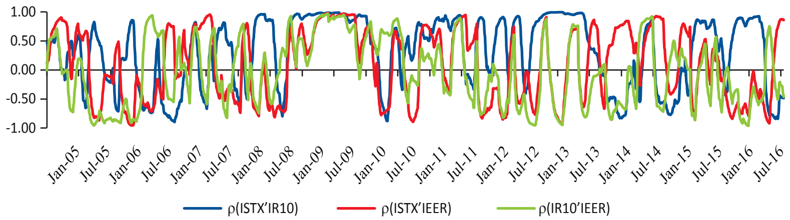


Figure 2. FSI's Indices Cross-Correlation

The summary statistics of the calculated FSI based on Equation (8) and Equation (9) are presented in Table 1. The summary statistics shows that the FSI's highest figure (stress period) is at 0.74, which occurred on 30 October 2008 during the collapse of Lehman Brothers prior to the 2008 Global Financial Crisis.

Table 1. FSI's Summary Statistics

Stat.	ISTX	IR10	IEER	FSI
Max	0.9987	0.99647	0.85004	0.74166
Min	0.00817	0.04733	0.00037	0
Mean	0.49886	0.50509	0.49469	0.14039
Std. Dev	0.24504	0.20865	0.17918	0.15077

The historical data of FSI is presented in Figure 3. Our event-analysis shows that FSI generally can explain shocks to the financial system since 2005. For example, it shows the shock of higher subsidized oil price in March and October 2005 as reflected by the relatively moderate hikes. Meanwhile, it shows substantial spikes during the 2008 Global Financial Crisis. It also explains a series of Euro Crisis events (2010 and 2011) as well as The Fed's tightening monetary policy (2013 and 2016).

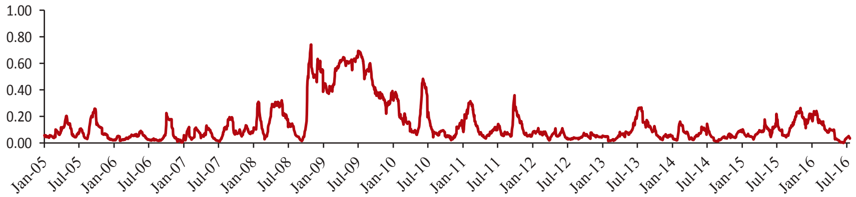


Figure 3. Indonesia's Financial Stress Indicator

In addition to event-analysis, we also compare FSI with an established financial stability indicator, i.e. FSSI. Our analysis shows that both indicators have positive and significant correlation with correlation parameter of 34%. This result confirms the complementary capability of FSI to measure financial system stability with high-frequency datra.

3.2. Threshold Calculation

After calculating the FSI, the next step is to determine the stress and normal levels of FSI. Two approaches are applied to separate the periods of high financial stress from periods with low and moderate stress levels. The first approach is the first order Markov Switching Autoregressive (MSAR), this is applied based on the assumption that FSI series are state-dependent and tend to cluster into specific levels across different regimes. For this estimation, we use the calculated FSI from January 2005 to July 2016. Figure 4 exhibits the distribution of FSI that tends to be skewed to the right and therefore suggesting that FSI may be separated into different regimes.

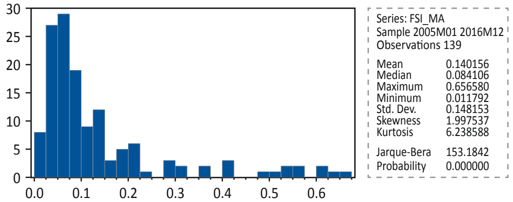


Figure 4. FSI's Histogram and Descriptive Statistics

We estimate several types of MSAR model for the FSI *(F_t)* up to three regimes *(S_t)*, as follows:

$$F_t = \alpha(s_t) + \beta(s_t)F_{t-1} + \sigma(s_t)u_t \text{ for } s_t = \{1,2,3\} \tag{10}$$

where every parameter $\alpha(S_t)$ intercept, $\beta(S_t)$ lagged of FSI, and $\sigma(S_t)$ are allowed to switch across regimes and u_t follows standard normal i.i.d process. The MSAR model also assumes that the stochastic process generating the states *st* follows an ergodic first order Markov chain with transition probabilities $p_{ij} = p(s_t = i|s_{t-1} = j)$ collected in the transition matrix P:

$$P = \begin{bmatrix} p_{1|1} & p_{1|2} & p_{1|3} \\ p_{2|1} & p_{2|2} & p_{2|3} \\ p_{3|1} & p_{3|2} & p_{3|3} \end{bmatrix} = \begin{bmatrix} p_{1|1} & p_{1|2} & p_{1|3} \\ p_{2|1} & p_{2|2} & p_{2|3} \\ 1 - p_{1|1} - p_{2|1} & 1 - p_{1|2} - p_{2|2} & 1 - p_{1|3} - p_{2|3} \end{bmatrix} \tag{11}$$

We find that FSI can be modelled as an MS(3)-DR(1), MS(3)-AR(1), MS(2)-DR(1), and MS(2)-AR(1).

MS(n)-AR(1) indicates that FSI can be modelled as a first order auto regression with all of the coefficients are allowed to switch across n regimes, meanwhile MS(n)-DR(1) suggests that FSI can be modelled as a first order auto regression with all the coefficients, except β (parameter for lagged of FSI), can vary across n regimes.

Table 2. Parameters Estimates from Different Specification of MSAR for FSI

	Variables	MS(3)DR(1)	MS(3)AR(1)	MS(2)DR(1)	MS(2)AR(1)
Regime 1	C	0.142 (0.020)***	0.453 (0.064)***	0.169 (0.028)***	0.036 (0.005)***
	FSI_MA(-1)	0.352 (0.053)***	0.182 (0.126)	0.493 (0. 050)***	0.364 (0.051)***
	LOG(SIGMA)	-2.673 (0.152)***	-2.581 (0.226)***	-2.176 (0.117)***	-3.691 (0.108)***
Regime 2	C	0.036 (0.005)***	0.034 (0.05)***	0.027 (0.005)***	0.100 (0.03)***
	FSI_MA(-1)	0.352 (0.053)***	0.374 (0.053)***	0.493 (0.050)***	0.751 (0.081)***
	LOG(SIGMA)	-3.728 (0.104)***	-3.704 (0.116)***	-3.673 (0.102)***	-2.329 (0.112)***
Regime 3	C	0.362 (0.040)***	0.107 (0.034)***		
	FSI_MA(-1)	0.352 (0.053)***	0.605 (0.103)***		
	LOG(SIGMA)	-2.423 (0.239)***	-3.042 (0.292)***		
SC		-2.70	-2.68	-2.66	-2.70
AIC		-2.98	-2.99	-2.80	-2.89

Note: standard error in parentheses; statistical significance in *** 1%, **5%, and *10%

As presented by Table 2, based on the significance of each variable among models and their criterions (SC and AIC), we consider MS(3)-DR(1) as the best model. Regime 1 represents the case when financial stress is on the intermediate level with a mean level around 0.215 (unconditional mean), this regime occurs 22.5% of the sample. Regime 2 represents the low financial stress with a mean level around 0.064 and can be considered as the dominant case since this regime occur 69% of the sample. Regime 3 represent the high level of financial stress with a mean level around 0.55 and occur only during the recent financial crisis (9% of the sample). Regimes classification of MS(3)-DR(1) for FSI presented on Figure 5. and Table 3 below.

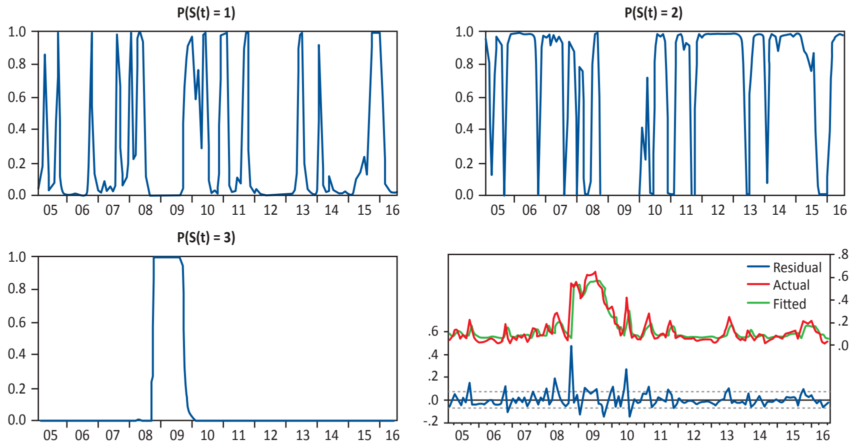


Figure 5. Smoothed Regime Probabilities, Fitted Values and Residuals from the MS(3)-DR(1) Model

Table 3. Transition Matrix and The Expected Duration of the MS(3)-DR(1) Model for FSI

Transition Prob	Regime 1, t	Regime 2, t	Regime 3, t
Regime 1, t-1	0.58	0.42	0
Regime 2, t-1	0.14	0.85	0.01
Regime 3, t-1	0.09	0	0.91

	Regime 1	Regime 2	Regime 3
Exp. Duration	2.4	6.54	11.34
Regime 2, t-1	0.14	0.85	0.01
Regime 3, t-1	0.09	0	0.91

The second econometric approach: threshold regression model (as introduced by Tsay (1989)) is used to determine the financial crisis levels that may affects the health of the real economy. FSI is assumed to be the relevant threshold variable determining regime shifts in the financial stress-economic activity relationship and can be represent as the following equations.

$$x_t = \begin{cases} c^H + \emptyset_1^H x_{t-1} + \emptyset_2^H x_{t-2} + \emptyset_3^H x_{t-3} + \varepsilon_t^H, if\ z_{t-d} > \tau,\ high\ stress\ regimes \\ c^L + \emptyset_1^L x_{t-1} + \emptyset_2^L x_{t-2} + \emptyset_3^L x_{t-3} + \varepsilon_t^L, if\ z_{t-d} \leq \tau,\ low\ stress\ regimes \end{cases} \quad (12)$$

with x_t is the two-dimensional vector of the endogenous variables (FSI and annual industrial production growth- as a proxy of economic growth), c^H and c^L the vector of intercepts, and $\emptyset_i^j$ the two matrices of the slope coefficients for states $j = H,L$ (with H and L standing for high-stress and low-stress regime) and lags $i = 1,2,3$.

We compute our preferred model that corresponds to a bivariate TVAR (FSI and annual growth of industrial production) with three lags and a threshold. Under this specification, the estimated threshold values of FSI is 0.1398. According to the result, FSI values below 0.1398 are considered low stress and FSI values above 0.1398 considered moderate to high financial stress. During low stress regimes, shocks in the FSI do not affect output significantly, on the other hand, intermediate and high stress regimes have a negative impact in industrial production.

Table 4. TVAR Estimation Results

Variables	FSI (-2) < 0.1398		FSI (-2) ≥ 0.1398	
	FSI_MA	G_IP	FSI_MA	G_IP
C	0.022	-0.029	0.063	0.035
	(0.026)	(0.019)	(0.026)	(0.010)***
FSI_MA(-1)	0.885	0.01	0.632	0.022
	(0.129)***	(0.062)	(0.128)***	(0.045)
FSI_MA(-2)	-0.296	1.211	0.198	-0.070
	(0.391)	(0.449)***	(0.147)	(0.055)

Table 4. TVAR estimation results *Lanjutan*

Variables	FSI (-2) < 0.1398		FSI (-2) ≥ 0.1398	
	FSI_MA	G_IP	FSI_MA	G_IP
FSI_MA(-3)	0.409	0.255	0.022	-0.003
	(0.206)**	(0.204)	(0.092)	(0.037)
G_IP(-1)	-0.150	-0.091	-0.659	0.273
	(0.259)	(0.156)	(0.489)	(0.108)**
G_IP(-2)	-0.002	0.415	-0.088	-0.111
	(0.279)	(0.164)**	(0.361)	(0.106)**
G_IP(-3)	0.048	0.309	-1.359	0.252
	(0.223)	(0.127)**	(0.510)***	(0.115)**

Note: standard error in parentheses; statistical significance in *** 1%, **5%, and *10%

4. CONCLUSIONS

We calculate FSI for Indonesian financial system using financial market data over the period 2005-2016. The following qualitative and quantitative analyses confirm the intuitiveness of FSI to be used as a complementary indicator in measuring financial system stability. The final part of the analysis shows some attempts in determining the threshold that separates between low stress period and high stress period.

This work has several implications on macroprudential analyses in Indonesia. First, it proposes a technical approach that relies on high-frequency data in determining a financial crisis. Second, the use of high-frequency data suggests that this indicator can be updated immediately to capture the most recent development in the financial system thus can act as an early warning model to complement the existing indicators. Finally, this indicator can also be utilized as an indicator to measure the effectiveness of macroprudential policy.

Indeed, further development is necessary particularly vis-à-vis capturing the dynamics in the banking system considering it accounts for 70%-80% assets of the whole financial system. Therefore, further study should focus on this aspect of development.

REFERENCES

Acharya, V.V., L.H. Pedersen, T. Philippon, M. Richardson. 2010. Measuring Systemic Risk. FRB of Cleveland Working Paper No. 10-02.

Adrian, T., M.K. Brunnermeier, 2011. CoVaR. NBER Working Paper No. 17454.

Albulescu, C.T. 2008. Assessing Romanian financial sector stability: The importance of the international economic climate. MPRA Paper No. 16581.

Brownlees, C.T., R. Engle. 2012. Volatility, correlation and tails for systemic risk measurement. Working Paper, NYU.

Cambón, M.I., L. Estévez. 2016. A Spanish Financial Market Stress Index (FMSI). The Spanish Review of Financial Economics 14 (1) 23–41.

Duprey, T., B. Klaus, & T. Peltonen. 2015. Dating systemic financial stress episodes in the EU countries. European Central Bank Working Paper Series No. 1873/December 2015.

Holló, D., M. Kremer, & M.L. Duca. 2012. CISS – A composite indicator of systemic risk in the financial system. European Central Bank Working Paper Series No. 1426/March 2012.

IMF-BIS-FSB. 2009. Guidance to assess the systemic importance of financial institutions, markets and instruments: Initial considerations. International Monetary Fund, Bank for International Settlements and Financial Stability Board.

Illing, M., Y. Liu. 2006. Measuring financial stress in a developed country: An application to Canada. Journal of Financial Stability 2, 243–265.

Karanovic, G., B. Karanovic. 2015. Developing an aggregate index for measuring financial stability in the Balkans. Procedia Economics and Finance 33, 3-17.

Tsay, R., S. 1989. Testing and modeling threshold autoregressive processes. Journal of the American Statistical Association 84 (405), pp. 231-240.

Prediction Based Portfolio Optimization Model Using Neural Networks with an Emphasis on Leading Stocks of NSE

Gajendra K. Vishwakarma* and Chinmoy Paul
Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad-826004, India
Email: vishwagk@rediffmail.com, chinmoy.gcc@gmail.com

Abstract

This work discusses most frequently traded stocks of National stock exchange of India. A prediction based portfolio optimization model is considered to present an ideal portfolio out of the considered stocks. Neural network has been used to predict stock returns and a risk measure is derived that has the same foundation as that of mean variance model. The architecture of the network is designed experimentally. The cross-correlation structure between the fluctuations of price for frequently traded stocks in national stock exchange has been studied for a period of last one year. The structure of interactions is studied using the spectral properties of the cross-correlation matrix. Spectral decomposition of the cross correlation matrix is used in finding the most influencing underlying stocks of NIFTY 50. The portfolio formed by the presented model is considered to verify the presence of influencing stocks as determined by the spectral analysis.

Keywords: Portfolio Selection; Time series prediction; Neural Network; Spectral Decomposition

1. INTRODUCTION

National Stock Exchange (NSE) and Bombay Stock Exchange (BSE) are the leading stock exchange in India. These two exchanges have brought about unparalleled transparency, speed and efficiency, safety and market integrity. Now with the growing digitalization and ease of banking facilities customer is slowly and gradually getting attraction towards online trading. Online trading accounts are the platform where a bank provides its customer the facility to invest in the share market without having the trouble to find a broker. Before making an investment an individual will like to optimize his portfolio in terms of minimizing risk and maximizing return. Selection of instrument for investment is a crucial task in financial theory and practice and it is basically aligned with the future performance of the selected instruments, mainly their expected returns. When investments are exposed to uncertainties, the investment selection framework must include a quantitative measure of the uncertainty of obtaining the expected return, i.e. a quantitative measure of risk.

The return of an index depends on the performance of the underlying constituents. A landmark in modern portfolio theory is the Mean Value model proposed by Markowitz (1952, 1991). In mean variance model, risk in an investment is minimized by optimal selection of instruments with low joint risk, which provides a mechanism of loss compensation known as efficient diversification. Portfolio optimization procedure consists in selecting a number of stocks, the participation of each stock that minimizes risk with a given level of return. Many other models are developed Elton et. al. (2009) using the fundamental assumption of Mean variance model. The assumption of the Mean Variance model is often violated by many real word examples. One such violation is deviation from normality by the returns of the stock Fama (1965).

The predictability of stock markets is still an open area in finance theory. The efficient market hypothesis (EMH), the theoretical framework that guides the discussion around this question, has been under empirical testing and reviewing for the past several decades Malkiel and Fama (1970), Fama (1991, 1998), Malkiel (2003). Market efficiency implies a random walk model for the prices of stocks, but pricing irregularities and predictable patterns like serial correlations, calendar effects, and even sports results effects do appear Malkiel (2003), Edmans et.al. (2007).

Forecasting in time series was customarily attempted by the linear models of time series analysis Box et. al. (2015). However, if the nature of time series is nonlinear prediction using linear models may lead to results far away from the expected. In recent past many machine learning approach such as neural network, support vector machine, fuzzy sets (White (1988), Huarng et. al. (2007)) for time series forecasting has come up. Among these approaches neural network approach is advantageous for forecasting and estimation of model parameters due its ability to estimate the functional relationship between input and output variables be it linear or non linear.

Fernandez and Gomez in (2007) presented a heuristic method based on artificial neural network to focus on the problem of tracing out the efficient frontier for the general mean variance model with cardinality and bounding constraints. Hopfield network had been used to the portfolio selection problem. All the experiments showed that the neural network model has given better solutions than the other three heuristic methods. Thawornwong and Enke (2004) presented a data mining approach to choose economic and financial variables and used the selected variables with predictive power as input of a neural predictor. The neural predictor is used in predicting the direction (sign) of future return movements. The predictions were used to implement a trading strategy for deciding either to invest on the SandP500 index portfolio or on the T-Bill (risk-free) during one month. The result of their study showed that the monthly returns obtained with this adaptive strategy were always higher than those of the (non-adaptive) compared methods.

Niaki and Hoseinzade (2013) used artificial neural network to forecast the daily price change direction of standard and poor's 500 (SandP 500) index. Factorial design of experiment considering 27 potential financial and economical variables as factors was applied to find the most influential features for the proposed ANN. The results of their proposed approach showed that the ANN that uses the most influential features is able to forecast the daily direction of SandP 500 significantly better than the traditional model. Sinha and Kovur (2014) in their work analyzed cross correlation between the fluctuations of 74 currencies. The Eigen value analysis of the cross correlation matrix presented that bulk of the Eigen values falls within the bounds predicted by random matrix. However, the ones deviating from the bulk over the bounds contain the important information about groups of strongly interacting components. This idea is used in finding the most influential stock among the underlying 50 stocks of the index Nifty-50 and is compared with the portfolio formed by portfolio selection model suggested by Freitas et. al (2009).

In this work, we present an approach to select optimal portfolio out of a given set of instruments. The portfolio optimization is based on the predictions using artificial neural network. Another approach based on the Eigen values of the cross correlation matrix is also presented. Both the method suggests an efficient portfolio out of a given set of instruments. A comparison between both the methods has also been carried out.

2. CROSS CORRELATION MATRIX FOR THE RETURN

To quantify the degree of association-ship between the underlying of Nifty 50 index, first the return for each underlying instrument is calculated

$$R_i(t,\Delta t) = P_i(t+\Delta t) - P_i(t) \tag{1}$$

where $P_i(t)$ is the price ith underlying security at time *t*. Prices considered here are the daily closing price. Δt is considered as one day. The standard deviation of the returns and normalized returns are obtained as $\sigma_i = \sqrt{\langle R_i - \bar{R}_i \rangle^2}$ and $r_i(t,\Delta t) = R_i / \sigma_i$ respectively.

After obtaining the normalized returns for all N underlying over a period of T days the cross-correlation matrix (C) between the normalized returns is computed as $C_{ij} = \langle r_i, r_j \rangle$. If the movement of the different underlying instruments is uncorrelated, the resulting random correlation matrix (referred to as a wishart matrix) has eigen values distributed according to Sengupta and Mitra (1999).

$$D(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda_{\min} - \lambda)}}{\lambda} \tag{2}$$

As $N \to \infty$ and $T \to \infty$ with $Q = (T/N) > 1$ the distribution of the eigen values of the normalized returns are given by $\lambda_{Max} = [1 + (1/Q)]^2$ and $\lambda_{Min} = [1 - (1/Q)]^2$. The data that has been analyzed the value of Q is 4.92 and $(\lambda_{Max}, \lambda_{Min}) = (1.45, 0.635)$. The distribution of the eigen value is shown in Fig.1.

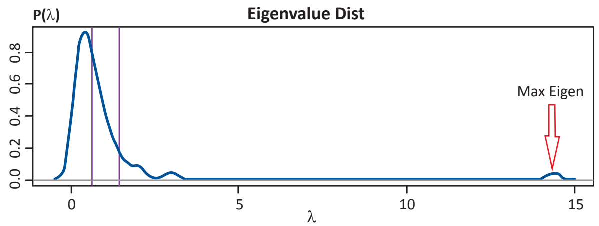


Figure 1: The Distribution of the Eigen Value

3. PREDICTION BASED ON OPTIMIZATION OF PORTFOLIO

Mean Variance theorem by Harry Markowitz has received considerable research attention Markowitz (1956), Hamza and Janssen (1996), Sharpe (1963), Konno and Yamazaki (1991), Sortino and Van Der Meer (1991) due to its computational feasibility, model simplifications and the development of risk measures. The prediction of future return in the context of portfolio selection has received little attention. In most of the models same prediction method is employed as in the case of Mean Variance theorem i.e. the mean of the past return. However, mean returns are expected to be verified only in case of long term predictions and show inadequacy in short term prediction of future return. Here an alternative procedure has been presented that describes prediction based portfolio optimization model using predicted returns as the expected return. Unlike mean variance theorem, this procedure uses the variance of the errors of prediction as risk measure.

3.1. *Expected return and risk of a stock*

Let the return of a stock and its expected returns are denoted by r_t and $\widehat{r}_t$ the relation between them is given by

$$r_t = \widehat{r}_t + \varepsilon_t \tag{3}$$

where r_t is the stock return at time *t*, $\widehat{r}_t$ is the predicted return for time *t* obtained at *t-1*and is ε_t the error at time *t*. Rearranging the terms, the error at time *t* can be presented as

$$\varepsilon_t = r_t - \widehat{r}_t \tag{4}$$

For an unbiased predictor the series of error $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n)$ prediction must be statistically independently and identically distributed with mean and variance given by

$$\mu = E(\varepsilon_t) = 0 \tag{5}$$

and

$$v = \sigma^2 = \frac{1}{n-1}\sum_{t=1}^{n}\varepsilon_t^2 \tag{6}$$

The variance equation gives the amount of uncertainty about the realization of predicted returns. Higher the variance higher is the risk. The variance component is used in modeling the optimization models for a portfolio.

3.2. *Expected return and risk of a portfolio*

A portfolio is a collection of *P* stocks and the corresponding weights (ω_i) in such a way that each of the weight ω_i, $i = 1, 2, \ldots P$, $0 \le \omega_i \le 1$ and $\sum_{i=1}^{p}\omega_i = 1$ represents fraction of portfolio invested in stock *i*. The expected return of the portfolio or the predicted return of the portfolio $\widehat{r}_p$ is the linear combination of expected return of each of the stocks and the corresponding weights.

$$\widehat{r}_p = \sum_{i=1}^{p}\omega_i\, r_i \tag{7}$$

With the assumption that return of each stock follows normal distribution, portfolio risk is considered as the variance of the joint distribution of the linear combination of expected return of each stock and its corresponding weights.

$$V = Var\left(\sum_{i=1}^{p}\omega_i r_i\right) = \sum_{i=1}^{p}\sum_{j=1}^{p}\omega_i\,\omega_j\,\psi_{\varepsilon ij} \tag{8}$$

Where V is the total portfolio risk, P is the number of stocks and ω_i, ω_j are participation of the stock i and stock j respectively. $\psi_{\varepsilon ij}$ represents the covariance between the expected returns of the stock i and stock j respectively and is given by

$$\psi_{\varepsilon ij} = \psi_{ij} = \frac{1}{n-1}\sum_{t=1}^{n}\varepsilon_{ti}\varepsilon_{tj} \tag{9}$$

Equation (9) can also be presented as

$$V = \hat{\sigma}_p^2 = \sum_{i=1}^{p}\omega_i^2\sigma_{\varepsilon i}^2 + \sum_{i=1}^{p}\sum_{i=1}^{p}\omega_i\omega_j\psi_{\varepsilon ij} \tag{10}$$

here the first term represents risk associated to each of the stocks to the portfolio and the second term represents interactive predictive risk of the presence of each pair of the stock i and j.

3.3. *Optimization of the portfolio*

Keeping all the assumptions intact the optimization model can be presented as

$$Minimize \quad V = \sum_{i=1}^{p}\omega_i^2\sigma_{\varepsilon i}^2 + \sum_{i=1}^{p}\sum_{i=1}^{p}\omega_i\omega_j\psi_{\varepsilon ij} \tag{11}$$

Subject to

$$\sum_{i=1}^{p}\omega_i\hat{r}_i = R_d \tag{12}$$

$$\sum_{i=1}^{p}\omega_i = 1 \;\; ; \; 0 \le \omega_i, \; i = 1, 2, \ldots, p \tag{13}$$

The formulated optimization problem minimizes the risk (Eq. 11) at a desired return (Eq. 12) and Eq. (13) represents the restriction of the model to purchase only.

4. EXPERIMENT

This section presents the experiment that has been carried out using the underlying constituents of Nifty-50 index. The method used to predict the returns along with the measures to verify the performance of the prediction based portfolio optimization model are also presented.

4.1. Data

We have considered daily closing prices of 50 companies listed under Nifty-50. This data is commonly available in the websites of BSE or NSE (https://www.nseindia.com, http://www.bseindia.com). A period of 1 year span has been considered starting from 17 December 2015 till 15 December 2016. A crucial point in collecting data for the constituents of Nifty -50 index is that, the index is reviewed after every six months and based on the decision of index policy committee the underlying constituent changes. Here we have considered the same set of index constituents that were present at during the period of July- December 2016. Historical data for the same set of constituents has been considered.

4.2. Prediction of returns

Backpropagation network has been used for predicting the returns Wong (1991), Zhang et.al. (1998), Ghiassi and Saidane (2005). The architecture of the network is designed experimentally. Depending upon the complexity of the problem suitable network architecture is designed. For checking the normality of the error of the predicted returns the last step of our proposed algorithm (Paul and Vishwakarma 2016) has been used. This imposed condition not only checks normality of the error of the predicted returns but also helps in finding suitable network architecture for a particular input data set.

Entire data set is divided in three parts, 50% of the data is used in train, 30% validation and 20% for testing. The use of training and validation is to control overfitting. Haykin (1999) presented incompetency in controlling overfitting for non linear time series prediction due to its structural regime change. Though there is no hard and fast rule available for controlling overfitting a better suitable technique like shifting the validation segment to other locations in time Pantazopoulos et. al. (1998) is used. The overfitting controlling procedure we have used is inspired in the Non-linear Cross Validation (NCV) proposed in Moody, J. (1994). The training and testing procedure described above is repeated for all 50 underlying of the Nifty-50 index by advancing a sliding widow of twenty days one day at a time.

4.2.1. Evaluation Matrics

The performance evaluation measures that are used in this study are *Mean Error (ME), Root Mean Square Error (RMSE)*, *Absolute percentage Error (APE)* and *Hit Rates Matrics (HRM)*.

Mean Error is the average of the difference of actual and the predicted returns and is given by

$$ME = \frac{1}{n}\sum_{t=1}^{n} r_t - \hat{r}_t \tag{14}$$

r_t and $\hat{r}_t$ are the realized and expected returns of a time series at time t and n is the length of the time series. ME is basically used to verify the assumption that the error of the expected return follow normal distribution with zero mean and a fixed standard distribution.

RMSE is used in comparing two time series in terms of its variability. It is well known that RMSE is badly affected by the presence of outliers, hence a time series with outlier will have more RMSE compared to other and is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(r_t - \hat{r}_t\right)^2} \tag{15}$$

The APE is defined as

$$APE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{r_t - \hat{r}_t}{r_t}\right| \tag{16}$$

APE is unit free and averages out very large values and very sensitive to very small changes in data. Hit rate H_R, H_{R^+}, H_{R^-} measures the rate of coincidence of actual return r_t and $\hat{r}_t$ the predicted return. H_R measures the percentage where both r_t and $\hat{r}_t$ have same signal and are different from zero. H_{R^+} measures the percentage where both are positive and H_{R^-} measures the percentage where both have same signal and are negative.

$$H_R = \frac{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ r_t\hat{r}_t > 0\\ 0 & otherwise\end{cases}}{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ r_t\hat{r}_t \neq 0\\ 0 & otherwise\end{cases}},\quad H_{R^+} = \frac{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ r_t > 0\ and\ \hat{r}_t > 0\\ 0 & otherwise\end{cases}}{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ \hat{r}_t > 0\\ 0 & otherwise\end{cases}},\quad H_{R^-} = \frac{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ f\ r_t < 0\ and\ \hat{r}_t < 0\\ 0 & otherwise\end{cases}}{\sum_{t=1}^{n} I_t = \begin{cases}1 & if\ \hat{r}_t < 0\\ 0 & otherwise\end{cases}}$$

The heat rate measure will give us the idea of the underlying stocks that that are with positive return, normal return and negative return. So, the search for the effective underlying can be traced out with the help of hit rate measures. The underlying time series with maximum hit rate should be included in the portfolio to achieve maximum return.

5. RESULTS AND DISCUSSION

Portfolio selection is an open problem and there are enormous research articles suggesting methods to select portfolio. The main target here is to find a set of underlying from a set of available instruments. The set of available instruments here are 50 underlying of Nifty-50 index. Various evaluation measures discussed in previous section i.e. ME, RMSE, APE, H_R, H_{R^+}, H_{R^-} are presented in Table-1 (appendix). Value of H_{R^+} above 50% indicates that the predictor achieved a performance above the pure chance of predicting the positive return of the market. Instruments with high percentages of H_{R^+} are highlighted in color. From the distribution of eigen value it is observed that bulk of the eigen values are beyond the defined limit and Maximum eigen value is around 10 times the upper bound. Eigen vectors corresponding to first four (magnitude wise) eigen value are presented in Fig-2 and forecasting return time series using neural network shown in Fig-3. If we closely consider the distribution of eigen vector corresponding to 1st eigen value, it may generally be stated that higher weights are given to those instruments which are having higher percentages of H_{R^+}.

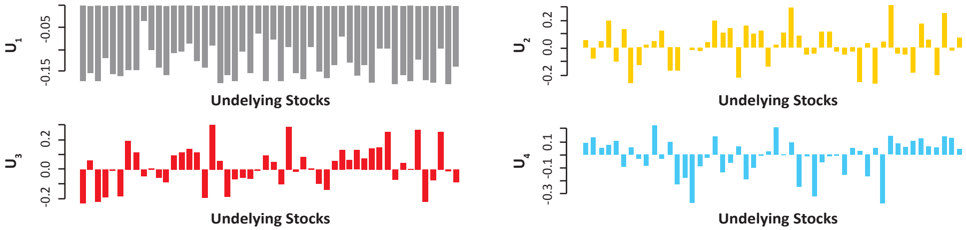


Figure 2: Eigen Vectors Corresponding to First Four Eigen Value

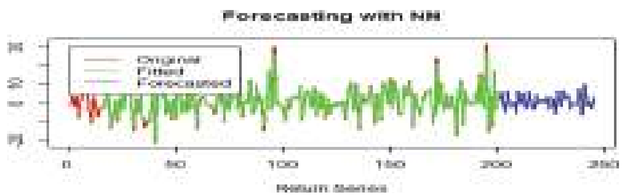


Figure 3: Forecasting Return Time Series using Neural Network

5. CONCLUSION

In this article underlying instruments of Nifty-50 index are analyzed to form a portfolio that can maximize return with minimum risk. Two different approaches one based on cross correlation structure of the return matrix and the other based on prediction using neural network are discussed. Maximum eigen value of the cross correlation matrix explains the maximum variability of the return data. The corresponding eigen vector gives higher weight to those underlying which are expected to contribute towards portfolio formation. Underlying instruments having higher weight and the instruments having high percentages of are H_{R^+} almost same. Hence based on the present study any one of the approach can be adopted for a portfolio selection in case of predefined set of underlying are available at hand The set of instrument considered in this study are from a well-known index. The return of the underlying time may depend on other potential influential features which are beyond the scope of this study. Only partial results based on H_{R^+} and cross correlation matrix are presented which provide an initial structure of a portfolio with expected positive return, the final step i.e. minimization of risk is under exploration.

REFERENCES

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). Time series analysis: forecasting and control. John Wiley and Sons.

Edmans, A., Garcia, D. and Norli, O. (2007). Sports sentiment and stock returns. The Journal of Finance, 62(4), 1967-1998.

Elton, E.J., Gruber, M.J., Brown, S.J. and Goetzmann, W.N. (2009). Modern portfolio theory and investment analysis. John Wiley and Sons.

Fama, E.F. (1965). Portfolio analysis in a stable Paretian market. Management science, 11(3), 404-419.

Fama, E. F. (1991). Efficient capital markets II. The journal of Finance, 46(5), 1575-1617.

Fama, E.F. (1998). Market efficiency, long-term returns, and behavioral finance. Journal of financial economics, 49(3), 283-306.

Fernández, A. and Gómez, S. (2007). Portfolio selection using neural networks. Computers and Operations Research, 34(4), 1177-1191.

Freitas, F.D., De Souza, A.F. and De Almeida, A.R. (2009). Prediction-based portfolio optimization model using neural networks. Neurocomputing, 72(10), 2155-2170.

Ghiassi, M. and Saidane, H. (2005). A dynamic architecture for artificial neural networks. Neurocomputing, 63, 397-413.

Hamza, F. and Janssen, J. (1996). Linear approach for solving large-scale portfolio optimization problems in a lognormal market. In IAA/AFIR Colloquium, Nürnberg, Germany (pp. 1019-1039).

Haykin, S. (1999). Neural Networks, A comprehensive Foundation Second Edition by Prentice-Hall.

Huarng, K.H., Yu, T.H.K. and Hsu, Y.W. (2007). A multivariate heuristic model for fuzzy time-series forecasting. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37(4), 836-846.

Konno, H. and Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. Management science, 37(5), 519-531.

Malkiel, B.G. (2003). The efficient market hypothesis and its critics. The Journal of Economic Perspectives, 17(1), 59-82.

Malkiel, B.G., and Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.

Markowitz, H. (1956). The optimization of a quadratic function subject to linear constraints. Naval research logistics Quarterly, 3(1-2), 111-133.

Markowitz, H.M. (1952). Portfolio selection. The journal of finance, 7(1), 77-91.

Markowitz, H.M. (1991). Foundations of portfolio theory. The journal of finance, 46(2), 469-477.

Moody, J. (1994). Prediction risk and architecture selection for neural networks. In From Statistics to Neural Networks (pp. 147-165). Springer Berlin Heidelberg.

Niaki, S.T.A. and Hoseinzade, S. (2013). Forecasting SandP 500 index using artificial neural networks and design of experiments. Journal of Industrial Engineering International, 9(1), 1-9.

Pantazopoulos, K.N., Tsoukalas, L.H., Bourbakis, N.G., Brun, M.J. and Houstis, E.N. (1998). Financial prediction and trading strategies using neurofuzzy approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28(4), 520-531.

Paul, C. and Vishwakarma, G.K. (2016): Back Propagation Neural Networks and Multiple Regressions in the Case of Heteroscedasticity, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2016.1212066.

Sengupta, A.M. and Mitra, P.P. (1999). Distributions of singular values for some random matrices. Physical Review E, 60(3), 3389.

Sharpe, W.F. (1963). A simplified model for portfolio analysis. Management science, 9(2), 277-293.

Sinha, S. and Kovur, U. (2014). Uncovering the network structure of the world currency market: Cross-correlations in the fluctuations of daily exchange rates. In Econophysics of Agent-Based Models (pp. 203-218). Springer International Publishing.

Classifying of Companies Listed in IDX LQ45

Maiyastri*
Jurusan Matematika, FMIPA, Universitas Andalas, Padang, Indonesia-maiyastri@gmail.com
Dodi Devianto
Jurusan Matematika, FMIPA, Universitas Andalas, Padang, Indonesia-ddevianto @gmail.com
Efa Yonedi
Jurusan Akuntansi, Fak. Ekonomi, Universitas Andalas, Padang, Indonesia

Abstract

Selecting companies for investment is very important. This research want to group companies listed in IDX LQ45 by using factor analysis and cluster analysis. The variables will be analyzed are profit, revenues, capital, liabilities, Net profit margin, dividend, and closing price. There are two groups of variables according of factor analysis namely; Volume Company and Performance Company. Based on result of factor analysis we classify companies. There two groups companies; the first is the efficient companies (6 companies) and the second is not efficient companies (39 companies).

Keywords: Classify, Factor analysis, Cluster analysis.
JEL: G11, G32

1. INTRODUCTION

LQ45 Index is an indicator of stock market indices Indonesia. LQ45 is a forum in which contains a company whose shares have liquidity and high market capitalization. Therefore, a company listed on LQ45 company is the reference point of investing in IDX. One benchmark is the selection of the company financial management, it can be seen from its financial reports.

This study is intended to create a grouping of companies listed in the LQ45 index is based on financial statements published by IDX any given period. The variables will be analyzed are profit, revenues, capital, liabilities, *Net profit margin*, dividend, and closing price. Dividend each year end, therefore, the data of this study using the company's financial report data in December 2015.

This study uses a multivariate analysis, therefore it is necessary to reduce the variables, the analysis used is factor analysis. The results of the analysis factors will be used in the grouping of companies listed in LQ45. The Clustering of company is started by selects the best grouping method. After that, the effectiveness of grouping selected is analyzed by Manova.

2. RESEARCH METHODS

2.1. Factor Analysis

According to Johnson and Wichern (1988) the purposes of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. Factors which obtained a smaller number of variables of origin, but he contains maximum information from the origin because of the variable formed from the variables are highly correlated with each other.It is an extension of principal component analysis.

2.2. Cluster Analysis

The basic objective of cluster analysis is to discover natural groupings of items. There is no assumptions are made concerning the number of groups or the the group structure. Grouping is done on the basis of similirities or distances (dissimilities) (Johnson and Wichern (1988)). In this research, we use the Euclidean Distance. The Euclidean distance between two p-dimensional observation $y = (y_1, y_2,\ldots,y_p)$" and $x = (x_1, x_2,\ldots,x_p)$' is

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})}.$$

Clustering methods is used is Hierarichical Clustering Metods. Hierarchical method is divided into two, namely:
1. Agglomerative hierarchical methods.
2. Divisive hierarchical methods.

The result of agglomerative hierarchical and divisive methods my be displayed in the form of a two-dimensional diagram known as ***dendogram***

This study concentrate on agglomerative hierarchical procedures, inparticular linkage methods. We shall use Average Linkage, Centroid Linkage, Complete Linkage, Mc Quity Linkage and Ward Linkage.

2.3. Multivariate Analysis of Variance (Manova)

MANOVA tests for the difference in vector means between two or more groups. The vector means are vector contains means all variables such as profit, revenues, capital, liabilities, *Net profit margin*, dividend, and closing price. If there are several groupings, we want to know the grouping which one is more efficient based vectors existing variable. The model for Manova is defined by:

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_{ij}$$

Where:

y_{ij} = vector observation

μ = vector overall mean

τ_i = vector treatment effect

ε_{ij} = vector random error, and $i=1,2, \ldots, k$ and $j=1,2, \ldots, n_{i.}$, k is number of groups and n_i number replication for each group.

The null hypothesis is : $H0: \tau_1 = \tau_2 = \ldots = \tau_k = 0$

2.4. Ratio to Determine the Goodness of Grouping Method

To know the goodness which method has the best performance, it can be used the standard deviation within cluster (S_W) and standard deviation between (S_B).

Formula standard deviation within cluster (S_W):

$$S_W = K^{-1}\sum_{k=1}^{K} S_k$$

where ; S_k = standard deviation of k-th group/cluster
k = numbers of gropus
S_W = standard deviationwithin group/cluster

And formula for standard deviation between cluster (S_B):

$$S_B = [(K-1)^{-1}\sum_{k=1}^{K}(\bar{X}_k - \bar{X})^2]^{1/2}$$

Where: $\bar{X}_k$ = mean of k-th group/cluster
$\bar{X}$ = overall mean

The method has the smallest ratio is the best method. Ratio is defined by: $R = \frac{S_W}{S_B}$.

3. ANALYSIS

This section will explain general information about the data, result of factor analysis, choosing the best clustering, and interpretation the cluster of the company listed in Lq45 index.

3.1. Data Description

Figure 1 give the general information about the data, the presentation use box-plot diagram. From the diagram, we seen almost all the variables homogenous, but any outliers present in variable revenues, capital and liabilites. Company number 8, 9, 12 and 12 have bigger liabilities than others.

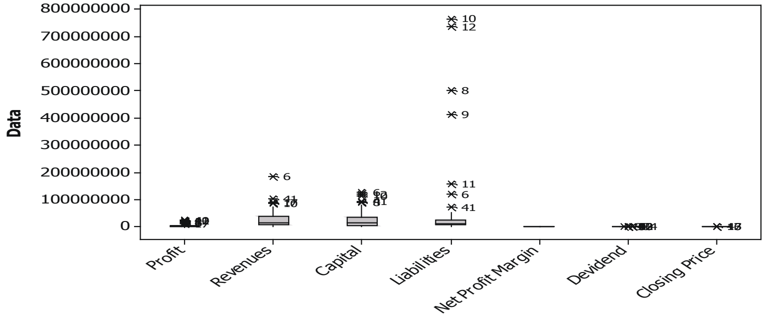


Figure 1. The Box-Plot Diagram for all Variables

3.2. Factor Analysis

The results of the factor analysis that there are two eigenvalues greater than 1, This means that there are two factors formed. Eigenvalues ordered from largest to smallest, with the criteria that the factors that have eigenvalues lesser than 1, there are not used in calculating the number of factors formed. The three first eigen values are 3.7712 1.2468 and 0.8757. The two factors is the most optimal amount can be obtained. Component factor shows the correlation between a variable with these two factors. The component factor after the rotation can be seen in Table 1.

Once it is known that two factors are the most optimal number, then Table 3 shows the distribution of seven variables on two factors formed. The numbres in this table are component factors that determine the correlation between a variable with factor 1 and factor 2. The process of determining which variables will include into the factors which, conducted by the comparison of the correlations for each variable. As seen from the table, the component factor can be interpretation. The Factor 1 consist of variables Profit, Revenues, Capital, Liabilities and Dividend. This factor can be named the performed company by internal indicator. The second factor formed from Net Profit Margin and Closing Price and it can be named the performed company by external indicator.

Tabel 1. Componen Matrix After Varimax Rotaion

	Factor-1	Factor-2
Profit	0,953302	0,118954
Revenues	0,736358	0,551838
Capital	0,927212	0,135995
Liabilities	0,877785	-0,23914
Net Profit Marginv	0,409869	-0,60777
Devidend	0,675139	-0,03739
Closing Price	0,179607	0,718004

3.3. Cluster Analysis

In this study, the distance matrix used the Euclidean distance and and agglomerative hierarchical procedures. For the linkage methods used Average Linkage, Centroid Linkage, Complete Linkage, Mc Quity Linkage and Ward Linkage. Performance cluster can be seen from the smallest ratio (R). If the ratio is less than 1 means that the diversity in the cluster is much smaller than the variation among clusters. It means the cluster is efficient. The value of ratio can be seen in Tabel 3.

Tabel 2. Cluster Analysis Recapitulation

No	Linkage Methods	Numbers of Cluster	Memberrs of Cluster			Ratio
			1st-Cluster	2nd-Cluster	3rd-Cluster	Sw/Sb
1	Average Linkage	2	39(others)	6(6,8,9,10,12,41)	-	0,387
3	Centroid Linkage	3	37(others)	2(16,17)	6(6,8,9,10,12,41)	0,334
4	Complete Linkage	3	38(others)	3(6,16,17)	5(8,9,10,12,41)	0,521
5	Mc Quity Linkage	2	39(others)	6(6,8,9,10,12,41)	-	0,387
6	Ward Linkage	2	39(others)	6(6,8,9,10,12,41)	-	0,387

Tablel 2 shows that the ratio of centroid linkage is the smallest (0.334). This value almost the same with three methods linkage (average, mc Quity and ward), it is 0.387. The highest ratio is 0.521 is from complete linkage. The last method is rejected. Therefore there are two candidates for clustering can be used for decision making grouping of companies listed in LQ45 index. The first clustering consist of 3 clusters and the second one 2 clusters. The dendogram for two clustering can be seen in Figures 2 and 3.

To make the accurate decision on the selects of clustering, we compare the performance clustering by Manova. This analysis tests the grouping which has a higher level of significance. The result of Manova shows that in Table 4.

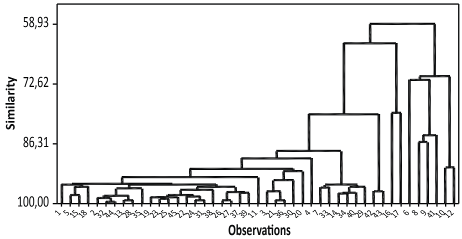


Figure 2. Dendogram of Centroid Linkage (3 Clusters)

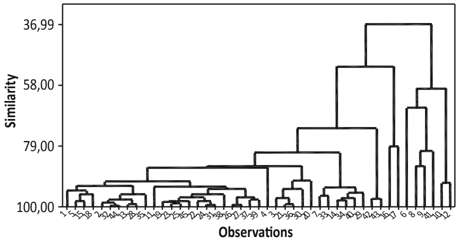


Figure 3. Dendogram of Average Linkage (2 Clusters)

Table 3. The Manova Results for Two Clustering

	Test		F statistic
First cluster	Wilks'	0,0872	48,92
Second Cluster	Wilks'	0,17927	96,143

The F statistic of second clustering (2 cluster) is 96,143, higher than the first cluster (3 cluster), it means the second clustering better than first clustering. To more explore the performance the both clustering is used the line chart of mean all variables for two clstering. The line chart in Figures 4 and 5 show that the second clustering sharper than the first, so it can be concluded that that the second clustering which produces 2 cluster better than the first clustering.

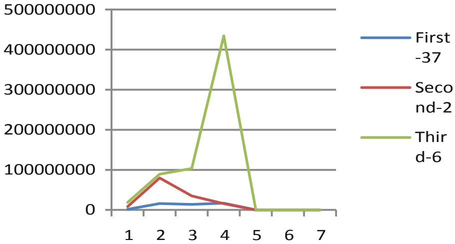


Figure 4. Line Chart First Clustering (3 clusters)

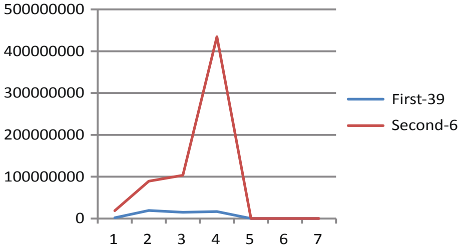


Figure 5. Line Diagram Second Clustering (2 cluster)

Based on all analysis has been done, it was decided that there are two cluster companies listed in LQ45 index. Members of the clusters are shown in Table 5. From the line chart is seen that the performance of the second cluster consists of 6 companies better and more efficient than the first cluster that consisting of 39 companies. Almost all of the financial performance for the average value of the second cluster is much higher.

Table 5. Classifying Companies

First Cluster (39 Companies)	1	AALI	Astra Agro Lestari Tbk	25	LPPF	Matahari Department Store Tbk
	2	ADHI	Adhi Karya (Persero) Tbk	26	LSIP	PP London Sumatera Tbk
	3	ADRO	Adaro Energy Tbk	27	MNCN	Media Nusantara Citra Tbk
	4	ANTM	Aneka Tambang (Persero) Tbk	28	MPPA	Matahari Putra Prima Tbk
	5	AKRA	AKR Corporindo Tbk	29	MYRX	Hanson International Tbk
	7	ASRI	Alam Sutera Realty Tbk	30	PGAS	Perusahaan Gas Negara (Persero) Tbk
	11	BBTN	Bank Tabungan Negara (Persero) Tbk	31	PTBA	Tambang Batubara Bukit Asam (Persero) Tbk
	13	BMTR	Global Mediacom Tbk	32	PTPP	Pembangunan Perumahan (Persero) Tbk
	14	BSDE	Bumi Serpong Damai Tbk	33	PWON	Pakuwon Jati Tbk
	15	CPIN	Charoen Pokphand Indonesia Tbk	34	SCMA	Surya Citra Media Tbk
	16	GGRM	Gudang Garam Tbk	35	SILO	Siloam International Hospitals Tbk
	17	HMSP	Hanjaya Mandala Sampoerna Tbk	36	SMGR	Semen Indonesia (Persero) Tbk
	18	ICBP	Indofood CBP Sukses Makmur Tbk	37	SMRA	Summarecon Agung Tbk
	19	INCO	Vale Indonesia Tbk	38	SRIL	Sri Rejeki Isman Tbk
	20	INDF	Indofood Sukses Makmur Tbk	39	SSMS	Sawit Sumbermas Sarana Tbk
	21	INTP	Indocement Tunggal Prakasa Tbk	40	TBIG	Tower Bersama Infrastructure Tbk
	22	JSMR	Jasa Marga (Persero) Tbk	42	UNTR	United Tractors Tbk
	23	KLBF	Kalbe Farma Tbk	43	UNVR	Unilever Indonesia Tbk
	24	LPKR	Lippo Karawaci Tbk	44	WIKA	Wijaya Karya (Persero) Tbk
				45	WSKT	Waskita Karya (Persero) Tbk
Second Cluster (6 companies)	6	ASII	Astra International Tbk	10	BBRI	Bank Rakyat Indonesia (Persero) Tbk
	8	BBCA	Bank Central Asia Tbk	12	BMRI	Bank Mandiri (Persero) Tbk
	9	BBNI	Bank Negara Indonesia (Persero) Tbk	41	TLKM	Telekomunikasi Indonesia (Persero) Tbk

5. CONCLUSIONS

Classifying of companies listed in LQ45 resulted in two cluster. The first is the efficient company, which consists of six companies, namely; Astra International Tbk, Bank Central Asia Tbk, Bank Negara Indonesia (Persero) Tbk, Bank Rakyat Indonesia (Persero) Tbk, Bank Mandiri (Persero) Tbk dan Telekomunikasi Indonesia (Persero) Tbk. The second is not efficient companies; that are Astra Agro Lestari Tbk, Adhi Karya (Persero) Tbk, Adaro Energy Tbk, Aneka Tambang (Persero) Tbk, AKR Corporindo Tbk, Alam Sutera Realty Tbk, Bank Tabungan Negara (Persero) Tbk, Global Mediacom Tbk, Bumi Serpong Damai Tbk, Charoen Pokphand Indonesia Tbk, Gudang Garam Tbk, Hanjaya Mandala Sampoerna Tbk, Indofood CBP Sukses Makmur Tbk, Vale Indonesia Tbk, Indofood Sukses Makmur Tbk,Indocement Tunggal Prakasa Tbk, Jasa Marga (Persero) Tbk, Kalbe Farma Tbk, Lippo Karawaci Tbk, Matahari Department Store Tbk, PP London Sumatera Tbk, Media Nusantara Citra Tbk, Matahari Putra Prima Tbk, Hanson International Tbk, Perusahaan Gas Negara (Persero) Tbk, Tambang Batubara Bukit Asam (Persero) Tbk, Pembangunan Perumahan (Persero) Tbk, Pakuwon Jati Tbk, Surya Citra Media Tbk, Siloam International Hospitals Tbk, Semen Indonesia (Persero) Tbk, Summarecon Agung Tbk, Sri Rejeki Isman Tbk, Sawit Sumbermas Sarana Tbk, Tower Bersama Infrastructure Tbk, United Tractors Tbk, Unilever Indonesia Tbk, Wijaya Karya (Persero) Tbk, and Waskita Karya (Persero) Tbk.

REFERENCES

Anderberg,M.R. (1973). *Cluster Analysis for Application*. Academic Press, New York.

Boja, C. (2011). Cluster Models Factors and Characteristcs. International Journal of Economic Practices and Theories, Vol. 1, No. 1, 2011 (July)

Hery. (2016). *Analisis Laporan Keuangan*. PT Grasindo, Jakarta.

Johnson, R.A. and D.W Winchern. (1998). *Applied Multivariate StatisticalAnalysis Ed – 4*. Prentice Hall Inc, New Jersey.

Kuah, A. T. H. (2002). Cluster Theory and Practice: Advantage for the Small Business Locating in a Vibrant Cluster. Journal of Research in Marketing and Entrepreneurship, Volume Four, Issue 3, 2002.

Lebart,L.,A.Morineau, dan L.M.Warwich. (1984). *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices. John Willey and Sons*, New York

This page intentionally left blank

IPS09: Sectoral Financial Account For Monetary Policy Making

German Households Portfolio Decisions and Balance Sheet Dynamics from a Monetary Policy Perspective
Christine Annuß

Revisions of Estimation Methods for Cash Holdings by Households and Corporations in Japan's Flow of Funds Accounts
Sayako Konno, Naoto Osawa, Ai Teramoto

Upgrading Financial Accounts with Central Balance Sheet Data – What's in It for Central Banks' Policy?
João Cadete de Matos, Lígia Maria Nunes

Development of the South African Institutional Sector Accounts
Joel Mokoena, Barend de Beer

German Households Portfolio Decisions and Balance Sheet Dynamics from a Monetary Policy Perspective

Christine Annuß
Deutsche Bundesbank, Frankfurt, Germany – christine.annuss@bundesbank.de

Abstract

Financial accounts provide valuable insights into households' financial portfolio behaviour and allow drawing implications with regard to the impact of monetary policy on the macroeconomy through households' balance sheets. The analysis of German households' portfolio decisions in the context of changing portfolio returns and evolving macroeconomic dynamics reveals a persistent liquidity preference and risk aversion, even in the current low-interest-rate environment. In fact, real portfolio returns do not turn out to be a major determinant of households' investment behaviour and the Eurosystem's Asset Purchase Programme seems to have induced household portfolio rebalancing only to a limited extent. Moreover, the reaction of household spending to changes in liquid asset holdings is different from the reaction to changes in less liquid asset holdings, debt or housing wealth. Therefore, these components of household balance sheets are likely to impact on consumption behaviour in different ways. Analyses suggest that monetary policy transmission through the household sector in Germany is different from that in Anglo-Saxon economies: Other things equal, rises in house prices reduce aggregate consumption, and lower interest rates are likely to be less effective in stimulating household spending.

Keywords: Wealth effects; Rate of return; Low interest-rate environment; Portfolio choice; Consumption.

1. INTRODUCTION

Nominal interest rates have fallen to historically low levels since the outbreak of the great financial crisis, particularly for bank deposits. Their nominal interest rates are now at zero or in some cases even in negative territory. More specifically, while many banks have increased the fees they charge for banking services, a few of them even started charging negative nominal interest rates on savers' transferable deposits. This has affected German households and their financial portfolios, as they traditionally hold a significant portion of their financial assets in the form of deposits. Thus, the property income they earn on a substantial part of their portfolios has shrunk to unprecedented levels in recent years. These developments have caused an ongoing public debate whether saving is still worthwhile when interest rates tend to zero or whether households could just as well lower their saving efforts[1].

In order to shed some light on this issue, however, a comprehensive conception of households' financial portfolio returns, which takes into account the complete portfolio structure including all major non-deposit investment instruments as well as controlling for the influence of inflation on purchasing power, is crucial. This real total portfolio return allows for the assessment of its determinants, including the macroeconomic and monetary policy environment as well as households' portfolio decisions. Moreover, the degree to which portfolio returns influence portfolio decisions can be analysed. This influence, as will be demonstrated, is less clear-cut due to additional, more important drivers like wealth, preferences and demographic factors which seem to play a crucial role in determining portfolio decisions.

The current low-interest-rate environment in Germany, moreover, is associated with increases in house prices. This raises the question of whether there is a transmission of these price increases via consumption on economic activity similar to the pre-crisis developments observed during the house price booms in the US, UK, Spain or in Ireland. While wealth levels and their prices generally affect

1 For examples of this public debate refer to Bindseil et al. (2015)

households' spending behaviour, the direction and size of this effect is conceptually ambiguous. This financial accelerator effect is likely to be time-varying, country-specific and to depend on factors such as the portfolio structure, credit market architecture and the general institutional setting. In this respect, Germany turns out to differ significantly from the Anglo-Saxon economies.

As financial accounts provide in-depth insights into households' investment and financing behaviour and as they display the relevant developments in the portfolio structure, they are a valuable and readily accessible source of information to address the issues raised above. German financial accounts were first published in 1955 (Stöß (2009) describes their evolution and methodology) and are regularly released on an annual and quarterly basis in different formats, among which Deutsche Bundesbank (2016)[2]. To the present day, financial accounts in Germany, which are compiled according to the ESA 2010, are used for economic and monetary policy analyses on a regular basis, such as in Deutsche Bundesbank (2012, 2015), Geiger et al. (2016) and Annuß and Rupprecht (2016). Drawing on some of these analyses, this paper illustrates the links between household balance sheets and the macroeconomy.

2. HOUSEHOLDS' REAL PORTFOLIO RETURNS AND PORTFOLIO DECISIONS

The nominal interest rate on bank deposits – which denotes the income from a financial investment in the form of interest payments in relation to the nominal value of the investment – is a particularly easy notion to grasp for private savers. However, nominal rates neglect the variation in purchasing power of interest income due to price level changes. Therefore, a more appropriate indicator for the rate of return on deposits is the real interest rate, which approximates the difference between the nominal interest rate and the expected rate of inflation[3].

As interest is just one form of income from a financial investment, other components need to be taken into account as well when considering portfolio returns. As opposed to income from bank deposits (and also from claims on insurance corporations and pension funds), income from securities is driven to a large extent by changes in their prices. Shares and certain types of investment funds, in addition, usually also accrue dividends. Together, these income components, as a percentage of the amount invested, represent the return on an investment. Adjusting the nominal return for inflation expectations yields the real investment return. However, data on households' inflation expectations is scarce and the intended investment horizon in most cases is unknown. Therefore, the realised inflation rate is used as an approximation in the remainder of this paper.

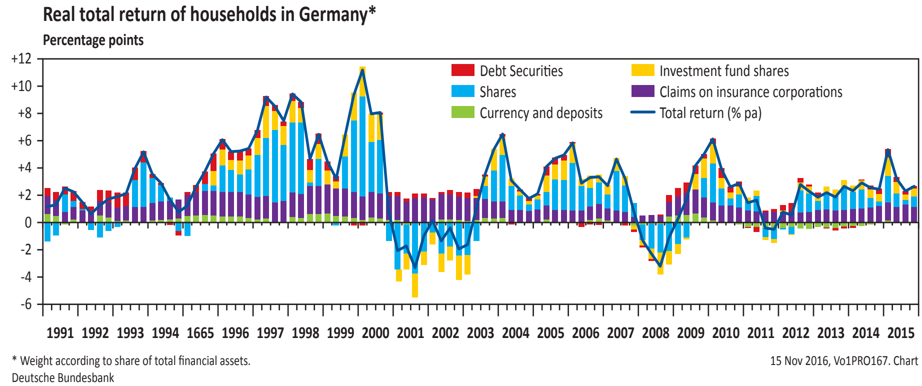


Figure 1. Real Total Return of Households in Germany

---

2 Time series can be downloaded from the Bundesbank website via the following link: http://www.bundesbank.de/Navigation/EN/Statistics/Time_series_databases/Macroeconomic_accountin g_systems/macroeconomic_accounting_systems_node.html?anker=GESAMTFINANZ

3 This can be expressed in terms of the equation $r_t = i_t - \pi_{t+1}$, which is based on the work of Irving Fisher, according to which the expected one-period real return equals the one-period nominal return adjusted for expected inflation over this period

Based on this concept, real returns on the financial assets held by households in Germany were calculated, taking into account the structure of households' financial portfolio. Details about the underlying assumptions, data sources and calculations can be found in Deutsche Bundesbank (2015). Weighting the individual asset returns with their respective portfolio share results in the real portfolio return, which is displayed in Figure 1. Contrary to what part of the public debate suggests, the real portfolio return has not been extraordinarily low, even in the current low-interest-rate environment: For the year 2015 the total real return amounted to 3.4 %, which lies above the long-term average. It should be noted that these results are merely backward-looking and do not attempt to predict future developments should the low-interest-rate environment persist. Nor do these results reveal anything about the distribution at household level.

The question of how households' financial investment behaviour is influenced by interest rates (or more generally: returns) in theory can be approached starting from the theory of money demand in conjunction with basic mechanisms posited in portfolio theory developed by Markowitz and Tobin as is done for example in De Santis et al. (2013) and De Bondt (2009). As households seek to optimise their portfolio structure and as returns affect the utility that a household can derive from holding a financial asset, the demand for certain types of financial asset at the microeconomic level entails a comparison of the anticipated risk-return ratios of different assets. A change in returns disturbs the portfolio equilibrium and thus induces portfolio rebalancing in order to restore the optimal portfolio structure.

Empirically, the link between returns and the portfolio structure at the macroeconomic level can be analysed in the context of a Financial Almost Ideal Demand System, which is employed for example by Ramb and Scharnagl (2011) or Avouyi Dovi et al. (2013). For German quarterly data from 1980 to 2015, such a multivariate system of demand equations, where different forms of investment are modelled as a percentage of financial assets, suggests that real returns do not have a clear-cut or substantial impact on the portfolio structure. While such econometric analyses at the macroeconomic level have methodological limitations and should be interpreted with caution[4], the analysis indicates that other factors, such as the age structure or the level of wealth, exert a comparatively clear influence on portfolio structure.

Similarly, Figure 2 does not show any major portfolio shifts during the last years which could be clearly attributed to portfolio returns. The sharpest increase is found in the portfolio share of transferable deposits, and a certain rebalancing from time and savings deposits seems to have been in place since 2009. This hints at a pronounced liquidity preference which seems to be a major driver of households' investment decision. Similarly, claims on insurance corporations and pension funds gained importance, albeit to a smaller degree. The fact that these claims tend to be perceived as comparably safe assets, taken together with the build-up of transferable deposits, suggests a marked degree of risk aversion. This corresponds with the rather weak engagement in capital markets, where only investment fund shares have recorded perceptible inflows since 2013, while the direct demand for shares remained subdued. A reason for this may be that households in Germany prefer to leave investments into what they see as potentially riskier assets to typically better informed professional investors. The demand for investment fund shares is likely to be related to a heightened awareness of returns in the low-interest-rate environment, despite households' risk aversion. Lastly, debt securities have been sold continuously, not least since the start of the Eurosystem's Asset Purchase Programme. As households' sales of debt securities have not significantly increased since the onset of the purchase programme, and as no major portfolio shifts have occurred, its impact on German households' financial portfolios is likely to be limited.

4 For instance, estimates of the partial effects are inaccurate as yields have a high degree of multicollinearity despite various model specifications. In addition, changes in the variability of the yields and/or in risk aversion overall are not modelled here

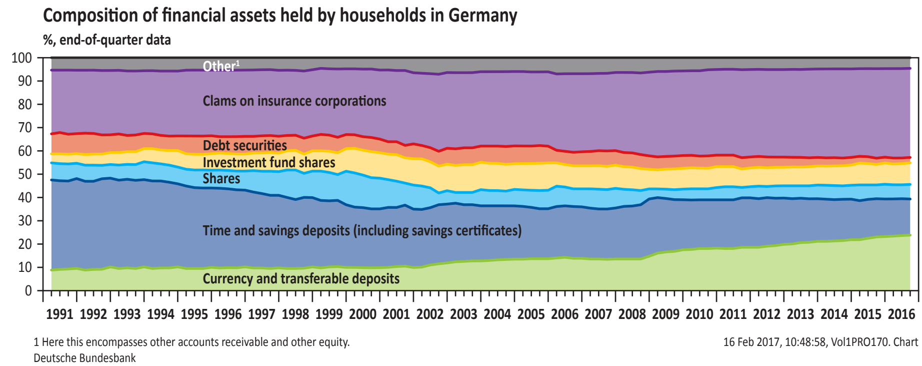


Figure 2. Composition of Financial Assets Held by Households in Germany

3. BALANCE SHEET DYNAMICS AND SPENDING BEHAVIOUR

Apart from the decision for or against certain financial assets, returns and – more generally – monetary policy can also exert an influence on households' decision on the income share which is used for saving. Figure 3 depicts the German household saving ratio in comparison to the real portfolio return, which at a first glance do not display any relationship with each other. In spite of the low-interest-rate environment, there has not been a distinct trend in the saving ratio since 2009 and its movements were confined to a range of only 1 percentage point. In fact, the correlation coefficient of the saving ratio and the portfolio return amounts to only 0.05, implying that there is no strong statistical relationship between those two variables.

Households are faced with the decision to use their income either for saving or for spending. This decision theoretically depends – apart from portfolio returns – on their intended consumption path as well as the structure and development of their balance sheets, where different kinds of assets and liabilities can have different implications for the reaction in saving/spending. Besides the financial assets discussed so far, balance sheets include housing assets, which played a major role in the context of the house price boom in the US, in Ireland or in Spain in the run-up to the financial crisis. Particularly in those economies, sharp increases in house prices were transmitted into economic activity via household spending (among other aspects). Reasons for this were, among others, that the share of households who own their residence was larger than for example in Germany and that banks' lending standards reflected house price developments and therefore were comparably loose, which allowed for home equity withdrawal on the part of households and additionally fuelled economic activity especially in the construction sector. Taken together, house prices formed an important part of the financial accelerator.

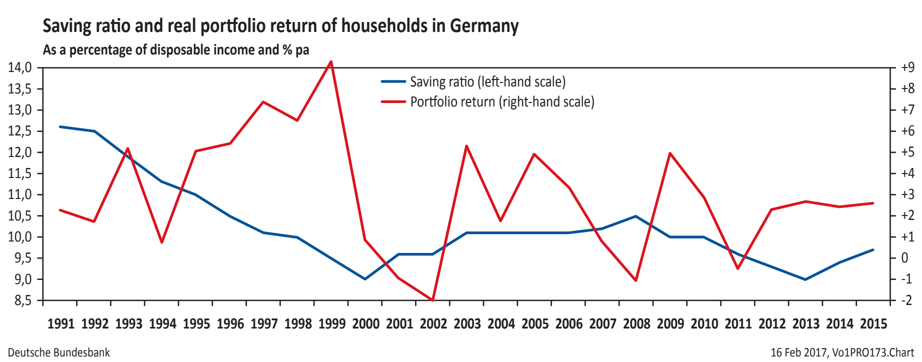


Figure 3. Saving Ratio and Real Portofolio Return of Households in Germany

House prices in Germany have increased markedly since 2010 and especially in urban agglomerations they resulted to some extent in an overvaluation of housing, as shown for example in Kajuth et al. (2013), Deutsche Bundesbank (2013) or Deutsche Bundesbank (2017). The financial accelerator ensuing from house prices can generally be influenced by the credit market structure, which in Germany differs substantially from that in the Anglo-Saxon economies, as detailed in Geiger et al. (2016). In particular, home equity withdrawal in Germany is effectively absent and lending standards for housing loans such as loan-to-value ratios tend to be more conservative. In this setting, an increase in house prices at a given income and given expectations with regard to income growth may imply that households who are willing to acquire owner-occupied housing need to increase their saving in order to honour the initial down payments. Similarly, rent payers anticipating higher rents in the future may react by cutting down on spending. This reasoning, which corresponds to a conventional income effect, contrasts with predictions based on substitution and wealth effects. In a six-equation system with German data for the period of 1981-2012, Geiger et al. (2016) find evidence for such an effect, indicating that increases in house prices in Germany do not translate into higher consumer spending. They attribute this relationship to the characteristics of the general institutional setting as well as the credit market structure. Closely related, given the high share of liquid assets in households' portfolio, they find that lower interest rates are less effective than in the US in directly stimulating household spending. As the period under review only lasts until 2012 and as such relationships are likely to be time-varying, the transfer of the results to the current situation should, however, be undertaken with caution.

Geiger et. al. (2016) further analyse the role of changes in other balance sheet components in household spending. In particular they present evidence that liquid assets among German households assume the function of a buffer stock and therefore can serve the precautionary saving motive, especially if access to unsecured credit is constrained and if risk aversion is high. This means that households in Germany build up liquidity buffers in more prosperous times which are used for consumption smoothing during temporary downturns. This matches the current liquidity preference mentioned above, which – compared to many other European countries – turns out to be rather strong.

4. CONCLUSION

With nominal interest rates at historically low levels, the income households in Germany earn on a substantial part of their portfolios has shrunk to unprecedented levels in recent years. Taking into account households' overall portfolios as well as the effect of price level changes on purchasing power, it can be shown, in contrast, that total real portfolio returns have not been extraordinarily low, even in the current low-interest-rate environment. Analyses based on German financial accounts data, however, suggest that portfolio returns are not a major determinant of households' investment behaviour, while other factors such as wealth, preferences and demographic factors seem to be more important drivers. Moreover, the current low-interest-rate environment in Germany is associated with increases in house prices which – in theory – can be transmitted into economic activity via household spending. For Germany, however, there is no clear evidence of financial accelerator effects as observed during the house price booms in some Anglo-Saxon economies or Spain. In particular, increases in house prices in Germany do not translate into higher consumer spending. This is likely to be due to differences in country-specific factors such as the household portfolio structure, the credit market architecture and the general institutional setting. Taken together, these findings imply that the direct transmission of monetary policy impulses via the household sector may be less effective in Germany than in some Anglo-Saxon countries. As a consequence, monetary policy seems to affect consumption more strongly by indirect channels (via investment and net exports) thereby altering households' disposable income.

REFERENCES

Annuß, C. and M. Rupprecht (2016). Anlageverhalten privater Haushalte in Deutschland: Die Rolle der realen Renditen. DIW Vierteljahrshefte zur Wirtschaftsforschung 1/2016, 95-109.

Avouyi Dovy, S., V. Borgy, C. Pfister, M. Scharnagl and F. Sedillot (2013). Households' financial portfolio choices: a comparison between France and Germany (1978-2009). In: B. Winkler, A. van Riet and P. Bull (eds.). A Flow of Funds Perspective on the Financial Crisis. Palgrave Macmillan.

Bindseil, U., C. Domnick and J. Zeuner (2015). Critique of accommodating central bank policies and the 'expropriation of the saver'. ECB Occasional Paper 161.

De Bondt, G. (2009). Euro area money demand: Empirical evidence on the role of equity and labour markets. ECB Working Paper 1086.

De Santis, R.A., C.A. Favero and B. Roffia (2013). Euro area money demand and international portfolio allocation: A contribution to assessing risks to price stability. Journal of International Money and Finance 32, 377-404.

Deutsche Bundesbank (2012). Long-term developments in corporate financing in Germany – evidence based on the financial accounts. Monthly Report, January, 13-27.

Deutsche Bundesbank (2013). The determinants and regional dependencies of house price increases since 2010. Monthly Report, October, 13-29.

Deutsche Bundesbank (2015). German households' saving and investment behaviour in light of the-low-interest-rate environment. Monthly Report, October, 13-31.

Deutsche Bundesbank (2016). Financial accounts for Germany 2010 to 2015. Special Statistical Publication 4, May 2016.

Deutsche Bundesbank (2017). Housing prices in Germany in 2016. Monthly Report, February. Forthcoming.

Geiger, F., J. Muellbauer and M. Rupprecht (2016). The housing market, household portfolios and the German consumer. ECB Working Paper 1904.

Kajuth, F., T.A. Knetsch und N. Pinkwart (2013). Assessing house prices in Germany: evidence from an estimated stock-flow model using regional data. Bundesbank Discussion Paper 46/2013.

Ramb, F. and M. Scharnagl (2011). Households' portfolio structure in Germany: Analysis of financial accounts data (1959-2009). ECB Working Paper 1355.

Stöß, E. (2009). Gesamtwirtschaftliche Finanzierungsrechnung der Deutschen Bundesbank. In: K. Voy (ed.). Kategorien der Volkswirtschaftlichen Gesamtrechnungen. Marburg, Metrop

Revisions of Estimation Methods for Cash Holdings by Households and Corporations in Japan's Flow of Funds Accounts[1]

Sayako Konno
Bank of Japan, Tokyo, Japan – sayako.konno@boj.or.jp
Naoto Osawa*
Bank of Japan, Tokyo, Japan – naoto.oosawa@boj.or.jp
Ai Teramoto
Bank of Japan, Tokyo, Japan – ai.teramoto@boj.or.jp

Abstract

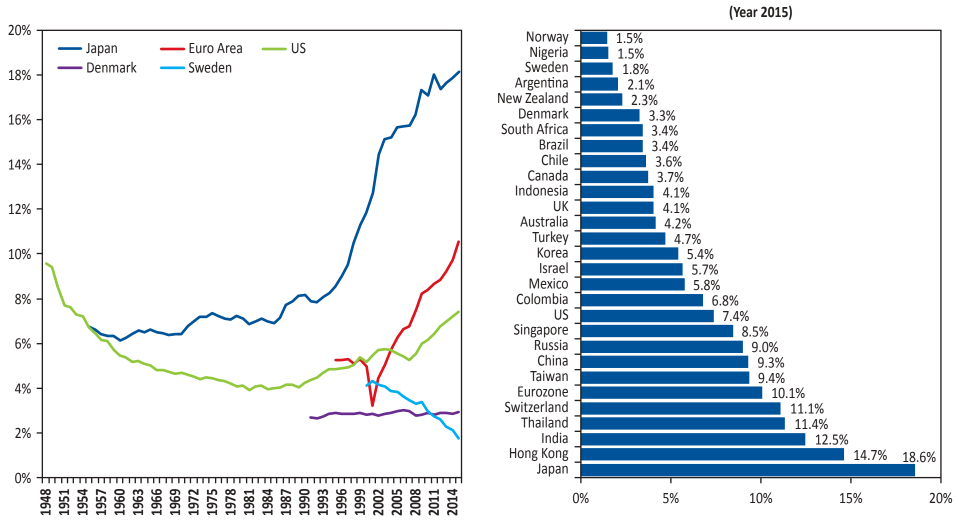
In recent years, cash holdings of the private sector in Japan have risen remarkably. While interest rates remain at a historically low level, the low opportunity cost of cash holdings appears to raise the substitutability of cash and bank deposits. Given the different motivations behind the cash holdings of households and corporations, Bank of Japan's Flow of Funds Accounts (FFA) publishes the amount of cash holdings — with the detail of 50 sectors and 57 transaction items — as an independent transaction item. These statistics of cash holdings across different sectors are attracting increasing public attention and analysis. Given data limitations, however, the FFA's previous estimation method applied the ratio for transferable deposits of the two sectors to the amount of cash holdings, most likely overestimating the cash holdings of corporations and in turn underestimating that of households. In an attempt to improve estimation accuracy, given that raw data and surveys regarding corporations are more readily available than for households, this paper proposes and discusses three alternative estimation methods of the cash holdings held by corporations, estimating household cash holdings as a residual. The first method uses total cash holdings and deposits data in the survey for corporations, subtracting deposits data in the deposits statistics after adjusting for differences in definitions of the two statistics; the second method uses a sole proprietorship survey — which, despite its significantly limited coverage, publishes cash holdings data — applying the ratio of cash holdings to sales proceeds for sole proprietorship to gross up with sales proceeds of corporations; and the third method uses raw data and anecdotal evidence of sales proceeds and cash holdings in the cash register, obtained from cash collection and delivery services companies.

Keywords: Substitutability of currency and deposits; Opportunity cost of currency; Low interest rates.

1. INTRODUCTION

In recent years, the amount of currency outstanding in the world's economies has diverged dramatically (Figure 1). On the one hand, in Scandinavian countries, notably in Sweden, the ratio of currency to nominal GDP has been low and declining, reflecting the attempt of financial institutions to reduce cash management costs and also the government's attempt to curtail illegal activities. On the other hand, in many other advanced economies such as the US, Euro Area and Japan, the ratio has been on the rise amid declining interest rates. Compared with other countries, this is particularly true of Japan, where the ratio has been sharply rising for a long period of time, apparently reflecting a more extended period of low interest rates that has resulted in the lower opportunity cost of cash and in turn a higher substitutability of cash and deposits.

1 The authors are grateful to colleagues at the Bank of Japan for their comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Bank of Japan or its staff members.



(Source) Rogoff (2016), UN, Denmark National Bank

Figure 1. Currency – GDP Ratio

In order to analyze these remarkable recent developments in Japan, there has been an increasing demand on statisticians from statistics end users to produce more disaggregated data, such as by sector, given for example that non-financial corporations (NFCs) and households are considered to have differing motives to have cash holdings. On the one hand, while industries such as retail trade, restaurants and amusement services hold cash with a transaction motive, others such as the manufacturing industry hold cash at a minimum level in principle, predominantly settling their transactions by trade credits. On the other hand, households hold cash for a transaction as well as a saving motive, both of which are in general stronger than that of NFCs.

Nonetheless, in attempting to respond to the statistics users' demand, the more disaggregated data statisticians explore, the more data limitations they usually encounter. In general, in most countries, the issuance of currency is easily obtained through the liabilities side of the central bank balance sheet while the holdings amount is also easily obtained through the assets side of the financial corporations' balance sheet. However, for sectors such as NFCs and households, data is not as readily available. In fact, until it was revised in September 2016, the Flow of Funds Accounts (FFA) of Japan, compiled and published by the Bank of Japan (BOJ) (which includes 50 sectors and 57 transaction items), estimated cash holdings of households (C_{HH}) and NFCs (C_{NFC}) by applying the ratio of transferable deposits (about 7 to 3) to proportionally divide the residual amount of outstanding cash (C_{CB} - C_{BK}). This residual amount is calculated by subtracting that of sectors principally including financial corporations from the issuance of currency amount, as depicted by the orange area in Figure 2[2]. While issuance of currency was valued at 100.3 trillion yen at the end of March 2016, the amount of households was valued at 60.7 trillion yen and that of NFCs was at 26.1 trillion yen.

2 Using notations in Figure 2, cash holdings of households are $C_{HH} = 0.7*(C_{CB}-C_{BK})$; those of NFCs are $C_{NFC} = 0.3*(C_{CB}-C_{BK})$. Figure 2 depicts asset-liability relationships of currency and deposits among different sectors: currency issued by central bank (C_{CB}) is held by commercial banks (C_{BK}), non-financial corporations (C_{NFC}), and households (C_{HH}); and deposits of non-financial corporations (D_{NFC}) and of households (D_{HH}) appear on the liabilities side of commercial banks (D'_{NFC}, D'_{HH}).

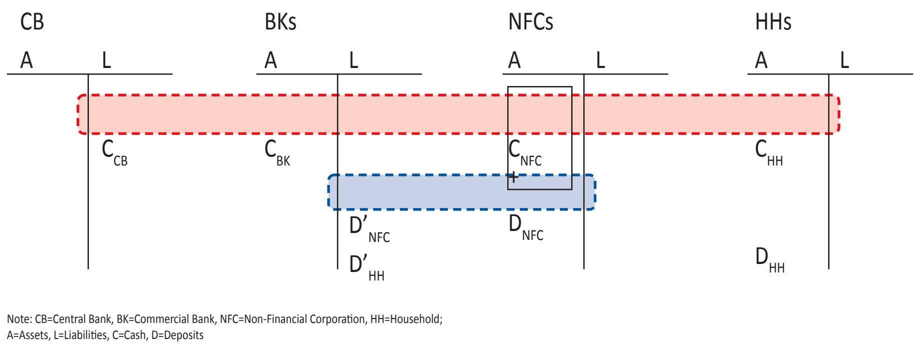


Figure 2. Cash and Deposits on Balance Sheet by Sector (Simple Illustration)

In an attempt to improve the estimation accuracy in the FFA, this paper proposes and discusses three alternative estimation methods of NFC cash holdings, all of which estimate cash holdings of households as a residual. These estimation methods make use of source data and anecdotal evidence regarding NFCs that are more readily available than that of households. The rest of the paper is organized as follows. Section 2 briefly explains the three alternative estimation methods. Section 3 presents estimation results and remaining methodological issues. Finally, Section 4 offers concluding remarks.

2. THREE ALTERNATIVE ESTIMATION METHODS

2.1. Use of Corporations' Financial Statements Statistics

In many countries, surveys of NFCs measure the sum of currency and deposits, but not currency itself. The liabilities side of financial institutions (mainly commercial banks), however, shows deposits by different individual sectors such as households and NFCs. The first alternative estimation method uses total holdings of cash and deposits (C_{NFC} + D_{NFC}) (from the survey for NFCs "Financial Statements Statistics of Corporations by Industry" published by the Ministry of Finance), and subtracts deposits (D'_{NFC}) (from "Amounts Outstanding of Deposits by Depositor" published by the BOJ), as depicted by blue in Figure 2. The following simple equation describes the estimation method:

$$C_{NFC} = [C_{NFC} + D_{NFC}] - D'_{NFC}$$

Note that insufficient adjustments for differences in definitions between the two sets of statistics as well as estimation errors inherent in the corporation survey — whose samples are used to infer population parameters — could lead to estimation errors of NFC cash holdings.

2.2. Use of Sole Proprietorship Statistics

In many countries, a survey to measure currency itself does not exist. As indicated in the previous subsection, surveys for corporations commonly measure only the sum of currency and deposits. However, in Japan a survey, albeit for sole proprietorship (that is, the "Unincorporated Enterprise Survey" published by the Ministry of Internal Affairs and Communications), measures currency by itself.[3] Taking into account the likely proportional relationship between sales proceeds and cash holdings, the second alternative estimation method, multiplies a) sales proceeds of NFCs at an industry level by b) the ratio of cash holdings to sales proceeds for sole proprietorship at an industry level to derive c) cash holdings

3 The "Unincorporated Enterprise Survey" selects limited samples of about 4,000 enterprises in four industries: manufacturing; wholesale and retail trade; accommodation and food services; and services. Note that these industry categories are applied to respective subcategories or similar ones in the later estimation process.

of NFCs at an industry level. Adding up cash holdings for selected “cash-oriented industries” yields total cash holdings of NFCs.[4] The following simple equation depicts this estimation method:

$${}_iC_{NFC} = {}_i(C/S)_{SP} * {}_iS_{NFC} \; ; \; C_{NFC} = \Sigma_i C_{NFC} \, ,$$

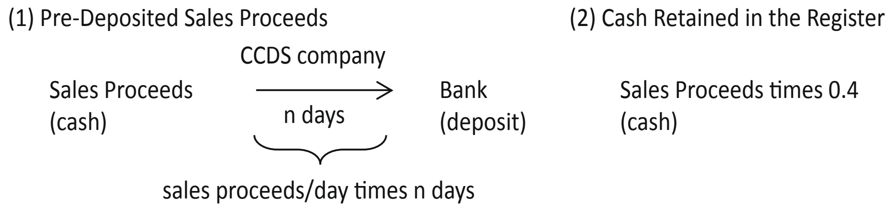
where S=Sales, SP=Sole Proprietorship, i=industry.

Note that this method could lead to estimation bias in both directions. On the one hand, the ratio of cash holdings to sales proceeds for sole proprietorship — which tends to make business transactions by cash rather than other means such as trade credits — is likely to be higher than average NFCs, leading to upward bias. A questionnaire item on cash in the survey for sole proprietorship includes checks as well as cash, also leading to upward bias. On the other hand, excluding “non cash-oriented industries” such as manufacturing from the estimation is likely to be too strong an assumption, leading to downward bias.

2.3. Use of Cash Collection and Delivery Services Companies

A direct survey of NFCs to measure cash holdings, albeit as a natural and promising estimation method, would impose a reporting burden and cost in order to conduct the survey and compile data. However, in Japan, cash collection and delivery services (CCDS) companies collect sales proceeds from corporations and deposit them at the bank, and deliver cash to corporations for transactions. As a less costly method, surveying CCDS companies could provide information regarding cash holdings of NFCs without needing to directly survey individual NFCs.

The third estimation method uses data and anecdotal evidence from CCDS companies; first, for each “cash-oriented industry” it, (1) calculates the interval in days at which CCDS companies collect sales proceeds to be deposited at the bank (sales proceeds collection interval), and (2) assumes the ratio of cash holdings for the purpose of business operations as that stored in the cash register to sales proceeds per day to be equal to 40% (cash retention ratio), as depicted in Figure 3[5]. Second, this method calculates the sales proceeds remaining at corporations before they are deposited (pre-deposit sales proceeds) for each individual “cash-oriented industry” by multiplying the sales proceeds collection interval by sales proceeds per day, which is derived using the corporations survey (“Financial Statements Statistics of Corporations by Industry”). It also calculates the balance in the cash register by multiplying sales proceeds per day by the cash retention ratio, that is, 0.4. Finally, adding up the amount of pre-deposit sales proceeds and cash retained in the cash register across “cash-oriented industries” yields total cash holdings of NFCs.



Note: CCDS = Cash Collection and Delivery Service

Figure 3. Pre-Deposit Sales Proceeds and Cash Retained in the Register

4 Based on information about cash collection and delivery services companies’ client corporations (see subsection 2.3), “cash-oriented industries” including retail trade, accommodation and food services, and amusement services are considered as holding large cash balances for their transactions. In contrast, “non cash-oriented industries” including manufacturing, construction, utilities, and, information and communications industries hold significantly less cash balances than “cash-oriented industries”. Note that “non cash-oriented industries” account for about 50% of all industries in terms of sales proceeds in 2014.

5 While sales proceeds collection intervals differ across industries, an average interval is about 3.3 days.

Note that this method could lead to estimation bias in both directions. On the one hand, as in the second method, excluding "non cash-oriented industries" from the estimation is likely to be too strong an assumption, leading to downward bias. On the other hand, selecting client corporations of CCDS companies tends to result in sampling corporations which make transactions in relatively large cash amounts (selection bias), leading to upward bias.

3. ESTIMATION RESULTS AND REMAINING ISSUES

These three alternative estimation methods, albeit mutually independent, in fact yield similar estimation results. The share of NFCs' cash holdings declines from the pre-revised 30% to 10%, revealing that households, not NFCs, hold most of the cash in the economy; cash holdings of NFCs are 9 trillion yen while those of households are 80 trillion yen in 2015. From a statisticians' point of view, the BOJ revised estimates of cash holdings at a higher level of accuracy, allowing researchers to analyze the outcome of these estimates.

Still, some challenges remain for the statisticians, as a result of directly estimating cash holdings of NFCs and consequently leaving those of households as a residual; the estimate of households may well contain estimation errors. For example, assuming overseas' currency holdings as zero is likely to overestimate households' cash holdings while estimating overseas' holdings poses a significant challenge for statisticians due to data limitations.[6]

A direct survey of households could overcome these challenges arising from possible estimation errors. Nonetheless, an estimate based on a survey for households grossed up to the population yields only about 10 trillion yen in 2014. This appears to be a significant underestimation when compared with 80 trillion yen of above, potentially reflecting underreporting by households and/or sample bias.

4. CONCLUDING REMARKS

This paper proposes and discusses three alternative estimation methods of NFC cash holdings by estimating household cash holdings as a residual, in an attempt to improve the estimation accuracy of Japan's Flow of Funds Accounts. Those methods make use of a combination of published statistics, source data and anecdotal evidence regarding NFCs, which are more easily available than that of households. Albeit mutually independent, all three methods yield similar estimates; the share of NFCs is revised from 30% to 10%. Equivalently, cash holdings of NFCs are 9 trillion yen while that of households are 80 trillion yen in 2015, revealing that households hold most of the cash in the economy.

Amongst these three methods, the BOJ has chosen the second, which uses the sole proprietorship survey. This method applies the ratio of cash holdings to sales proceeds for sole proprietorship, grossing up with sales proceeds of NFCs to derive cash holdings of NFCs. On the one hand, the other two methods pose some practical challenges for statisticians: the third method of using cash collection and delivery services companies would impose a reporting burden and survey costs for periodical estimation updates; and the first method which uses two sets of statistics, that is, corporations' financial statement statistics and deposit statistics, tends to lead to relatively large estimation errors as a result of subtracting two independent statistics whose definitions can never completely converge even after meticulous adjustments. On the other hand, the second method would benefit from the use of published statistics, making periodical updates of estimates relatively less costly.

REFERENCES

Rogoff, K. (2016). The curse of cash. Princeton University Press.

6 While the portion of overseas' cash holdings is estimated to be some 40% in the US, that of Japan is considered to be significantly smaller, given that Japanese Yen is not nearly as internationalized as the US Dollar.

Upgrading Financial Accounts with Central Balance Sheet Data – What's in it for Central Banks' Policy?[1]

João Cadete de Matos*
Banco de Portugal, Lisbon, Portugal – jcmatos@bportugal.pt
Lígia Maria Nunes
Banco de Portugal, Lisbon, Portugal – lmnunes@bportugal.pt

Abstract

Good statistics are a precondition to good policy-making. Thanks to their comprehensiveness and methodological soundness, financial accounts provide a powerful tool in helping to assess the influence of monetary policy actions on the different economic sectors, in a context characterized by their increased financial interconnectedness and threats to financial stability. To achieve fully integrated and consistent financial accounts, *Banco de Portugal* benefits from the richness in statistical content of its Central Balance Sheet Database which is based on census data submitted by virtually all resident corporations through the so-called *IES – Informação Empresarial Simplificada* (literally meaning "Simplified Corporate Information") reporting scheme. Based on our experience, this paper aims to illustrate the advantages and potential uses of corporate accounting data by financial accounts compilers, namely in the processes of improving the consistency and for data quality control purposes.

Keywords: Accounting information; Micro databases; Sectoral accounts; Non-financial corporations.
JEL Classification: E52; E58

1. INTRODUCTION

The global financial crisis of 2008 highlighted the need to better identify the build-up of risks in the financial sector and to understand financial connections among the resident sectors of an economy and between them and their counterparties in the Rest of the World.

Under this context, several initiatives were taken aiming the enhancement of the availability of financial statistics, especially of those related to financial stability issues and systemic risk assessment. This effort was, from the onset, clearly inserted within the general financial accounts framework through the encouragement to the development of integrated sectoral accounts and the compilation of flow of funds schemes (*vd.* recommendation 15 of the G20 data gaps initiative). Financial accounts, an integrated part of the national accounts, are a simplified statistical representation of the financing structure and net financial assets of the various institutional sectors which allow an overview of the uses of the financial surpluses and the way deficits are financed. Complemented with counterpart information, from-whom-to-whom matrices become possible to be compiled. On the basis of these matrices, it is additionally possible to build flow of funds schemes recognised by showing the interlinkages between institutional sectors and, therefore, as a powerful tool to support decision making processes at a macroeconomic level.

The definition of such ambitious statistical output, given the demand for complete information for all institutional sectors, financial instruments and counterparts, was accompanied by the recognition of the challenges imposed by the lack of so detailed data, especially for some institutional sectors, as the Non-Financial Corporations (NFC) and the Households (HH) (Tissot, 2016).

As a way to overcome this concern, the collection of micro data has been increasingly encouraged and their benefits are being globally recognised: the information required to compile integrated sectoral

1 The authors would like to thank Carla Ferreira, Fábio Albuquerque and Pedro Cordeiro for their valuable contributions to this paper. This analysis, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or of the Eurosystem

accounts and flow of funds schemes can be easier derived from granular administrative datasets, given that these data sources have generally a good coverage of the relevant economic agents.

At this respect, the Statistics Department has a remarkable experience in managing administrative databases, which are taking on a predominant role in the achievement of fully integrated and consistent national financial accounts. Among them we highlight the Securities Statistics Integrated System (SSIS)– a security-by-security and investor-by-investor database of both securities holdings and issues; the Central Credit Register (CCR) – which contains granular information on e.g. credit exposures; and the Central Balance Sheet Database (CBSD) – an economic and financial database based on annual and quarterly accounting data on individual Portuguese NFC, which will be discussed in greater detail in the Section 3.

The next section of this paper presents briefly the experience of *Banco de Portugal* (the Bank) as a financial accounts compiler and Section 4 addresses the use of CBSD data to the compilation of national accounts.

2. THE PORTUGUESE SOLUTION TO COMPILE NATIONAL FINANCIAL ACCOUNTS

Following a protocol signed in 1998, in Portugal, the responsibility of the production and dissemination of national accounts is shared between the Portuguese National Statistical Institute, for the non-financial accounts, and the Bank, in charge of the financial accounts compilation. This last process is carried out, on a quarterly basis, by a multidisciplinary team denominated *Estrutura de Missão das Contas Financeiras* (EMCF). This successful arrangement was put in place at the Statistics Department by the end of 2009, and has proved the importance of the involvement of all the divisions of the Department. The EMCF is chaired by the National Financial Accounts Head of Unit and encompasses both national financial accounts experts – permanently allocated to these tasks – and experts from the different underlying primary statistics. Within this new format, all team members become stakeholders of national financial accounts statistics and therefore also actively engaged in collectively contributing to the end- product: for instance, experts from the Central Balance Sheet Statistics Unit provide not only primary data but also are specifically responsible for the compilation of the NFC sector account, and more generally co-responsible for national financial accounts (Matos, 2016a).

Financial accounts data include both the financial transactions and stocks of the different institutional sectors. Its compilation is done on a quadruple-entry basis, whereby each transaction is recorded for the two institutional sectors involved and as a change in both assets and liabilities. In practice, this is achieved by constructing highly detailed from-whom-to-whom matrixes with information on creditor and debtor sectors, financial instrument and assets/liabilities. For the flow of funds representation, the availability of the information on a from-whom-to-whom basis is crucial.

The most important internal data required for the compilation of this output are monetary and financial statistics, balance of payments and international investment position statistics, CBDB and securities statistics. Regarding external sources, information for general government accounts is one of the main inputs. As mentioned above, micro-databases have an unquestionable valuable to ascertain counterparts and construct from-whom-to-whom matrixes, allowing the Bank to go beyond its statistical reporting obligations.

3. THE CENTRAL BALANCE SHEET DATABASE: THE IES AND THE ITENF

As mentioned, the Central Balance Sheet Database of *Banco de Portugal* is an economic and financial database on Portuguese NFC. The data sources used to feed the CBSD are based on annual and quarterly accounting data on an individual basis.

For annual data the CBSD is based on information on the annual accounts of corporations reported within a scope of Annexes A and R of IES – *Informação Empresarial Simplificada* (literally meaning "Simplifies Corporate Information"). IES was an innovative solution launched in 2007, as a result of a joint effort by four public entities in Portugal: the Ministry of Justice, Portuguese Tax and Custom Authority, the Portuguese National Statistical Institute and the Bank. Formerly, in order to fulfil their statutory obligations, corporations were obliged to remit, in separate and independent reports, nearly the same information about their annual accounts to the aforementioned four public entities, in four different moments in time and according to four different formats. The submitted data was not completely harmonized, once each public entity had different requirements (Figure 1).

IES has brought about several advantages for all stakeholders involved. Firstly, it has contributed significantly to streamline companies reporting requirements, decreasing their reporting burden and also avoiding redundancies, by allowing companies to fulfil all the different reporting obligations through one single paper-free report. The statement containing the annual business accounting data is submitted online by each company, once a year, with a delay of about seven months after the end of the reference period. Secondly, it came to make it simpler to those public entities, since they no longer directly request the annual data included in IES. Data is now more "friendly", i.e., it is now much easier to conduct analysis and guarantee the quality of the data because it is reported online and in a harmonized template. The information collected through IES is chiefly of an accounting nature, based on the financial statements and the respective annexes set out in the accounting standards. Additionally, it also comprises a range of data with further detail on the activity and situation of the corporations, as necessary for statistical purposes.

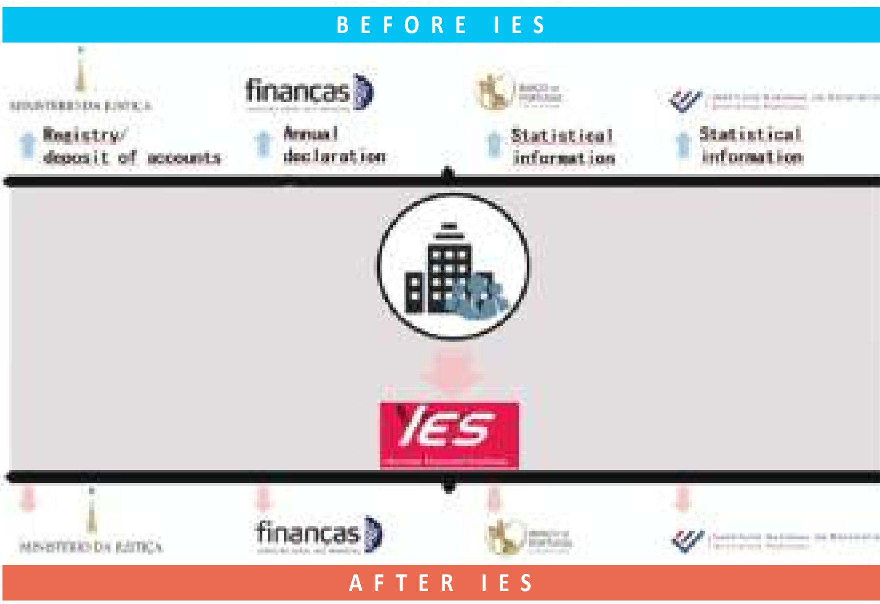


Figure 1. The Reporting of Corporate Information before and after IES

The CBSD annual output includes a very significant observed component obtained via data submitted under IES. In the last years a coverage of about 95% of the total corporations was achieved. This observed component also allows the estimation of the residual component for non-response, which aims to obtain the main variables of the balance sheet and profit and loss account for corporations that have not fulfilled the reporting requirements or in cases of delayed delivery of IES reporting.

The starting point for the treatment of this last component is the information available in the reference population of the NFC sector and in the census databases managed by the Bank. The first step consists in obtaining an estimate for total assets of corporations not reporting to IES, by resorting to IES information on the same corporation for a previous period, or information on that corporation in other *Banco de Portugal* databases (SSIS and CCR). The second step corresponds to estimate a set of relevant information on these corporations not reporting IES, using as ancillary information the NACE – Rev.2 and turnover from the reference population, SSIS and CCR data and the total asset estimates in the previous step. In the end, estimations are calibrated so as to maintain, to the extent possible, the initial classification of the corporations by quantitative stratum (turnover and total assets).

This information is indispensable for the extrapolation process and conciliation between annual and quarterly data. This procedure makes it possible to complete the annual database but, give the small weight of the non-response component, it does not introduce changes in the development of the main indicators observed in responses to the IES.

For quarterly data, the CBSD is feed with information reported through the Quarterly Survey to NFC (here in after denominated ITENF, from the Portuguese designation *Inquérito Trimestral às Empresas Não Financeiras*), a statistical operation jointly developed between *Banco de Portugal* and the Portuguese National Statistical Institute, with a main objective of collecting a range of accounting variables related to the activity and financial situation of a sample of companies. In recent years, significant improvements have been introduced in the methodology associated with the definition of the ITENF's final sample. The current approach, since it does not correspond to a classical sampling design, makes necessary to estimate probabilities of selection of corporations, which are key to calculate the extrapolation factor assigned to each corporation.

The extrapolation procedure is based on the account variables reported by the respondents of the ITENF (Figure 2). In addition, the probabilities of selection used are those previously calculate to determine the extrapolation factor of each corporation, as well as data on the reference population on the NFC sector, updated with information for all relevant variables. This extrapolation procedure makes possible to obtain estimates for variables of interest, for total corporations in the sampling-frame. However, total output for the corporations in the sampling frame does not substantially differ from the population of NFC, in terms of total assets and turnover.

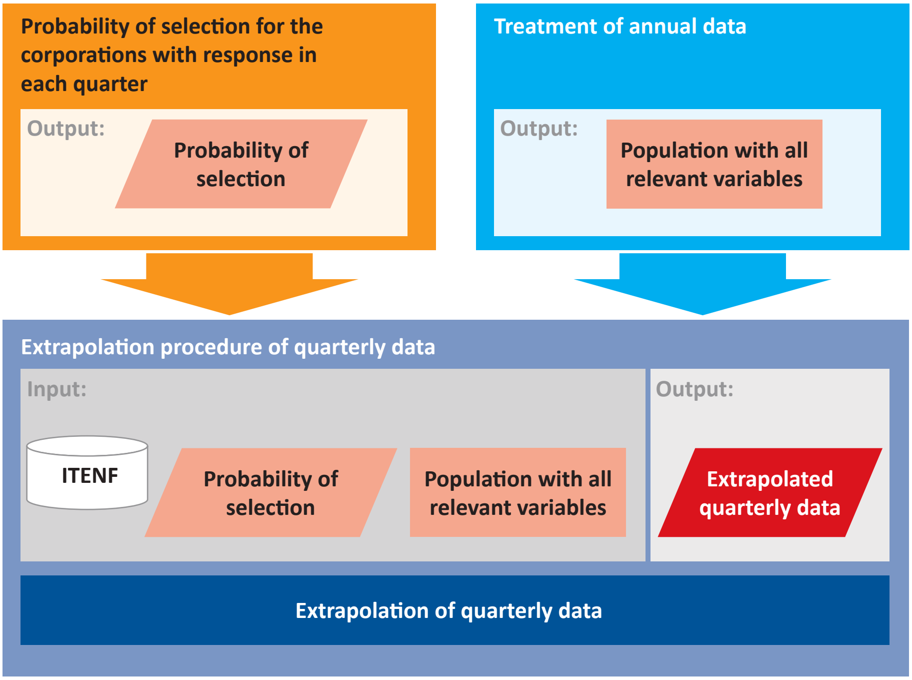


Figure 2. Extrapolation Procedure for Quarterly Data

Extrapolated data are thus used as data for total NFC, which is undoubtedly a great plus to the production of NFC statistics and also to the elaboration of different and flexible statistical products – inter alia, quarterly national financial accounts, which will be discussed in the next section.

Given the existence of both annual and quarterly data sources, different values are obtained at year-end between annual and quarterly indicators. In this context, a conciliation procedure is used, which is known as benchmarking. The data sources used in benchmarking are the outputs of procedures regarding annual and quarterly data. This conciliation method makes it possible to adjust the quarterly time-series obtained by extrapolating ITENF to the annual time-series obtained from IES, which are considered as benchmarks. In order to carry out this adjustment, a set of constraints to be fulfilled by final data are defined:

- Aggregation constraints which ensure that the quarterly time-series is consistent with the annual time-series. They are defined according to the type of variable: for stock variables, the value at the end of the fourth quarter shall be equal to the value at the end of the year and for flow variables, the value of the annual time series shall correspond to the accumulated value in the year for the quarterly variable.
- Contemporaneous constraints of accounting balance which ensures that an accounting balance condition between the balance sheet and the profit and loss account is met in each period. This condition is particularly relevant in the case of variables in the ITENF, where the extrapolation procedure may give rise to imbalances between the balance sheet and the profit and loss account, since the balance sheet variables (stocks) and activity variables (flows) are obtained through the implementation of different procedures. On the other hand, the estimate based on aggregation constraints alone does not ensure intra-annual data balance.
- Assessment of final data: the final result for each aggregate consists in a range of quarterly time series based on a balance sheet and a profit and loss account without imbalances, which combine the annual value obtained in IES with the intra-annual dynamics resulting from the extrapolated ITENF.

4. NATIONAL FINANCIAL ACCOUNTS COMPILATION USING CBSD DATA

The information derived from the above described extrapolation procedure has assumed an increasingly role in the compilation of quarterly national financial accounts, as a direct source for NFCs account and an indirectly contributor to the HH account compilation.

There are two main complementing elements to compile NFCs and HHs accounts: counterpart information and own data sources. Counterpart information refers to the appropriation of information from other sectors, in the cases where it is deemed of a superior quality, and whenever the counterpart is NFC or HH. Typically, both the NFC and the HH sector lie at the bottom of the hierarchical chain of counterpart information. This means that these sectors normally take the information of other sectors as given and incorporate it directly. More specifically, compilation of NFCs' accounts takes on board counterpart information from:

- Financial corporations, i.e., balance sheet statistics from Monetary Financial Institutions (and from Other Financial Institutions);
- General government statistics;
- Rest of the World account, i.e., balance of payments and international investment position statistics.

In other words, NFC compilation takes on board the counterpart information of all other sectors/statistical domains except for HHs. In turn, HH account compilation uses counterpart information of all other sectors. The second approach pertains to the use of data sources which are specific to these two sectors.This is where the information derived from CBSD comes to play in terms of NFCs' account compilation. Finally, the SSIS provides information on securities holdings and issuance for the two sectors.

The interesting element of CBSD data is that it is not only an own source for NFCs, but it also provides indirectly contributes to the HH account compilation as counterpart information, to the extent that some types of operations between the NFCs and the HH sector are collected. There are two ways in which the information used from CBSD to compile the NFCs' account is also relevant for the HH sector:

- It helps identify HH equity stakes in NFCs, i.e., HH equity assets in NFCs' equity liabilities. In a country such as Portugal, where the structure of NFCs is heavily tilted towards small and medium- sized corporations, these figures represent an important share of HH equity holdings as well as of NFCs' equity liabilities.
- Secondly, loans received/granted by NFCs and granted/received by HHs are also collected in IES and are therefore used to determine assets and liabilities of NFCs and HH loans.

5. CONCLUSIONS

The achievement of Portuguese fully integrated and consistent financial accounts, was only made possible due to the richness of the statistical content of the CBSD run by the Bank. This huge potentiality is due to the relevant data on the population of corporations in Portugal for a relatively long period, as a result of the excellent coverage of IES, which allows the Bank to access to new and more complete information on the Portuguese NFC and, more recently, by the appropriate methodology for the selection of corporations within the scope of the ITENF and the respective extrapolation procedure.

Alone or combined with other information, CBSD data also proved to be a great value for pursuing the statistical central bank statutory obligations. The use of this information for the compilation of the NFC and HH sectors' financial accounts is an example of the statistical possibilities of such census information.

REFERENCES

Banco de Portugal (2013). "Statistics on non-financial corporations of the Central Balance Sheet Database: Methodological notes". Supplement to the Statistical Bulletin (2|2013)

Brites, M. (2013). "Efficient ways of dealing with accounting data from enterprises". *Banco de Portugal*

Cordeiro, P., Magalhães, C. & Poiares, R. (2016). "Quarterly time-series from Central Balance Sheet Database". *Banco de Portugal*

Lima, F. & Drumond, I. (2016). "How to keep statistics' customers happy? Use micro-databases!". *Banco de Portugal*

Cadete de Matos, J. (2015). "Using financial accounts to better understand sectoral financial interlinkages". *Banco de Portugal*

Cadete de Matos, J. (2016a). "Innovative solutions in compiling financial accounts". *Banco de Portugal*

Cadete de Matos, J. (2016b). "Using census information on non-financial corporations to compile national financial accounts". *Banco de Portugal*

Cadete de Matos, J. (2016c). "Reaping the benefits of using integrated micro-data for statistical purposes... and beyond". *Banco de Portugal*

Cadete de Matos, J. (2016d). "Brief note about the use of Census information on non-financial corporations to compile national financial accounts". *Banco de Portugal*

Development of the South African Institutional Sector Accounts

Joel Mokoena*
South African Reserve Bank, Pretoria, South Africa – Joel.Mokoena@resbank.co.za
Barend de Beer
South African Reserve Bank, Pretoria, South Africa – Barend.deBeer@resbank.co.za

Abstract

The paper discusses the work undertaken by the South African Reserve Bank to further the development of integrated economic accounts for the South African economy. This is done from the vantage point of national interest but also as part of South Africa's commitment to the G-20 Data Gaps Initiatives. One of the main benefits foreseen from this undertaking is the improved insight it ought to provide into financial vulnerabilities and risks associated with institutional sector balance sheets. It could be a useful tool for identifying potential stress-points in the financial sector as well as the transmission channels through which these risks could be propagated to the real economy. In addition, it also potentially provides a data set that ought to be of value in better understanding the transmission of monetary policy. Much work revolved around the measurement and estimation of the stock of non-financial assets and financial assets and liabilities for the various institutional sector balance sheets with specific emphasis on financial instrument detail and further detailed breakdowns by institutional sectors which were generally lacking. In addition to institutional sectors and financial instruments, further developments entail the presentation of these sets of accounts in terms of currency composition, maturity analysis, and on a from-whom-to-whom basis. The paper examines the concepts and methods used to compile the integrated economic accounts which follow the standard System of National Accounts 2008 framework. It discusses the South African Reserve Bank's approach to the balancing model, including the estimation of nominal holding gains and losses and financial transactions as well as the application of market valuation principles in the financial balance sheets. The public-sector debt statistics is highlighted as an example of valuation issues between nominal and market value - a mix of nominal and market values would lead to financial asset and liability inconsistencies as institutional sector counterparty asset positions are usually recorded at market values in their balance sheets whilst the liabilities of certain financial instruments are more commonly recorded at nominal value in the accounts of the debtor sector. The paper discusses selected results of the analysis of preliminary high level institutional sector balance sheet data – both non-financial and financial positions and broken down by type of non-financial category and financial instrument as well as by institutional sector. The objective of this analysis is to show the estimated size of institutional sectors in the South African economy and their positions as net lenders or borrowers, and financial positions of institutional sectors and the rest of the world on a from-whom-to-whom basis. Thereafter some areas for future development work which will lead to the dissemination of a more complete set of integrated economic accounts for South Africa will be highlighted. This includes the extension of the dataset to 2015 and later, to most recent quarterly data.

Keywords: Integrated economic accounts, Non-financial assets, Financial instruments.
JEL Classification Numbers: E01, E02, H63

1. INTRODUCTION

The South African Reserve Bank established a project to further develop the integrated economic accounts for the South African economy in 2015. This is in the interest of providing reliable and relevant statistics for national policy formulation and to fulfill South Africa's international statistical commitments to the G-20 Data Gaps Initiatives. One of the main benefits foreseen from this undertaking is the improved insight it ought to provide into financial vulnerabilities and risks associated with institutional sector balance sheets. It could be a useful tool for identifying potential stress-points in the financial sector as well as the transmission channels through which these risks could be propagated to the real economy. In addition, it also potentially provides a data set that ought to be of value in better understanding the transmission of monetary policy.

The further development of the integrated economic accounts entails a harmonised set of macroeconomic accounts that link the real and financial spheres of the economy. These additional data sets are intended to provide quarterly data for the period from 2010 to the present. Much work on the development of integrated economic accounts revolved around the measurement and estimation of the stock of non-financial assets and financial assets and liabilities for the various institutional sector balance sheets with specific emphasis on financial instrument detail and further detailed breakdowns by institutional sectors which were generally lacking. In addition to institutional sectors and financial instruments, further developments entail the presentation of these sets of accounts in terms of currency composition, maturity analysis, and on a from-whom-to-whom basis.

This paper draws heavily on the work of de Beer et al. (2016) on the compilation of fully integrated economic accounts for South Africa. Section 2 introduces concepts and methods used to compile the integrated economic accounts. Section 3 discusses selected results of the analysis of preliminary high level institutional sector balance sheet data. Section 4 concludes by discussing some areas for future development work which will lead to the dissemination of a more complete set of integrated economic accounts for South Africa.

2. METHODOLOGY FOR COMPILING THE INTEGRATED ECONOMIC ACCOUNTS

In the integrated economic accounts, the real economy is represented by the current and capital accounts as well as the non-financial asset and accumulation accounts. The current account comprises the production, generation and allocation of income, distribution and redistribution of income accounts, as well as its utilisation. Currently only these accounts and high level financial assets and liabilities and non-financial assets of households are published on an annual basis by the South African Reserve Bank. However, this information needs to be extended in terms of institutional sector breakdown as well as to a quarterly frequency. The coverage of the financial sphere of the economy is obtained through the development of a full set of financial balance sheets and accumulation accounts.

2.1. Institutional sectors

One of the anchor parameters underlying the project is institutional sector delineation. The institutional sector delineation applied in the integrated economic accounts is obtained from a combination of delineation contained in the SNA 2008 as well as national requirements. Currently, various data sources are used for the different institutional sectors, except for the private non-financial corporate sector which is measured using derived data from the other counterparty sectors.

2.2. Non-financial assets

Non-financial assets consist of both produced and non-produced assets. The produced non-financial assets in turn comprise fixed assets used in production, inventories and valuables - see Table 2. Valuables are not included in the scope of the project due to data constraints. Non-produced non-financial assets consist of natural assets and assets created through legal agreement from which the owners can extract economic benefits. Natural assets include land and mineral reserves whereas leases, licenses and permits are created through legal agreements. Currently, only the market value of land underlying dwellings, buildings other than dwellings and other structures is reported.

The market value of land - whether residential, industrial or different types of farmland - differs markedly and sufficient price data are not available to make accurate estimations. The market value of South Africa's various national parks and other recreational holiday retreats as well as other land belonging to general government further complicates estimations. Data constraints also inhibit the estimation of mineral and energy reserves.

2.3. Financial instruments

Financial assets and liabilities are valued at current market prices as at the balance sheet date. However, some datasets are only available in nominal terms which pose a problem. Figure 1 illustrates the divergence between nominal and market value due to valuation effects in the stock position of long-term debt securities of national government as an example.

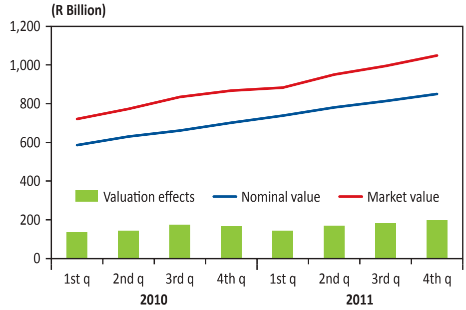


Figure 1.
Outstanding Nominal and Market Value of National Government Long-Term Debt Securities

Unfortunately most of the institutional sector data available only relates to stock positions, in particular in the case of financial assets and liabilities, which places a heavy burden on the South African Reserve Bank to gauge the contribution of valuation effects as opposed to actual transactions and other changes in volume. Future development work will focus in this area.

2.4. Balancing

The accumulation accounts of the financial balance sheets are divided into financial transactions, other volume changes and revaluation[1] accounts. The financial balance sheets and accumulation accounts measurement system has three major components, namely the building block component, the balancing component and the revaluation estimation component. The building block component deals with the source data inputs from contributing institutional sector compilers. Data sourcing is conducted through the utilisation of financial asset and liability templates for each institutional sector which in effect produces unbalanced balance sheets for the institutional sectors and which is based on templates provided by the International Monetary Fund. These financial asset and liability templates are introduced into the balancing system where balancing assumptions are made based on a detailed hierarchy of sources. Once the balancing assumptions have been applied to the unbalanced financial asset and liability templates the balanced financial asset and liability templates for each institutional sector is generated. This is the current status of the project and rendered the data shown and discussed in this paper.

3. PRELIMINARY RESULTS[2]

The further development of the integrated economic accounts into all its dimensions is a project that will stretch over an extended period of time. The first milestone, to develop a high-level quarterly core dataset for 2010 and 2011, has been achieved.

1 The revaluation account is also referred to as nominal holding gains and losses.
2 These are experimental results subject to future adjustment.

3.1. Non-financial assets

Total produced non-financial assets including underlying land in the domestic economy amounted to R9,4 trillion at the end of December 2011 as indicated in Table 1.

Table 1. Non-financial Asset Stock Positions at Market Prices, 31 December 2011 (R millions)

	Non-financial corporates	Financial corporates	General government	Households	Total domestic economy
Total produced assets	3.673.489	294.319	1.660.095	1.905.073	7.532.976
Total fixed assets	3.112.793	278.497	1.656.494	1.872.662	6.920.446
Dwellings	243.334	10.070	237.628	1.660.831	2.151.863
Buildings other than dwellings	733.769	113.287	312.168	66.309	1.225.533
Other structures	983.612	7.198	938.444	68.365	1.997.619
Machinery and equipment	1.042.563	119.189	141.076	62.126	1.364.954
Cultivated biological resources	8.781			14.659	23.440
Intellectual property	100.734	28.753	27.178	372	157.037
Inventories	560.696	15.822	3.601	32.411	612.530
Non-produced					
Total underlying land	563.191	33.113	525.811	760.363	1.882.478
Dwellings	85.424	5.331	114.941	744.668	950.364
Buildings other than dwellings and other structures	477.767	27.782	410.870	15.695	932.114
Total real estate*	2.523.906	163.668	2.014.051	2.555.868	7.257.493
Dwellings	328.758	15.401	352.569	2.405.499	3.102.227
Buildings other than dwellings and other structures	2.195.148	148.267	1.661.482	150.369	4.155.266
Total produced including underlying land	**4.236.680**	**327.432**	**2.185.906**	**2.665.436**	**9.415.454**

*Fixed assets and underlying land.

The total produced fixed assets accounted for roughly 74 per cent of non-financial assets, with the largest contribution, about 39 per cent, originating from non-financial corporations. The household sector dominated non-produced assets, with 98 per cent related to dwellings. Of all four institutional sectors, financial corporations held the smallest portion of non-financial assets in the domestic economy.

3.2. Financial assets and liabilities

The overall size of the market value of the balanced stock of financial assets and liabilities held by both residents and non-residents in the South African economy amounted to R22 trillion as at 31 December 2011 as shown in Table 3, or about seven times gross domestic product recorded for the year. The distribution by value of financial assets and liabilities by institutional sector and financial instrument is shown in Figure 2.

Three important observations emerge from viewing the financial asset and liability data sets. The first is that the overall values of the financial assets and liabilities of the financial corporate sector are fairly similar. The second observation is that the financial liabilities of the household sector are significantly less than its financial assets. Thirdly, the financial liability position of the non-financial corporate sector is markedly larger than its financial asset position.

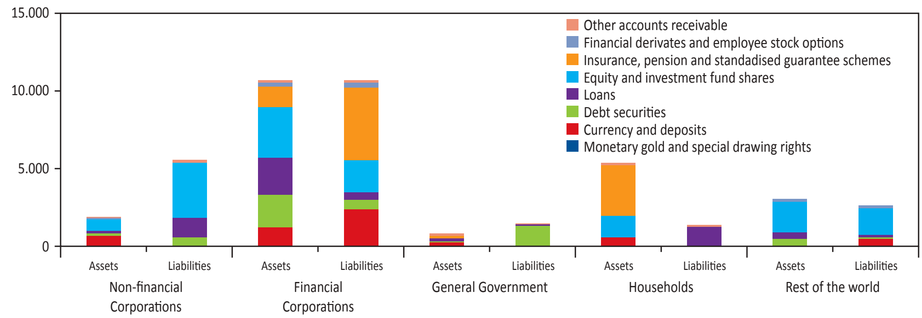


Figure 2. Market Value of Total Financial Assets and Liabilities by Institutional Sector and Financial Instrument, 31 December 2011 (R billions)

Additional insight into the financial instrument-specific make-up of the financial assets and liabilities of the institutional sectors reveal that the majority of the financial assets of the non-financial corporate sector, financial corporate sector, and the rest of the world are concentrated in equity and investment fund shares or units. On the other hand, a significant portion of household assets are in the form of pension claims on the financial sector while the general government has a fairly small financial asset position. On the liability side, both the non-financial corporate sector and the rest of the world had a high exposure to equity and investment fund shares or units; the financial corporate sector to insurance, pension and standardised guarantee schemes; households to loans; and general government to debt securities.

3.3. From-whom-to-whom positions

The analysis done thus far does not reveal any information on institutional sector-by-sector linkages or the extent of these linkages. It is precisely these linkages that are of importance to understand the possible implications for financial stability. The balanced financial asset and liability stock positions of the four domestic institutional sectors as well as the rest of the world, on a from-whom-to-whom basis, as at 31 December 2011 is shown in Table 3. For the resident institutional sectors, aggregated (i.e. non-consolidated) data are presented. This means that intra-sectoral positions, transactions, nominal holding gains and losses as well as other volume changes are not eliminated.

Total financial assets amount to just more than R22 trillion, with total financial liabilities amounting to just less than R22 trillion. The difference of R51 billion, relates to the monetary gold assets of the central bank which does not have a counterparty liability.

The horizontal view of Table 3 provides total financial assets of each institutional sector vis-à-vis the institutional sector against which these claims are held. For example, of the R10,7 trillion financial assets of the financial corporate sector R2,5 trillion represents claims against the non-financial corporate sector and R4,1 trillion represents claims against the financial corporate sector itself. Similarly, the vertical view of Table 3 provides an institutional sector counterparty breakdown of an institutional sector's liabilities.

Important to note is that the rest of the world sector is compiled from the perspectives of foreign economies – which basically results in the inverse of South Africa's international investment position. The financial assets of the rest of the world issued by the rest of the world are not covered and will thus always be zero due to the fact that this data are not relevant from South Africa's perspective. Holdings of financial instruments by non-residents (vis-à-vis resident sectors as debtors) are shown as South African liability positions (rest of the world asset positions) in the rest of the world balance sheet, while acquisitions and disposals by non-residents of financial instruments issued by residents are shown as financial transactions in the rest of the world financial account. Similarly, holdings of financial instruments by residents vis-à-vis non-residents are shown as South African asset positions (rest of the world liability positions) in the rest of the world balance sheet. Thus, the rest of the world financial asset

Table 2. From-whom-to-whom market value positions of total financial assets and liabilities between resident institutional sectors as well as the rest of the world, 31 December 2011 (R billions)

Institutional sectors		Liabilities by resident institutional sector and residency						
		Non-financial corporations	Financial corporations	General government	Households	Rest of the world	Total	
Assets by resident institutional sector and residency	Non-financial corporations	60	966	15	77	848	**1 965**	**22 020**
	Financial corporations	2 525	4 075	1 002	1 270	1 773	**10 696**	
	General government	162	537	89	74	1	**863**	
	Households	1 070	4 149	102	0	73	**5 394**	
	Rest of the world	1 738	1 007	355	0	0	**3 101**	
	Total	**5 555**	**10 734**	**1 563**	**1 421**	**2 696**	**-405***	
		21 969						51**

* This value is equal to the net international investment position excluding monetary gold; which has no counterparty classification in the financial balance sheets.
** This is the value of monetary gold held by the South African Reserve Bank as at 31 December 2011; it equates to the difference between the balanced assets and liabilities because monetary gold has no counterparty classification in the financial balance sheets.

position of R3,1 trillion and financial liability position of R2,7 trillion in Table 2 reflect the inverse of South Africa's international investment position. The positive difference of R0,4 trillion indicates South Africa's negative international investment position.

4. CONCLUSION

The data shown and discussed in this note are the outcome of the project thus far and should be treated as preliminary and experimental. In addition, it should be noted that certain source data need to be broadened and improved to close material data gaps. Future development work will focus on the utilisation of the balanced asset and liability stock positions to estimate nominal holding gains and losses at detailed financial instrument level where after financial transactions will be derived as the residual. This data set will facilitate detailed from-whom-to-whom accounts of stock positions, transactions (flow of funds), revaluations and other volume changes.

The further development of integrated economic accounts for South Africa is an ongoing process. The first phase, discussed above, has laid the foundation and the subsequent phases will build on this. A first priority is to expand the dataset to cover the period from 2012 to 2016. Further effort will be directed at expanding the institutional sector and financial instrument coverage as well as improving the data coverage and quality. There is a significant advantage in using the sector accounts approach contained in the SNA 2008 framework as it ensures data consistency and international comparability. This allows for a systematic understanding of the important relationships in the South African economy, such as the relationship between economic flows in the real and financial spheres as well as the issue of financial interconnectedness and linkages among the various components in the economic value chain.

REFERENCES

De Beer, B., Kock, M., & Wagner, H. (2016). Note on the further development of integrated economic accounts for South Africa – institutional sector balance sheets. *Quarterly Bulletin*, No. 282, December. Pretoria: South African Reserve Bank: 93–102.

United Nations (2009). System of National Accounts 2008 (2008 SNA), UN, New York.

IPS10: Statistical Theory And Methods

Extracting More Value from Confidentialised Tabular Data

Jarod Y. L. Lee
University of Technology Sydney, Sydney, Australia – Yan.L.Lee@student.uts.edu.au

Abstract

National Statistical Agencies and other data custodians increasingly rely on confidentialised tabular data to provide data access while protecting privacy. Data are confidentialised by randomly perturbing the cells while maintaining the perturbed grand total. This introduces noise, making small cell counts unreliable and limits the ability of analysts to fully unlock the actionable information in the data. In this presentation, we propose a modeling framework that utilizes the concept of statistical sufficiency to safeguard confidential information. We illustrate the methodology via a multilevel analysis of the Australian unemployment data obtained from the TableBuilder, an online dissemination tool by the Australian Bureau of Statistics that allows the public to create user-defined frequency tables from the census data based on a chosen geographical classification. Incorporating the statistical sufficiency modeling framework into such data extraction tools allows more detailed analysis to be carried out at a finer geographical level. This is a joint research with Professor Peter Green, Professor James Brown and Professor Louise Ryan from the University of Technology Sydney.

Keywords: Privacy; Generalized linear models; Generalized linear mixed models; Conjugacy.

Detail-Preserving Unsupervised Ensemble Clustering

Siow Hoo Leong
Universiti Teknolog MARA, Kota Samarahan, Sarawak, Malaysia – leongsh@sarawak.uitm.edu.my

Abstract

In unsupervised ensemble clustering, when the data is generated from different sources, there are a number of issues to be considered in order to preserve the details: how to generate ensembles, what the consensus function is, and how to integrate multiple results. This paper proposes an algorithm that addresses all these issues. It generates probabilistic structure ensembles using mixture model with improvement in initialisation. The consensus of the structure ensembles is assessed through a modified model selection criterion, which has the advantage of recovering structure details. The overall information is integrated using a mixture model based structure combination function. The performance of the algorithm is evaluated with simulation and real data. It produces promising results and performs better than the existing similar algorithm.

Keywords: Structure ensemble; Consensus function; Model selection; Gaussian mixture model.

This page intentionally left blank

IPS12: Recent Advances in Failure Time Data Analysis

Regression Analysis of Informatively Interval-censored Failure Time Data

Jianguo (Tony) Sun*; Tao Hu; Ling Ma; Peijie Wang; Hui Zhao

Abstract

Interval-censored failure time data occur in many fields such as demography, economics, medical research and reliability, and many inference procedures on them have been developed (Chen et al., 2012; Sun, 2006). However, most of the existing approaches assume that the mechanism that yields interval censoring is independent of the failure time of interest and it is clear that this may not be true in practice. In this talk, we will discuss this latter situation and present some inference procedures for the problem.

Keywords: Interval censored; Informative censoring.

Regression Analysis of Case K Interval-censored Failure Time Data in the Presence of Informative Censoring

Peijie Wang
Jilin University, Changchun, China – wangpeijie@jlu.edu.cn

Abstract

Interval-censored failure time data occur in many fields such as demography, economics, medical research and reliability and many inference procedures on them have been developed (Chen et al., 2012; Sun, 2006). However, most of the existing approaches assume that the mechanism that yields interval censoring is independent of the failure time of interest and it is clear that this may not be true in practice (Ma et al., 2015; Zhang et al., 2007). In this paper, we consider regression analysis of case K interval-censored failure time data when the censoring mechanism may be related to the failure time of interest. For the problem, an estimated sieve maximum likelihood approach is proposed for the data arising from the proportional hazards frailty model and for estimation, a two-step procedure is presented. In the addition, the asymptotic properties of the proposed estimators of regression parameters are established and an extensive simulation study suggests that the method works well. Finally, we apply the method to a set of real interval-censored data that motivated this study.

Keywords: Case K interval-censored data; Informative censoring; Proportional hazards model; Sieve maximum likelihood estimation.

This page intentionally left blank

CPS12:
Health & Social Statistics (2)

Application of Higher Order Markov Chain for Modeling Immunological Status of HIV Infected Patients
Amit Nirmalkar, Mohan Kale, Manisha Ghate, Ramesh Paranjape, Bharat Rewari

Return Time Distribution Based Analysis of Dengue Virus
Trupti Vaidya, Mohan Kale, Vaishali Waman, Pandurang Kolekar, Urmila Kulkarni-Kale

The Association of Road Traffic Injuries with Injury Severity Score at Region 11 Southern Province in Thailand
Natthika Sae-Tae, Sampurna Kakchapati, Apiradee Lim

Application of Higher Order Markov Chain for Modeling Immunological Status of HIV Infected Patients

Amit P. Nirmalkar*
National AIDS Research Institute, Pune, India - anirmalkar@gmail.com

Mohan M. Kale
Savitribai Phule Pune University, Pune, India - mohan.kale80@gmail.com

Manisha V. Ghate
National AIDS Research Institute, Pune, India - mghate@nariindia.org

Ramesh S. Paranjape
National AIDS Research Institute, Pune, India - rameshsp1952@gmail.com

Bharat B. Rewari
National AIDS Control Organisation. New Delhi, India - drbbrewari@yahoo.com

Abstract

The immunological parameter CD4 cell count is a significant predictor of disease progression among Human Immunodeficiency Virus (HIV) infected patients. The main objective of this analysis is to study HIV disease progression and to compute higher order of Markov chains to look for dependency with earlier immunological states. The data on CD4 cell count measurement was extracted from database. The analysis is restricted to patient who have maintained their follow-up for 5 year of study visits with data on CD4 cell count measurement at baseline and at every 6 month time interval. Hence, we have 418 patients' data on CD4 cell count measurement with none of the record missing. While modeling the progression of the disease, we defined the disease in terms of different states i.e. state 1: CD4+ count > 500; state 2: 350 < CD4+ count 500; state 3: 200 < CD4+ count 350; state 4: CD4+ count 200. We have excluded the death state (absorbing state) as the data is not available. The states are assumed to be finite and patient cannot be in more than one state at any one time. The patients can be in any of the state for random amount of time before entering into new state i.e. the points in time when the event is observed or may remain in the same state for indefinite amount of time. A Discrete Time Markov Chain (DTMC) approach was employed to observe CD4 cell count transitions and to estimate transition probability matrix (TPM) using Maximum Likelihood Estimation (MLE). The one step TPM indicates that the probability of patients staying in same state is higher compared to other transitions [State 1 (P_{11}) =77.21%; State 2 (P_{22}) = 45.23%; State 3 (P_{33}) = 52.83% and in State 4 (P_{44}) = 47.60%]. A negligible percent of patient transitions from State 1 to State 4 (P_{14}) were (0.65%) and 1.13% transitions from State 2 to State 4 (P_{24}). A reverse transition could also be observed i.e. the transitions from State 1 to State 2 (P_{12}) are 18.16% and its reverse transitions is from State 2 to State 1 (P_{21}) are 34.97%. Similarly, transitions from State 4 to State 3 (P_{43}) are 41.95% and State 3 to State 2 (P_{32}) are 27.53%. To look for extent of dependency on earlier states, we carried a testing order of a Markov chain analysis. Akaike Information Criterion (AIC) and Likelihood Ratio Test (LRT) is used for the best model selection among the set of models. Results based on likelihood ratio statistic revels presence of higher order Markov dependence ($r > 1$) i.e. the future evolution of disease state is depending on preceding states with the current state based on CD4 cell count categories. The analysis shows that, in this cohort the government Antiretroviral Therapy (ART) access program has been successful in maintaining immunological status of patient.

Keywords: Akaike information criterion, Disease progression, Likelihood Ratio Test, Maximum likelihood estimation

1. INTRODUCTION

Markov model is a useful tool in medical research to study the disease progression see Sonnenberg & Beck (1993). It has application in modeling chronic diseases like hypertension, cancers, diabetes, sepsis and in human immunodeficiency virus infection and acquired immune deficiency syndrome (HIV/AIDS) see Sato & Zouain (2010); Rangel-Frausto *et al*. (1998); Guihenneuc, Richardson, Longini (2000), where the risk is continuous over time and event may happen more than once. These models are particularly suited to analyze ordered clinical processes and can be applied to estimate and predict rates of HIV disease progression. Markov models have been successfully used in the analysis of HIV infection see, for example, Guihenneuc, Richardson , Longini (2000); Longini Jr. *et al*. (1989); Bwayo *et al*. (1995); Hendriks *et al*. (1996); Kousignian et al. (2003) but none of the articles were focused on higher order Markov chain.

In this research, the immunological parameter CD4 cell count is used to study disease progression and to verify whether the future clinical states are dependent on more than one preceding states or not i.e. dependency on earlier states, among HIV infected patients. Here, we have discritized the CD4 cell count into four states and employed a Markov chain theory.

2. METHODS

Data: The data was extracted from patients enrolled in National AIDS Research Institute -Pune Municipal Corporation -National AIDS Control Organization (NARI- PMC-NACO) ART clinic situated at Model colony clinic, Pune, India.

Inclusion criteria: The patients whose age was more than 15 years and currently receiving ART were included in the analysis. We restricted our analysis to patients who have maintained their follow-up for 5 year of study visits with data on CD4 cell count measurement at every 6 month time interval. With this, we have base- line CD4 cell count and additionally two records of CD4 cell count every year that gives eleven CD4 cell count measurement transitions over five year period with none of the record missing.

Sample size: With the above criteria we had data of 418 patients on eleven CD4 cell count measurements to analyze.

Ethics approval: To use this data, we had obtained the approval from the National AIDS Control Organization and Ethics Committee National AIDS Research Institute, Pune for the retrospective analysis of the data. Patient records/information was anonymized and de-identified prior to analysis.

Classification of states: While modeling the progression of a disease, the first step is to define the disease in terms of different states. Referring to the study by Giuseppe *et al*. (2007), we retained the same classification of states excluding the death state, as we did not have information on death status of the patients under the present research study.

The states defined for our analysis were state 1(S1): CD4 cell count > 500; state 2 (S2): 350 < CD4 cell count 500; state 3 (S3): 200 < CD4 cell count 350 and state 4 (S4): CD4 cell count 200.

Statistical analysis: Maximum Likelihood Estimation (MLE) was used to estimate transition probability matrix and to determine order of Markov chain. Akaike Information Criterion (AIC) and Likelihood Ratio Test (LRT) were used for the best model selection among the set of models.

3. RESULTS

To estimate one step TPM on the basis of the given data, for all the patients' we counted the number of transitions made by the patients in the form of 4 4 transition count matrix. This gives the number of observed transitions in the data, e.g. there were 4180 contributions (observed transitions) to the likelihood function from these 418 patients (Table 1).

Table 1: The One-Step Transition Count Matrix

State	State 1	State 2	State 3	State 4	Total
State 1	1067	251	55	9	1382
State 2	341	441	182	11	975
State 3	13	307	589	106	1115
State 4	17	57	297	337	708
Total	**1538**	**1056**	**1123**	**463**	**N = 4180**

Maximum likelihood estimation: The one step transition probabilities are computed which are given in the following transition probability matrix.

P^ =

State	1	2	3	4
1	0.7721	0.1816	0.0398	0.0065
2	0.3497	0.4523	0.1867	0.0113
3	0.1013	0.2753	0.5283	0.0951
4	0.024	0.0805	0.4195	0.476

In the above matrix all the entries are non-zero hence it confirms that the underline Markov chain is ergodic or irreducible.

The above estimated transition probability matrix reveals that probability of patient staying in same state is higher compared to transitioning to another state

(Staying in State 1 (P_{11}) = 77.21%; State 2 (P_{22}) = 45.23%; State 3 (P_{33}) = 52.83% and in State 4 (P_{44}) = 47.60%). A negligible percent of patients transition from State 1 to State 4 (P_{14}) = (0.65%) and 1.13% transition from State 2 to State 4 (P_{24}) was observed.

We also observed that the process is reversible i.e. patients transitions from deterioration state to improvement state or vice-versa is possible with significant probability. The transition from State 1 to State 2 (P_{12}) is 18.16% and its reverse transition was from State 2 to State 1 (P_{21}) was 34.97%. Similarly, State 4 to State 3 (P_{43}) was 41.95% and from State 3 to State 2 (P_{32}) was 27.53%. Additionally, (P_{22} +P_{21}) was 80.20% and (P_{33} +P_{32} +P_{31}) was 90.43%. This is possible when patients are on regular follow-up and adhered to prescribed medication. These results indicate that a good adherence with ART and follow-ups may lead to stabilization of patient's health, recover their immune status to some extent and can show increment in their survival time.

Testing order of Markov chain models

We had computed the log likelihoods for Markov chains with an order ranging from 0 to 4 using the relative frequencies. A Markov chain with highest likelihood indicates fourth order Markov chain (Table 2). A Markov chain with fourth order depends on current state and the three preceding states. This makes the higher order Markov chain better than the lower order to discover and handle different trends. This is because it has better memory and considers more substance. In practice we need to check whether the higher order Markov chain is significantly better see, Singer (2016). The problem with higher order Markov chains is that the number of parameters to be estimated are very large.

Table 2: Log Likelihoods for Varying Order of Markov Chains

Order	Log Likelihood
0	-6220.35
1	-3984.88
2	-3331.34
3	-2775.70
4	-2197.48

With our data with four states, the fourth order Markov chain requires 768 parameters to be estimated. Considering the sample size of 418 data points the number of parameter estimation seems to be very large. Also, before jumping to the conclusion that underlining process is of order 4, we will run through the model selection criterion.

Statistical model selection using Akaike information criterion (AIC)

Akaike H (1977), derived a criterion which is now called the Akaike Information Criterion (AIC) given by

$$AIC = -2^{*} \ln(L) + 2K$$

Where k is the number of parameters in the statistical model and L is the maximized value of the likelihood function for the estimated model.

For best model we need to consider minimum AIC value and the minimum value is for order 4 (Table 3). The result is consistent with that of likelihood maximization. Encouraged by this fact we had proceeded to apply likelihood ratio test to verify the best fit.

Table 3: Akaike Information Criterion for Varying Order of Markov Chain Models

Order	Number of parameters (k)	AIC
0	3	12446.69
1	12	7993.76
2	48	6758.69
3	192	5935.40
4	768	5930.96

Likelihood ratio test

The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit between two models. A relatively more complex model is compared to a simpler model to see if it fits a particular dataset significantly better i.e. how many times more likely the data are under one model than the other see, Huelsenbeck, Hillis, Nielsen (1996).

The method for using the likelihood ratio for model comparison was adopted using Akaike's Information Criterion (AIC). The method uses the likelihood ratio statistics and modifies it by a penalty term. The model with the lowest value is to be preferred. We now define method (k order, r order, m states) as follows

$$AIC(k) = D - 2^{*}(m_r - m_k)^{*} (m- 1)$$

Given our log likelihood values obtained above we set r=4 (fourth order) and test whether we should probably prefer a lower order model even though the higher order chain results in a lower log likelihood value. AIC computation based on likelihood ratio test statistics are presented in Table 4.

Table 4: AIC Scores Using Likelihood Ratio Test

Order	D	AIC
0	8046	3456
1	3575	-961
2	2268	-2052
3	1156	-2300

We want to prefer a model with lowest AIC value. Looking at the AIC scores we can see that a Markov chain of order 3 should be preferred in comparison to the other models.

4. CONCLUSIONS

This manuscript is an attempt to use higher order Markov chains to model HIV infected patients CD4 cell count who are on ART. If the past and current information of a process are known, the statistical behavior of future evolution of the process is determined by its present state, and therefore, the states of the past and the future are conditionally independent. The higher order Markov chains help to observe the memory properties of CD4 cell count transitions. It is observed that the disease progression is not completely memoryless which establishes the need of frequent follow-ups.

Our results indicate presence of third order Markov chain i.e. it has two preceding CD4 cell count states in memory along with the present state before entering into next improvement state (higher CD4 cell count category) or deterioration state (lower CD4 cell count category).

To conclude, the government ART access program has been successful in maintaining immunological status of patients.

Acknowledgements

The authors acknowledge and thank National AIDS Control Organization, Government of India, supported NARI-PMC ART link center situated at Model colony, Pune. The authors would like to thank the study participants for their contribution.

REFERENCES

Akaike, H. (1977). On entropy miximisation principle. *In: Proc.Symp.on Applications of Statistics ed.* P.R.Krishnaiah. 27-47, Amsterdam, The Netherlands.

Bwayo, J.J., Nagelkerke, N.J., Moses, S., Embree, J., Ngugi, E.N., Mwatha, A., Kimani, J., Anzala, A., Choudhri, S., Achola, J.,O., et al. (1995). Comparison of the declines in CD4+ counts in HIV-1-seropositive female sex workers and women from the general population in Nairobi, Kenya. *Journal of acquired immune deficiency syndromes and human retro- virology*. 10(4):457-461.

Giuseppe, D., D'Amico, G., Girolamo, A., Janssen, J., Iacobelli, S., Tinari, N.and Manca, R. (2007). A stochastic model for the HIV/AIDS dynamic progression. *Mathematical Problems in Engineering*, Volume 2007, Article ID 65636, 14 pages.

Guihenneuc C, Richardson S, Longini IM (2000). Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4+ cell decline. *Biometrics*. 56:733-741.

Hendriks, J.C., Satten, G.A., Longini, I.M., van Druten, H.A., Schellekens, P.T., Coutinho, R.A., van Griensven, G.J. (1996). Use of immunological markers and continuous-time Markov models to estimate progression of HIV infection in homosexual men. *Acquired Immune Deficiency Syndrome*.10(6):649-656.

Huelsenbeck JP, Hillis DM, Nielsen R. (1996). A Likelihood-Ratio Test of Monophyly. *Systematic Biology*. 45 (4): 546. doi:10.1093/sysbio/45.4.546

Kousignian, I., Autran, B., Chouquet, C., Calvez, V., Gomard, E., Kat- lama,C., Rivire, Y., and Costagliola, D. (2003). Markov modeling of changes in HIV-specific cytotoxic T lymphocyte responses with time in un- treated HIV-1 infected patients. *Statistics in Medicine*. 22:1675-1690.

Longini, I.M. Jr., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F., Hethcote, H.W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*. 8(7):831-843.

Rangel-Frausto, M.S., Pittet, D., Hwang, T., Woolson, R.F., and Wen- zel, R.P. (1998). The Dynamics of Disease Progression in Sepsis: Markov Modeling Describing the Natural History and the Likely Impact of Effective Antisepsis Agents. *Clinical Infectious Diseases*. 27:185-190.

Sato, R,C, & Zouain, D.M. (2010). Markov Models in health care. *Einstein.* 8(3 Pt 1): 376-379.

Singer, P. (2016). Order estimation for Markov Chain Models. Accessed on 31 December 2016. http://www.philippsinger.info/?p=44

Sonnenberg, F.A. & Beck, J.R. (1993). Markov models in medical decision making: a practical guide. *Medical Decision Making*.13(4):322-338.

The Return Time Distribution Based Analysis of Dengue Virus

Trupti Vaidya*
Department of Statistics, Savitribai Phule Pune University, Pune, India: tvaidya23@gmail.com

Mohan Kale
Department of Statistics, Savitribai Phule Pune University, Pune, India.: mmkale@unipune.ac.in

Vaishali Waman
Bioinformatics Centre, Savitribai Phule Pune University, Pune, India: vaishali.p.waman@gmail.com

Pandurang Kolekar
Bioinformatics Centre, Savitribai Phule Pune University, Pune, India: pandurang.kolekar@gmail.com

Urmila Kulkarni-Kale
Bioinformatics Centre, Savitribai Phule Pune University, Pune,India: urmila.kulkarni.kale@gmail.com

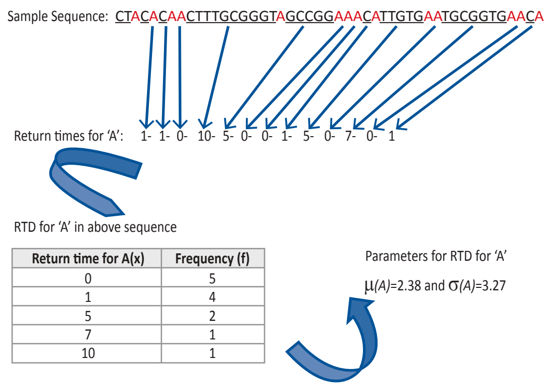
Abstract

When an outbreak of epidemic is observed it is important to track the virus, its family, origin, genotype, and its serotype to counter it. Statistical techniques play a crucial role in genomics or bioinformatics. Return time distribution (RTD) is an important statistical technique which can help to identify the pattern in Deoxyribonucleic Acid (DNA), Ribonucleic acid (RNA), gene, or protein sequences of virus. The dengue virus (DENV) is the cause of dengue fever. It is a single positive-stranded RNA virus of the family *Flaviviridae*. Four serotypes of the virus have been found. Currently, no human vaccine is available. Several vaccines are under development by private and public researchers. One needs to study Dengue virus evolution as well as its specio-temporial variations along with Biodiversity. It is in this context, the current research highlights the use of RTD's of structural and non-structural regions for comparison among the serotypes isolates. It involves comparisons of structural and non-structural regions through distributional properties via modeling of distributions of the RTD's. It is observed that the RTD's of non-structural and structural regions show different probability distribution. The entire analysis is based on identifying stochastic pattern namely probability law for RTD. Hence, collection of RTDs at mono, di, tri and tetra level can be viewed as signature of each sequence. The parameters of the fitted distribution for various serotypes differ across the serotypes, hence, parameter value consideration can provide a reasonable separator. Also, considering the structural and non-structural sequences Markov order was determined and homogeneity test between two Markov chains was performed. The analysis which was carried out showed that structural regions are more likely to evolve faster than non-structural region.

Keywords: Serotype, Stochastic, Structural, Non-structural Proteins.
Journal of Economic Literature (JEL) classification: C63

1. INTRODUCTION

When an outbreak of epidemic is observed it is important to track the virus, its family, origin, genotype, serotype to counter it. This will not only help in deciding line of treatment but also to study behaviour of virus and explore the chances of mutation. Hence it is important to identify the pattern in DNA, RNA, gene or protein sequences of virus. One such technique to identify the pattern is known as return time distribution (RTD) i.e. returns time between nucleotides (measured as number of sites). The figure below shows the computation of return time distribution:



Return time for A(x)	Frequency (f)
0	5
1	4
5	2
7	1
10	1

Figure 1: The Computation of Return Time Distribution (RTD) of 'A' and its Parameters in Sample Sequence at k = 1

Return times are random variables and bring stochastic nature to the sequence. The natural statistical question is what type of distribution is an appropriate model for RTD or which distribution is a good fit for return time distribution. In case of DNA sequence, one can have four (4) RTD's each for nucleotide A, G, C, and T at monolevel. Similarly for k^{th} nucleotide combination there are 4^k return time distributions.

Since biological sequences need not be of same length in the comparative genomic studies, the first step is to make alignment of these sequences. Alignment based methods are time complex, computationally laborious and put extra pressure on resource because of which it may result in high complexity of the algorithm. Therefore, attempts are being made to come up with alignment free analogue of these techniques which consumes less time than its alignment counterpart and provides almost the same and equally accurate results. Hence, a new thought brings in alignment free approach, consuming less time and has minimum level of complexities which is based on internucleotide distance.

Kolekar *et.al.*,(2012) has proposed return time distribution method which is alignment free method and has successfully employed on Mumps, Dengue etc. The analysis focus on fitting distributions to this obtained RTDs.

Dengue virus is a mosquito-borne single positive-stranded RNA virus of the family *Flaviviridae.* Four serotypes of the virus have been found, all of which can cause the full spectrum of disease. Its genome is about 11000 bases that codes for three structural proteins (capsid protein C, membrane protein M, envelope protein E) and seven non-structural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, NS5). Currently, no human vaccine is available for the treatment of disease caused by Dengue virus. Several vaccines are under developmental stage by various private and public organizations/researchers. Developing a vaccine against the disease is challenging. With four different serotypes of the dengue virus that can cause the disease, the vaccine must immunize against all four types to be effective. In order to overcome the lack of human vaccine problem, one needs to study Dengue virus evolution as well as its specio-temporial variations along with Biodiversity (Waman *et.al.,*(2016)). These studies can be done on the basis of complete genome, sub genomic regions and protein sequences. For the present study, it was decided to use structural and nonstructural regions for the comparison among the serotypes isolates etc. These sub genomic regions are chosen because many of the biochemical properties such as immunity etc., are associated with them or decided by the variations in the sub genomic regions.

2. METHODOLOGY

Initially, the method of analysis explored in this research was tried on SH gene sequence of Mumps virus A to M. It was observed that the frequency of occurrences of distributions acts like a good classifier for the genotype classification. This classifier has higher resolution at tri and tetra mers when value k is larger. Encouraged by the results obtained in earlier work it is decided to apply the same technique for an organism/virus which has a sequence much longer than SH gene sequence. Hence, Dengue virus was selected as a test case for the present work. The analysis of RTDs of Dengue virus for non- structural as well as structural regions was planned and carried out. It is known that these two regions shows different variability patterns hence separate analysis of these is required.

The data is obtained from Dengue Virus Variation Resource at National Centre for Biotechnology Information and curated at Bioinformatics Centre, Savitribai Phule Pune University, Pune, India. The data consists of RTDs for structural and non-structural sequences of Dengue viruses. There are 3186 structural sequences and 3186 nonstructural sequences of Dengue viruses used in the present work. The analysis is carried out at monolevel.

The analysis aims to carry out probabilistic modelling via RTDs, to study the probabilistic behaviour of structural and non-structural regions for the comparison among the dengue serotypes, to compare evolution rate by determining order of Markov chain and further testing homogeneity of two Markov chains.

The techniques such as distribution fitting, Markov chain ordering and test of homogeneity of sequences to study the above aspects have been applied.

Table 1 below shows the Return time Distribution of A:

Table 1: Return Time Distribution of A D3 Virus with Accession Number JF937641

Return time (xi)	Frequency(fi)
6	25
11	4
3	82
7	16
9	10
2	99
12	2
15	2
8	6
4	39
1	160
0	220
10	10
13	3
5	40

3. DATA ANALYSIS

3.1. Distribution fitting

If one has some data generated by some random process on a random variable or on random vector and wants to know what type of pattern it reveals, then distribution can be used to explain the variation in the data and the process carried out is fitting of distribution. The steps in distribution fitting can be as follows:

- Use of graphical techniques.
- Choice of the distributions from families of distributions.
- Testing goodness of fit.

The RTD of each of Structural and Non- structural sequences of Dengue viruses is analyzed to recognize the pattern using easy fit software with Kolmogorov Smirnov (K-S) criterion and Anderson Darling (A-D) criterion. The distribution which has rank 1 is selected as the best model for the given data. Structural region shows geometric distribution as model for RTD whereas non-structural has negative binomial distribution (n>1). The table below shows the parameters of negative binomial and geometric distribution fitted to non-structural and structural regions respectively:

Table 2: Parameters of Negative Binomial Distribution

Seq. Accession No.	Type	Parameters	Mean	variance
GU131697	D1	n=3 p=0.14051	18.35	2.57847
FJ478455	D2	n=3 p=0.12531	20.94	2.62407
FJ182011	D3	n=3 p=0.11819	22.38	2.64543
AY618991	D4	n=3 p=0.11463	23.17	3.389941

Table 3: Parameters of Geometric Distribution

Seq. Accession No.	Type	Parameters	Mean	variance
GU131697	D1	p=0.05311	17.83	335.6966
FJ478455	D2	p=0.04768	19.97	418.9
FJ547089	D3	p=0.04397	21.74	494.4912
EU529692	D4	p=0.04607	20.71	449.4487

From the parameters of both the distributions it can be seen that the mean recurrence time of RTDs of D1 is less as compared to D2, D3 and D4. This implies that D1 has less mutation and higher evolutionary rate as against D2, D3 and D4.

3.2. Markov Chain Ordering

Now for this part of analysis instead of RTDs the original sequence of viruses was considered. As all nucleotides are connected with each other it is a finite Markov chain. As at least one state is ergodic this implies that it is ergodic Markov chain.

A procedure for the determination of the order of a Markov chain by Akaike's Information Criterion (AIC) has been developed by Tong(1975). The AIC is defines as

AIC = (-2)log(Maximum likelihood)+2(Number of independent parameters in the model)

This statistic is introduced as a measure of deviation of the fitting model from the true structure. Given several models, the procedure envisages adoption of the model that minimizes the AIC and is called Minimum AIC Estimation.

In this analysis order of Markov chain is determined using DNA Ginie software (developed in-house). The order of Markov chain is 6 for structural regions and 7 for non-structural regions. This implies that in structural regions appearance of nucleotide depends on earlier 6 nucleotides while in non-structural it depends on earlier 7 nucleotides. Thus the probability of occurrence of particular nucleotide at key position is less in structural case and high in non-structural. Less the probability of occurrence the higher the evolutionary rate. Hence, structural regions are more likely to evolve faster than non-structural region. As order remains same for all structural and non-structural regions it was thought to carry out homogeneity testing of two Markov chains.

3.3. Test of homogeneity testing of two Markov chains

To test the null hypothesis that the ordered realisation comes from a Markov chain with a given transition i.e. matrix $P^0=(p_{jk}{}^0)$; suppose that the null hypothesis is

H_0: $P=P^0$

Then, for large N, and for p_{jk}, the statistic is $\sum_{k=1}^{m} \frac{n_j(\hat{p}_{jk}-p_{jk}{}^0)^2}{p_{jk}{}^0}$ j=1,2,..m

is distributed as χ^2 with m-1 d.f (degrees of freedom).

where $\widehat{p_{jk}} = \frac{n_{jk}}{\sum_{k=1}^{m} n_{jk}} = \frac{n_{jk}}{n_{j.}}$

The test of homogeneity showed that no two structural and non-structural sequences are homogeneous.

4. CONCLUSIONS

It is observed that RTD's of non-structural and structural regions show different probability distribution. Structural region shows geometric distribution as model for RTDs whereas non-structural has negative binomial distribution (n>1). As all structural and non-structural regions are modeled by single distribution (geometric or negative binomial), this establishes RTDs limitation to provide separator within serotypes. Parameters ranges of the geometric or negative binomial distribution for various serotypes differs across the serotypes hence parameter value consideration can still provide a reasonable separator. The order of Markov chain is 6 for structural regions and 7 for non-structural regions. This means that structural regions are more likely to evolve faster than non-structural region. This fact confirms known virological observation hence the method is validated. The test of homogeneity was performed on structural regions which showed that no two Markov chains are homogeneous.

The analysis can further be extended to higher order mers and the work is under progress.

REFERENCES

Waman V, Kale M , Kasibhatla S.,Kulkarni U(2016) Population genomics of dengue virus serotype4: Insight into genetic structure and evolution, Arch Virology, 161(8): 2133-48.

Waman V, Kale M , Kasibhatla S.,Kulkarni U, Ramthithkar M(2016),Analysis of genotype diversity and evolution of dengue virus serotype2 using complete genome, Peer J.

Kolekar P, Waman V, Kale M, Kulkarni U(2016) RV-Typer: A web server for genotyping of Rhinovirus using alignment free approach. PLOS-ONE

Kolekar P, Kale M , Kulkarni U, Nilesh Hake (2014) WNV Typer: A server for genotyping of West Nile viruses using an alignment-free Method based on return time distribution. Journal of Virological Methods P.41-55

Kolekar P, Kale M, Kulkarni U (2012) Alignment free Distance Measure Based on Return Time Distribution for sequence Analysis: Applications to clustering, Molecular Phylogeny And Subtyping, Molecular Phylogeny Evolution, Vol 65, p. 510-522

Kolekar P, Kale M, Kulkarni U(2011) Genotyping of Mumps Viruses Based on the SH gene: Development of a server using Alignment-free and alignment based methods. Immunome Research, Vol 7 No.3:4 p.1

The Association of Road Traffic Injuries with Injury Severity Score at Region 11 Southern Province in Thailand

Natthika Sae-Tae*
[1]PhD Candidate, Program in Research Methodology, Prince of Songkla University – nattika99@gmail.com

Sampurna Kakchapati*
Post-doctoral Public Health Scientist, Program in Research Methodology, Prince of Songkla University– kck_sampurna@yahoo.com

Apiradee Lim*
Assistant Professor, Program in Research Methodology, Prince of Songkla University – api_45@hotmail.com

Abstract

The high rates of road traffic injuries and the complexity of management, focusing on the treatment and outcome of injuries, brought about the development of injury scores. The aim of the study was to access factors associated with Injury Severity Score in Nakhon Si Thammarat and Surat Thani Province of Thailand. We conducted a secondary analysis of road traffic injuries in Nakhon Si Thammarat and Surat Thani Province of Thailand from 2008 to 2013. Data were obtained from injury surveillance of Office of Disease Prevention and Control region 11th. Chi-squared test was used to examine the association between risk factors and Injury Severity Score. Logistic regression models were applied to identify the determinants of Injury Severity Score. The severe score due to road traffic injuries was 3.9% (95% CI, 3.7%-4.1%). Multivariate logistic regression revealed that gender, age groups and Injury severity score were significantly associated with road traffic injuries. Injury Severity Score were higher among males (AOR=1.4; 95% CI=1.2-1.5), aged groups above 60 years (AOR=2.6; 95% CI=1.5-4.4), Divers road user (AOR=1.2; 95% CI=1.1-1.3) and safety belts use (AOR=2.3; 95% CI=1.0-6.6). The oldest drivers involved in road traffic injuries have the highest risk for severe injury and fatality. Their efforts are needed teaching knowledge to aging for prevent control road traffic injuries in this region. Thus urgen education on the effect of safety belts is recommended.

Keywords: Road traffic injury, Region 11, Injury Severity Score

1. INTRODUCTION

World Health Organization defines road traffic injury as a fatal or nonfatal injury incurred as a result of a collision on a public road involving at least one moving vehicle and pedestrians. It is a major but neglected public health challenge that requires concerted efforts for effective and sustainable prevention (Tiruneh et al, 2014). Individuals, families and the country's economy have suffered great loss due to traffic injuries. It has been the cause of many untimely deaths, incapacitating injuries and financial loss due to both property damage and a loss in productivity resulting in adverse effects on a countries economy. Thailand, which is a heavily motorized country, has one of the best, wide spread and connected transportation systems in Asia. However the road safety and traffic injuries rates have become a major concern to the country (Bajracharya, 2013)

The high rates of road traffic injuries and the complexity of management, focusing on the treatment and outcome of injuries, brought about the development of injury scores (Kondo *et al*, 2011). Injury scores quantitatively summarize injury severity and have played a major role in the management of road traffic injuries in the developed countries due to an improved triage (Saad *et al*, 2016). An injury severity score is used to summarize a person's injuries using a single number. These methods are important in trauma research and are mostly used to study mortality on a group level. Injury severity scores can be used to compare the effectiveness of trauma care in different regions and time periods. They can be applied to all trauma patients (Tamsen, 2015).

The Nakhon Si Thammarat hospital and Surat Thani hospital are the tertiary hospitals in the Southern Region of Thailand. In this region, limited studies of factors associated with road traffic injuries, in tertiary care hospitals have been documented and a study in needed to conduct a study on road traffic injuries and its determinants in these provinces. With this view and importance, this study assessed severity with road traffic injuries in the region and various factors associated with it. This study will provide baseline data to policy makers and other stakeholders on traffic injuries in tertiary hospital in these provinces. Moreover the study will also recommend various interventions to improved road safety in the country.

2. MATERIALS AND METHODS

A retrospective analysis of road traffic morality in Tertiary Hospital of Nakhon Si Thammarat and Surat Thani of Thailand was conducted. Data of road traffic injuries from 2008 to 2013 were obtained from injury surveillance from Office of Disease Prevention and Control region 11th. The independent variables were age, gender, occupation, road user type, type of vehicle, key risk factors (speed, helmet, alcohol, safely belt and mobile. Injury Severity Score (ISS). Abbreviated Injury Scale categorization and calculation of ISS were done from the anatomic description of the injuries. ISS was classified into non- severe and severe.

Bivariate analyses were conducted to examine the association of demographic, key risk factors and vital signs with road traffic mortality using chi-squared tests. Logistic regression analyses were performed to determine variables associated with road traffic mortality defined by combinations of the determinants, using the additive model:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

In this model, P is the expected probability of road traffic injuries, α is intercept, x_i through x_k are determinants variables, β_i are regression coefficients. Sum contrasts were used to obtain confidence intervals for comparing each proportion with the overall proportion. As it is necessary to construct specific contrasts for logistic regression, this can be accomplished by using weighted sum contrasts rather than treatment contrasts (Tongkumchum and McNeil, 2009) where the first level is left out from the model to be the reference. The advantage of using appropriately weighted sum contrasts is that each proportion can be compared with the overall proportion rather than with a specified reference group. The computed 95% confidence intervals provide a way of classifying the levels of each factor according to whether each corresponding confidence interval exceeds, crosses or is below the overall proportion (Kongchouy and Sampantarak, 2010). The confidence intervals compare percent of injury severity score in each category of a factor with the overall percent. All statistical analysis and graphs were carried out by using R program.

3. RESULTS

About 10,012 injuries occurred from 2008 to 2013, 27.0 percent (2,705) injury severe score due to traffic injuries. Table 1 examines the association between background characteristics and injury severity score of road traffic injuries in Nakhon Si Thammarat and Surat Thani Province. Province, gender, age, road users, vehicle type was significantly associated with road traffic injury severity score. Road traffic level score was severe in Nakhon Si Thamarat (29.0%) than compared to Surat Thani Province (24.9%). Similarly males had severe (29.1%) compare to female (22.0%). Road traffic injury significantly increased with increasing age and severe was found to be higher above 60 years (27.3%) followed by less than 60 years (11%). Driver had more severe than passenger (28.5%). Among the victims who were either driver, pedestrians were suspected to have not used alcohol (27.1%). But the victims who were use medicine (33.3%). Another driver had more not use safety belts and helmet (27.1%) (Table 1).

The median ISS was four among vulnerable road users and one among non-vulnerable road users (p=0.001). Differences in proportion of province, gender, age groups, road user and use of safety belts those between with severe injury. (Table 2). On bivariate analysis, vulnerable road users were at higher risk of sustaining severe injury compared to province (OR1.23, p=0.001). Male has severe score injury compared to gender. (OR 2.9, p=0.01) and at night-time (OR 2.4, p=0.004) were also risk factors for severe injury, while being in a crash while travelling in a van or bus at the time of the crash was protective

(OR 0.5, p=0.02) (Table 3). In multivariate logistical regression, vulnerable road users, road crashes in rainy weather, and night time crashes remained independent risk factors for severe injury (Table 2).

Table 1: Bivariate Analysis of Factors Association Characteristics with Injury Severity Score

Variables	Injury severe score			
	Non-severe (N=7,307)(%)	Severe (N=2,705)(%)	Chisq	P value
Province			21.04	< 0.001
Nakhon Si Thammarat	3,646 (71.0)	1,490 (29.0)		
Surat Thani	3,661 (75.1)	1,215 (24.9)		
Gender			53.52	< 0.001
Male	4,983 (70.9)	2,049 (29.1)		
Female	2,324 (78.0)	656 (22.0)		
Age			18.56	< 0.001
≤60 yrs	130 (89)	16 (11)		
>60 yrs	7177 (72.7)	2,689 (27.3)		
Road Users			31.87	< 0.001
Driver	5,325 (71.5)	2,122 (28.5)		
Passenger	1,982 (77.3)	583 (22.7)		
Use of Alcohol			0.18	0.67
Yes	736 (73.6)	264 (26.4)		
No	6,571 (72.9)	2,441 (27.1)		
Use of Medicine			0.11	0.735
Yes	12 (66.7)	6 (33.3)		
No	7295 (73)	2699 (27)		
Use of safety belts			3.59	0.058
Yes	35 (87.5)	5 (12.5)		
No	7,272 (72.9)	2,700 (27.1)		
Use of Helmet			0.15	0.7
Yes	419 (73.8)	149 (26.2)		
No	6,888 (72.9)	2,556 (27.1)		
Use of Mobile			0.02	0.881
Yes	3 (60.0)	2 (40.0)		
No	7,304 (73.0)	2,703 (27.0)		

Table 2: Multivariate Logistic Regression of Factors Associated with Injury Severity Score

Variables	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
Province				
Suratthani	1			
Nakhorn Sri Thammarat	1.23 (1.13,1.35)	1.22 (1.11,1.33)	< 0.001	< 0.001
Gender				
Female	1			
Male	1.46 (1.32,1.61)	1.38 (1.24,1.53)	< 0.001	< 0.001
Age group				
≤ 60 years	1			
> 60 years	3.04 (1.81,5.13)	2.59 (1.53,4.39)	< 0.001	< 0.001
Road User				
Passenger	1			
Driver	1.35 (1.22,1.5)	1.17 (1.05,1.31)	0.006	0.006
Use of safety belts				
Yes	1			
No	2.6 (1.02,6.64)	2.59 (1.01,6.64)	0.047	0.026

From univariate analysis, province, gender, age, injury severity score for road traffic injury were found to be statistically significant factors (p-value < 0.05). So these determinants were used for multivariate analysis.

4. DISCUSSION

This study shows that when injured, vulnerable road users are at risk of sustaining more severe injury compared to vehicle occupants. Road traffic accidents continue to bear a substantial burden in Thailand. This present study assessed the road traffic injury and factors associated with injury severity score in Nakhon Si Thammarat and Surat Thani Province of Southern Thailand. This showed the level severe score due to road traffic accidents was 27.0 percent of patients in this region of Thailand.

Majority of those injured in the study were males which is consistent with findings from other studies and in other low-income and middle-income countries (Osoro *et al*, 2011). This could possibly be due to the greater exposure to traffic of the males compared to females as drivers or riders and as frequent travelers in motor vehicles for work-related activities. Males had severe score compare to females.

The road crash victims in this study were aged group more than 60 years. Similar age distribution of road crash victims was reported in other studies in Hanrahan at Wisconsin country epidemiologic review. Some promising alternative approaches to age triggered assessments for the identification of unsafe older drivers are being developed. For example, the tiers assess driver's visual, mental, physical abilities, knowledge, and perceptual response time. The impact of driving cessation on the elderly must be taken into consideration. Losing the ability to drive may limit the freedom and independence of seniors. Forced cessation of driving may lead to social isolation and depression in the elderly, which is why physicians, state licensing agencies, and others should work together to help the elderly drive safely for as long as possible (Hanrahan *et al,* 2009).

5. CONCLUSIONS

In this study, injuries sustained by divers were more likely to be severe compared to passenger. These vulnerable road users may help protect them from being hit by automobiles. Severe injuries were also more likely to male and realize to safety. Road traffic injuries are preventable and measures such as use of reflective helmets users and enhancing road visibility with street lights and reflective paint could help reduce the severity of road traffic injuries.

REFERENCES

Hanrahan, R. B., Layde, P. M., Zhu, S., Guse, C. E., & Hargarten, S. W. (2009). The association of driver age with traffic injury severity in Wisconsin. *Traffic injury prevention*.

Kongchouy N, Sampantarak U. (2010) Confidence interval for adjusted proportions using logistic regression. *Modern Applied Science*.

Kondo, Y., Abe, T., Kohshi, K., Tokuda, Y., Cook, E. F., & Kukita, I. (2011). Revised trauma scoring system to predict in-hospital mortality in the emergency department: Glasgow Coma Scale, Age, and Systolic Blood Pressure score. *Critical care*.

Osoro, M. E., Ng, Z., Oundo, J., Omolo, J., & Luman, E. (2011). Factors associated with severity of road traffic injuries, Thika, Kenya. *Pan African medical journal*.

Saad, S., Abo-zied, A. S., Siam, W. A., & El-Ghoul, Y. (2016). Assessment of the use of TRISS scoring system in polytraumatized patients in Suez Canal university hospital, Ismailia, Egypt. *International Surgery Journal*.

Tamsen, F., Logan, F. K., & Thiblin, I. (2015). Homicide Injury Quantification: Correlations and Reliability of Injury Severity Scores Applied to Homicide Victims. *Homicide Studies*.

Tiruneh, B. T., Dachew, B. A., & Bifftu, B. B. (2014). Incidence of road traffic injury and associated factors among patients visiting the emergency department of Tikur Anbessa specialized teaching hospital, Addis Ababa, Ethiopia. *Emergency medicine internationa*.

Tongkumchum P, McNeil D.(2009). Confidence intervals using contrasts for regression model. *Songklanakarin J Sci Technol*.

Yongchaitrakul, T., Juntakarn, C., & Prasartritha,T. (2012). Socioeconomic inequality and road traffic accidents in Thailand: comparing cases treated in government hospitals inside and outside of Bang

This page intentionally left blank

CPS13:
Macroeconomic Statistics (3)

Discrepancies in International Trade Statistics between Trading Partners

Mostafa Mahmoud Abd El-Naby
Central Agency for Public Mobilization and Statistics (CAPMAS) Cairo, Egypt - mos.capmas@hotmail.com

Abstract

The phenomenon of the differences in foreign trade statistics between national sources and trading partners is one of the common global phenomena facing the quality of foreign trade statistics. The reasons for this phenomenon are the different methodologies and mechanisms which used in the collection and processing of international trade data, in addition to some technical problems in the data sources, such as a different degree of inclusiveness in the data. So many countries and organizations are conducting studies to reconcile data between national sources and exports and imports statistics provided by trading partners which called "Mirror Exercises". This paper discusses the Discrepancies in imports and exports data between Egypt and the biggest trade partners China and Italy for 2015 at the total level. Simple mirror exercise and demonstrating some main methodological points (Trade system, Classification, Currency Conversion Factor, Valuation, Partner) in each country was did. Data source was United Nations Commodity Trade Statistics Database (UN comtrade).We found that the differences in the mentioned methodological points had a great effect on making big discrepancies between Egypt and its biggest trading partners China and Italy. The method of descriptive analysis was used.

Keywords: Mirror exercise, Trade system, Valuation.

1. INTRODUCTION

Trade statistics play an important role as information base for enterprises, federations, governmental bodies and other users. Most experienced users seek information from several sources. Accordingly, there is confusion among users, when large discrepancies are displayed in the trade between country A and country B, be it on the aggregated or the detailed level, and in particular if the results are contradictory. There is a need to explain the reasons, and to highlight the most important ones.

The strong belief that one's own figures are always correct has shown its deficiency. Instead the invitation to an open dialogue seems a productive way to accomplish results, i. e. to explain the discrepancies and thereby improving the quality of international trade statistics.

The main purpose of this paper is to present the most important methodological issues and its affection on making discrepancies between Egypt and its most important trading partners China and Italy.

The rest of the paper divided into (4) sections. The first is keywords definitions ,the second is the main Methodological points which can lead us to the main reasons of Discrepancies in imports and exports data between Egypt ,China and Italy for the year 2015, the third is Data comparison (mirror exercise) to determine the percentage of differences between the mirror data of the three countries.

2. DEFINITIONS

- Mirror exercise: It is to make comparisons to the international trade data between two countries or more (at international level) or between two destinations or more from the producers of the international trade data in the same country (at national level).

  Provided that the comparison of both sides at the same level of detail and the same reference period. Preferably be of the same data source on both sides.

  At International level it means "making reconciling data with trading partner" which we make comparison between exports of country A to country B correspond to imports of country B from country A.

From the point of view of producers of trade statistics, i.e. normally the national statistical institutes, this fact provides an excellent possibility for examining the results in the mirror, i.e. to examine how well exports of country A to country B correspond to imports of country B from country A.

- Trade system: There are broadly two approaches, closely linked with customs procedures, used for the measurement of international trade in goods. These are the general trade system and the special trade system.

  The general trade system is the wider concept and under it the recorded aggregates include all goods entering or leaving the economic territory of a country with the exception of simple transit trade. In particular, all goods which are received into customs warehouses are recorded as imports at that stage whether or not they subsequently go into free circulation in the Member State of receipt. Similarly, outgoing goods from customs warehouses are included in the general trade aggregates at the time they leave the Member State.

  The special trade system, on the other hand, is a narrower concept. Goods from a foreign country which are received into customs warehouses are not recorded in the special trade aggregates unless they subsequently go into free circulation in the country of receipt (or are placed under the customs procedures for inward processing or processing under customs control). Similarly, outgoing goods from customs warehouses are not recorded as exports.

- Valuation:

  When you buy or sell goods across national boundaries, you and the other party must have a clear understanding of the terms for moving those goods to their destination. CIF and FOB are commonly used agreement models for international shipping. Each type of agreement specifies which party is responsible for the goods and the point at which responsibility transfers from the seller to the buyer.

  FOB stands for Free On Board. With the FOB type of shipping agreement, the seller or shipper arranges for goods to be moved to a designated point of origin. Normally this is a port because FOB and other INCOTERM contracts are mainly intended for maritime shipping. However, FOB contracts are also used for inland and air shipments. Delivery is accomplished when the seller releases the goods to the buyer. FOB contracts stipulate that this occurs when the goods cross the rail of the ship.

  When a CIF — Cost, Insurance and Freight — shipping agreement is used, the seller has responsibility for the cost of the goods in transit, providing minimum insurance and paying freight charges to move the goods to a destination chosen by the buyer. From the point of delivery at the destination, the buyer assumes responsibility for unloading charges and any further shipping costs to a final destination.

3. METHODOLOGICAL POINTS

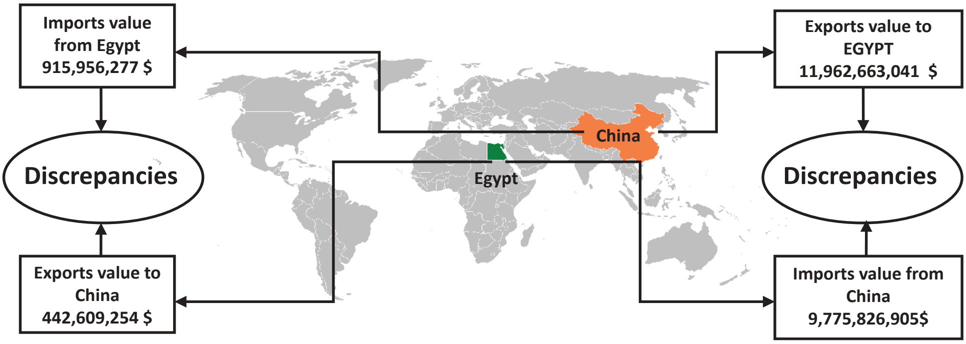
We present the most common items in the methodologies which any difference in each one in country to another can make a big differences in the international trade figures between countries. We got the data of this table from United Nations Commodity Trade Statistics Database (UN comtrade) explanatory and publication notes for the year 2015 of each country as the following:

Table 1. Comparison between Egypt,China and Italy in The Main Methodological Points

<table>
<tr><th>s</th><th colspan="2">Subject</th><th>Egypt</th><th>China</th><th>Italy</th></tr>
<tr><td>1</td><td colspan="2">Trade System</td><td>General</td><td>General</td><td>Special</td></tr>
<tr><td>2</td><td colspan="2">Reported Classification</td><td>HS2012</td><td>HS2012</td><td>HS2012</td></tr>
<tr><td>3</td><td colspan="2">Reported Currency</td><td>USD</td><td>USD</td><td>EUR</td></tr>
<tr><td rowspan="2">4</td><td rowspan="2">Currency Conversion Factor</td><td>Imports</td><td>1.000000</td><td>1.000000</td><td>1.109214</td></tr>
<tr><td>Exports</td><td>1.000000</td><td>1.000000</td><td>1.108475</td></tr>
<tr><td rowspan="2">5</td><td rowspan="2">Valuation</td><td>Imports</td><td>CIF</td><td>CIF</td><td>CIF</td></tr>
<tr><td>Exports</td><td>FOB</td><td>FOB</td><td>FOB</td></tr>
<tr><td rowspan="2">6</td><td rowspan="2">Partner</td><td>Imports</td><td>Consignment</td><td>Origin</td><td>Origin</td></tr>
<tr><td>Exports</td><td>Last Known Destination</td><td>Last Known Destination</td><td>Last Known Destination</td></tr>
<tr><td>7</td><td colspan="2">Publication Notes</td><td>Data for this year is published for the first time; the data was received from/through the Central Agency for Public Mobilization And Statistics on 11 May 2016. (Update date 20-9-2016)</td><td>Data for this year is published for the first time; the data was received from/through the International Trade Center (ITC) on 12 May 2016. (Update date 25-5-2016)</td><td>Data for this year has been published before and we are publishing now newly received updated data; the data was received from/through the Italian National Statistics Institute on 17 November 2016. (Update date 8-12-2016)</td></tr>
</table>

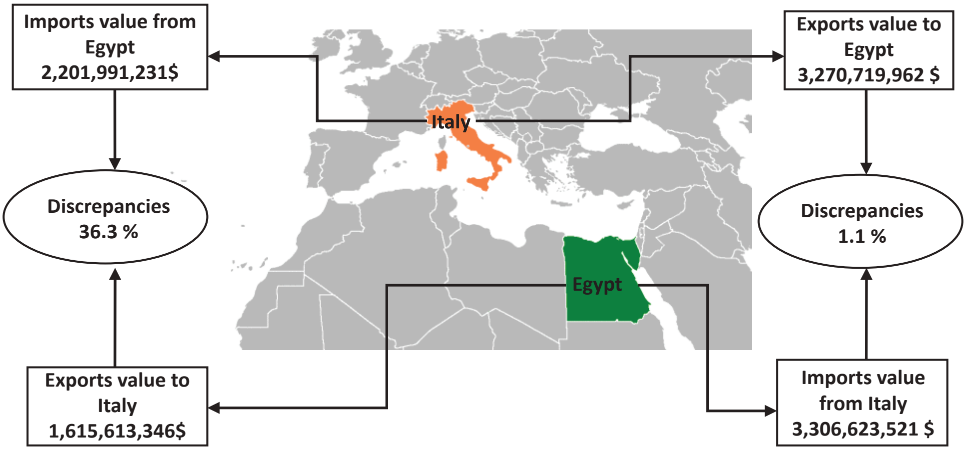
4. DATA COMPARISON (MIRROR EXERCISE)

In this important section we demonstrate the figures of total imports and exports valued in US Dollar ($) in the year 2015 between (Egypt and China),(Egypt and Italy). We calculate percentage of discrepancies to know how the differences in the most common items in the methodologies between (Egypt and China),(Egypt and Italy) from table (1) effected on the mirror discrepancies as shown in the two following forms



Form 1. Mirror Exercise between Egypt and China at the Total Level of Imports and Exports in 2015

We noticed that the differences percentage between Egyptian exports to China and Chinese imports from Egypt is 106.9%. It's very high and the causes need more analysis. There are many causes can lead us to this huge percentage.We think not only the difference between consideration of import country can cause this huge discrepancies percentage.



Form 2. Mirror Exercise between Egypt and Italy at The Total Level of Imports and Exports in 2015

We found that the differences percentage between Egyptian imports from Italy and Italian exports from Egypt is 1.1%. It's an ideal case. We can accept differences percentage till 10% or 15%.

5. CONCLUSION

We conclude that there are mirror discrepancies between Egypt and China due to differences in the methodologies adopted by the two countries especially consideration of Import country. The causes of discrepancies between Egypt and Italy due to differences in reported currency and then currency conversion factor in addition to using different trade system and consideration of import country too. The causes are multiple and would produce differences even if countries were fully to adopt the international recommendations. So countries should do in depth studies for reconciling data (mirror exercise) with trading partners to decrease discrepancies to enhance data quality and dependability.

REFERENCES

Anne B.Dahle, Jens Thomasen and Hans Kristian Ostereng (eds.) 1998, The Mirror Statistics Exercise between the Nordic Countries 1995,.page.5, Statistics Norway, Department for Economic Statistics

Henri Tyrman (2013), Quality report on external trade statistics in Southern & Eastern Mediterranean countries, Adetef – Medstat III , Paris, November 2013, ISBN 978-2-37009-005-8

Nicola Veronese and Henri Tyrman (2009), Methodologies and Working papers, MEDSTAT II: Asymmetry in foreign trade statistics in Mediterranean partner countries" 2009 edition, European Commission, EUROSTAT, Office for Official Publications of the European Communities, Luxembourg, July 2009, ISBN 978-92-79-13750-1

United Nations Commodity Trade Statistics Database (UN comtrade),
http://comtrade.un.org/data/
https://comtrade.un.org/db/mr/daExpNoteDetail.aspx?r=818
https://comtrade.un.org/db/mr/daExpNoteDetail.aspx?r=156
https://comtrade.un.org/db/mr/daExpNoteDetail.aspx?r=381

http://smallbusiness.chron.com/difference-between-cif-fob-39152.html

http://ec.europa.eu/eurostat/statisticsexplained/index.php/Glossary:General_and_special_trade_systems

The Impact of Asean Economic Integration on Indonesia Foreign Direct Investment: A Panel Gravity Model Approach

Imansyah
Badan Pusat Statistik (BPS) Kabupaten Nunukan,
Jl. Ujang Dewa, Nunukan, Indonesia, Telp. +6282137869674, E-mail: imansy@bps.go.id

Abstract

ASEAN leaders agreed to put Foreign Direct Investment (FDI) as a major component in economic development of ASEAN countries and making it as one of the principal objectives of ASEAN in realizing the ASEAN Economic Community (AEC) at the end of 2015. However, the condition of investment climate in Indonesia that continue to get worse will reduce investor's interest for investing in Indonesia. This research is considered important is done to encourage the influx of Foreign Direct Investment (FDI) to Indonesia. It is interesting to gauge empirically, Whether or not ASEAN economic integration has had any impact on Indonesia FDI inflows. In addition, this paper also analyzes the macroeconomics determinants of bilateral FDI flows between Indonesia and home country, both pull factors and inhibiting factors of FDI flows such as GDP per capita, Inflation ratio and real interest rate. The data used in this study were collected from bilateral relation between Indonesia and 21 home countries whose dominant share FDI to Indonesia from 2005 to 2013. Analysis method conducted by augmented gravity model of Indonesia FDI, based on panel data. The gravity model reveals that economic integration of ASEAN has created impact of investment diversion on Indonesia FDI. It shows that there was a competition among ASEAN countries in attracting FDI. Other economic integration such as European Union (EU) and Indonesia's cooperation with Japan in ASEAN-Japan Comprehensive Economic Partnership (AJCEP) have created impact of investment creation on Indonesia FDI. The economic determinants such as GDP of the home country and Indonesia have a positive impact on Indonesia FDI. Distance and home country real interest rate have a negative impact on Indonesia FDI. In order to increase FDI inflow to Indonesia, Indonesia government should strengthen cooperation with investors from European Union (EU) and Japan as well as continue to improve cooperation with other integration cooperation countries and does not just depend on ASEAN countries. Indonesia government also should improve the physical and social infrastructures to drive the productivity and economic efficiency. It will increase the GDP and also attract more investors. Low interest rate policy should be considered.

Keywords: Economic integration, FDI, Gravity model
JEL Classification: C23, E22, F36

1. INTRODUCTION

As a developing country, Indonesia needs a lot of funds to pursue high economic growth. Insufficiency of domestic capital cause Indonesian government create a policy of inviting foreign direct investment. The attraction of FDI is important for Indonesian government to an outward-looking development strategy in contemporary global economy. They bring in new (risk-sharing, non-debt-creating) capital flows, foreign exchange, easy access to foreign markets, and technology transfer (Panayotou, 2000). Furthermore, FDI is expected to be the closing gap between economic development needs with available resources, which eventually became the driving force of economy growth. This opinion is evidenced by the results of studies showing that FDI has a positive effect on economic growth, especially in developing countries (Borensztein et al., 1998).

ASEAN leaders agreed to put Foreign Direct Investment (FDI) as a major component in the economic development of ASEAN countries and making it as one of the principal objectives of ASEAN in realizing the ASEAN Economic Community (AEC) at the end of 2015. Increasing FDI in ASEAN countries is realized through a policy of free capital flows among ASEAN countries. Free capital flow is also useful to obtain a more efficient financial transactions, financing development in a country, facilitating international trade, supporting the development of the financial sector and boosting economic growth (Ministry of Commerce of the Republic of Indonesia, 2015).

Indonesian Government wants to make the moment of ASEAN Economic Community as an opportunity to increase inflow of FDI to Indonesia. However, the condition of investment climate in Indonesia that continue to get worse wiil reduce investor's interest for investing in Indonesia. According to the World Bank doing business levels in 2014, Indonesia was ranked 120 of 189 economies in the world. The bussiness doing ratings of Indonesia lower than other ASEAN countries such as Singapore, Malaysia, Thailand, Brunei Darussalam, Vietnam and Filiphina. Moreover, according to data of Indonesian Investment Coordinating Board (BKPM) 2014, Indonesia's FDI decreased by 0,003% compared to the previous year. Indonesia's FDI decrease from 28,62 billion US$ in 2013 to 28,53 billion US$ in 2014. If the condition persists in the long term, it will affect economic conditions in Indonesia that led to the decline in economic growth in the country.

Therefore this research is considered important is done to encourage the influx of Foreign Direct Investment (FDI) to Indonesia. It is interesting to gauge empirically, Whether or not economic integration agreement approved by Indonesia has had any impact on Indonesia's FDI inflows. In addition, this paper also analyzes the macroeconomics determinants of bilateral FDI flows between Indonesia and home country, both pull factors and inhibiting factors of FDI flows such as GDP per capita, Inflation ratio and real interest rate originating from countries source FDI (home country) and Indonesia againts Indonesia FDI.

2. LITERATURE REVIEWS

2.1. Theoretical Literature

United Nation Conference on Trade and Development (UNCTAD) defines economic integration as a deal was done to facilitate international trade and cross-country movement of production factors. Holzman (1976) defines economic integration was a situation in which the prices of all similar goods and similar factors in two regions were equalised. This made the two regions in essence one region or market. This definition implies that economic integration was the realization of factor price equalization between two region. It implicity assumes that there are no barriers to the movements of goods, service and factors between the two region and that there are institutions that facilitate those movements. Jovanovic (2006) defines the concept of integration is a complex notion which must be defined with care. In General, economic integration can be defined as a process and a means by which a group of countries strives to increase its level of walfare.

Bilateral and regional Free Trade Agreements (FTA) formation attracts long-term, risk-sharing investment flows by creating a more integrated marketplace within which multinational corporations can enjoy a regional division of labor with low transaction costs and exploit economies of scale. A multinational corporation that believes an FTA will lead to greater economic dynamism may be compelled to invest more in one of the members, thus resulting in "investment creation." An FTA may induce more FDI flows into the region by multinationals that are headquartered outside the region. An FTA may also induce intrabloc investment by multinationals with a regional origin.

However, if the multinational decides to invest in the member country not because of a perceived increase in dynamism but because it will now have preferential access to the FTA market, then we have "investment diversion." In other words, although investing in an outsider country might have been more cost effective, the multinational diverts investment to the FTA because of this regional accord. Motivation would be the same as in "tariff hopping" FDI (Plummer, Cheong and Hamanaka, 2010).

2.2. Empirical Literature

Ismail, Smith and Kugler (2009) examined the effects of the economic integration of ASEAN toward FDI using gravity models. They found ASEAN-5 invest in each other less than they invested in new ASEAN members. Results from extra-regional-FDI revealed that European countries increase investment in ASEAN than any other region in the sample. Moreover, further enquiry also found that USA and Japan invested more in ASEAN5 than in new ASEAN member. The market size, income per capita for both source and host countries are positively related with FDI in both panel.

Ridwan (2009) examined the impact of economic integration against ASEAN investment using gravity model. He found that GDP of home country and host country, real interest rate, home country population and Regional Economic Integration (REI) dummy of AFTA and NAFTA had a positive impact on ASEAN FDI. Meanwhile, population and real interest rate of host country, distance, tax and REI of EU, Mercosur and CIND (China and India) had a negative impact on ASEAN FDI.

Plummer and Cheong (2008) using a gravity model is expanded to capture the influence of regional economic integration towards the flow of FDI in the EU, NAFTA, MERCOSUR, and ASEAN. Their search resulted three important conclusions: (1) regional integration has a positive and significant influence towards FDI, which is a combination of the effects of investment diversion and creation (2) investment diversion effects happen in some cases, and thus need to get attention, particularly among developing countries that are not part of regional members with developed countries (3) FDI act as substitutes for trading, although in some cases are complement to trade.

2.3. Hypothesis

Based on the literature review and previous studies, the hypothesis proposed in this study are:

a. GDP per capita both of Indonesia and home country have a positive impact on Indonesia FDI.

b. Indonesia real interest rate, home country real interest rate, inflation ratio (home country/Indonesia) and distance have negative impact on Indonesia FDI.

c. Regional economic integration such as ASEAN, EU and Indonesia's cooperation with Japan, South Korea, China, Australia have a positive impact on Indonesia FDI.

3. METHODOLOGY

3.1. Data

This research uses panel data from bilateral relation between Indonesia and 21 home country whose dominant share FDI to Indonesia from 2005 to 2013. The countries are European Union (Germany, Holland, Italy, Switzerland, France, England, Spain, Belgium), ASEAN (Singapore, Malaysia, Thailand, Philippines), Japan, China, Hongkong, South Korea, Turkey, US, Canada, Seychelle and Australia. The data obtained from the Indonesia Investment Coordinating Board (BKPM), World Bank and International Financial Statistics (IFS). Meanwhile, the distance is calculated based on the distance between the capital of home country and Indonesia obtained from www.infoplease.com.

3.2. Gravity Model

The Newtonian law of gravitation states that two celestial bodies are subjected to a force of attraction that is directly proportional to their mass and indirectly proportional to their distance. In the 1860s, H. Carey was the first to apply the Newton's law of universal gravitation to the study of human behaviour and subsequently the so – called "gravity equation or model" has been widely used in the social science. Tinbergen (1962), use a simple form of gravity model of bilateral trade in analyzing bilateral trade flows.

Linnemann (1966) suggested the model, which describes the flow of goods from one country to another in terms of supply and demand factors (income and population). Anderson (1979) assumed product differentiation and Cobb-Douglas preferences. Bergstrand (1985) concluded that price and exchange rate variation have significant affects on aggregate trade flows. Deardorff (1995) derived a gravity model into FDI for the first time. The basic Model developed by Deardorff is :

$$FDI_{ij} = \frac{(GDP_i x\ GDP_j)}{distance} \tag{1}$$

If the equation is modified in the form the logarithmic, then retrieved empirical equation as follows:

$$\text{Ln } FDI_{ij} = \text{Ln } (GDP_i) + \text{Ln } (GDP_j) - \text{Ln (Distance)} \tag{2}$$

based on the empirical Equation 2, then the model specification equation FDIij are built on this research are:

$$\text{Ln}\ (FDI_{ijt}) = \alpha_0 + \alpha_1 \text{Ln}(GDPP_{it}) + \alpha_2 \text{Ln}(GDPP_{jt}) + \alpha_3 \text{Ln}(\text{D}_{ij}) + \alpha_4 \text{R}_{it} + \alpha_5 \text{R}_{jt} + \alpha_6 \text{Rinf}_{ijt} + \alpha_7 \text{ASEAN} + \alpha_8 \text{EU} + \alpha_9 \text{Japan} + \alpha_{10} \text{China} + \alpha_{11} \text{South Korea} + \alpha_{12} \text{Australia} \quad (3)$$

Where:

FDIijt	:	FDI inflows from home country (i) to Indonesia (j) with respect to year (t),
GDPPit, GDPPjt	:	Gross Domestic Product per capita with respect to year (t)
Dij	:	Distance between capital of home country (i) and Indonesia,
Rit, Rjt	:	Real interest rate with respect to year (t)
Rinfijt	:	Inflation Ratio (Home Country/Indonesia) with respect to year (t)
ASEAN	:	Dummy variable, value 1 while home country (i) is ASEAN country
EU	:	Dummy variable, value 1 while home country (i) is EU Country
Japan	:	Dummy variable, value 1 while home country (i) is Japan
China	:	Dummy variable, value 1 while home country (i) is Cina
South Korea	:	Dummy variable, value 1 while home country (i) is South Korea
Australia	:	Dummy variable, value 1 while home country (i) is Australia.

Estimation of the gravity equation Indonesia FDI using regression data panel can being estimated by common effects model and random effects model. The fixed effects model cannot be used in this study because of the gravitaty equation built in this research contain a time-invariant variables (Baltagi, 2008). The selection of the best model among the common effects and random effects obtained by Breush Pagan (LM Test). Normality assumption testing with the Jarque-Bera test. multicolinierity test assumptions by Variance Inflation Factors (VIF).

4. RESULTS AND DISCUSSION

4.1. Indonesia FDI

In general, Indonesia FDI increased from 1990 to 1995. The decline in non-oil imports during the last quarter of 1995 led to FDI decline in 1996. The Asian economic crisis of 1997 led to the source of investment from Asian countries has decreased. Indonesia FDI realization in 1997 fell by 30 percent compared to the previous year.

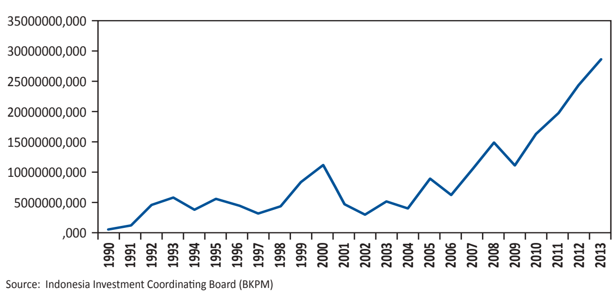


Figure 1. Indonesia FDI Inflows from 1990 to 2013

The implementation of regional autonomy and fiscal decentralization in Indonesia in 2001 led FDI dropped by 58 percent compared to previous year. This occurred due to the implementation of regional autonomy led to increased extortion, lack of transparency and efficiency thus resulting in higher cost of investment in Indonesia. Indonesia FDI decreased in 2004, 2006 and 2008. This is due to rise in oil prices in 2005 and 2008, followed by a rise in the price of fuel by the Indonesian government as well as the global economic crisis of 2008. Oil is an important component for the industry. Rising oil prices be additional costs for manufacturers in the production so as to reduce investment in Indonesia.

The largest source of Indonesia FDI come from Japan with 958 projects and value of 4,71 billion US$ or 16,47 percent of total Indonesia FDI in 2013. Indonesia 's second Largest source of FDI come from Singapore with 1,592 projects and value of 4,67 billion US$ or 16,32 percent of the total Indonesia FDI in 2013. Indonesia 's third Largest source of FDI come from US with 210 projects and value of

Table 1. Top Ten Sources of Indonesia Foreign Direct Investments Inflow

No	Country/Region	Project	FDI Inflows to Indonesia	
			Value (million US$)	Percent of total FDI
(1)	(2)	(3)	(4)	(1)
1	Japan	958	4712,89	16,47
2	Singapore	1592	4670,78	16,32
3	US	210	2435,75	8,51
4	South Korea	807	2205,48	7,71
5	England	231	1075,80	3,76
6	Netherlands	233	927,81	3,24
7	British Virgin Island	307	785,71	2,75
8	Mauritius	55	779,99	2,73
9	Malaysia	574	711,26	2,48
10	Taiwan	158	402,64	1,41

Source: Indonesia Investment Coordinating Board (BKPM)

2,43 billion US$. Basic on Table one can be concluded that the average Japanese investment projects in Indonesia have a greater value than the average of other state projects.

4.2. Model Estimation

Based on the test results of the Breush Pagan (LM Test), we obtained the value of the test statistic is 146,57 which larger than value of χ (0.05; 1) = 3,84, so it can be concluded that random effects approach is better used to describe the influence of the independent variables to the variation value of Indonesia FDI. Jarque-Bera test shows that residual Gaussian. Testing the assumption of non-multicolinearity also fulfilled. The estimation result is displayed as the table below.

Table 2. Gravity Estimation Result for Indonesia Foreign Direct Investment

Dependen variabel: FDI Inflows from home country to Indonesia (Ln FDI_{ijt})				
Independent Variables	Coefficient	Standard Error	t-stat	P-value
(1)	(2)	(3)	(4)	(5)
C	40,584*	7,4493*	5,4479*	0,0000*
GDPP *home country* (Ln $GDPP_{it}$)	0,5270*	0,1217*	4,3302*	0,0000*
GDPP Indonesia (Ln $GDPP_{jt}$)	1,4186*	0,3001*	4,7276*	0,0000*
Distance (Ln D_{ij})	-3,7837*	0,8422*	-4,4923*	0,0000*
Home country real interest rate (Ln R_{it})	-0,0549*	0,0258*	-2,1310*	0,0345*
Indonesia real interest rate (Ln R_{jt})	-0,0318	0,0448	-0,7110	0,4780
Inflation ratio (Ln $Rinf_{ijt}$)	0,4135	0,4473	0,9244	0,3565
ASEAN	-3,2768*	1,5861*	-2,0660*	0,0403*
EU	2,5191*	0,8520*	2,9568*	0,0035*
Japan	3,3897*	1,5003*	2,2593*	0,0251*
China	1,3009	1,5455	0,8417	0,4011
South Korea	2,7496	1,5066	1,8251	0,0697
Australia	0,4367	1,4956	0,2920	0,7706
R-square	0,3414			
F-stat	7,6030			
(P-value)	0,0000			

Notes: * significant at 5 percent

Over-all F-test show that model was significant at five percent level, while R-square is 0,3414 which means that 34,14 percent of Indonesian FDI variation value could be explained by all of independent variable in the model, while 65,86 percent of the rest is explained by other variables outside the model.

4.3. The Impact of Economic Integration

ASEAN economic integration has created impact of investment diversion on Indonesia FDI. ASEAN economic integration has lowered Indonesia FDI of 3,28 percent, cateris paribus. This result shows that there was a competition among ASEAN countries in attracting FDI. This happens due to similarity resources owned by the ASEAN countries, mainly on primary commodities. Moreover, Indonesia doing business level which lower than Singapore, Malaysia, Thailand, Brunei Darussalam, Vietnam and

Philippines made investors from other countries prefer to invest in these countries. Economic integration of Europe has created impact of investment creation on Indonesia FDI. It showed a high attraction of the investors from EU against investment market in Indonesia.

Indonesia's cooperation with Japan in ASEAN-Japan Comprehensive Economic Partnership (AJCEP) has created impact of investment creation on Indonesia FDI. This occurs due to the high attractiveness of Japan investors against investment markets in Indonesia. According to the Japan Bank for International Cooperation (JBIC), there were 30 percent of investors Japan argued that Indonesia's market very attractive for investment purposes and 80 percent of the 500 respondents to Japan investor stated that Indonesia's consumption market will be greater in the future.

Indonesia's cooperation with China in ASEAN-China Free Trade Area (ACFTA), South Korea in ASEAN-Korea Free Trade Area (AKFTA) and Australia in ASEAN-Australia-New Zealand Free Trade Area (AANZFTA) haven't been able to increase the flow of Indonesia FDI. It can be seen from the probability value dummy integration China, South Korea and Australia are greater than 5 percent. This happens because the integration of the ACFTA, AKFTA and AANZFTA began running effective since January 1, 2010 so that the impact of economic integration hadn't given significant impact to Indonesia FDI, considering FDI is an investment that is long term.

4.4. The Impact of Economic Factor

4.4.1. GDP per Capita

Home country GDP per capita has positive and significant impact on Indonesia FDI. Every increase of home country GDP per capita by 1 thousands US$/people then Indonesia FDI will increase by 0,52 percent, when cateris paribus a high level of national income will increases investor revenues, and further high-income investors will increase their ability to invest. This causes an increase in capital outflows from home country to Indonesia so that Indonesia FDI also increased.

Indonesia GDP per capita has positive and significant impact on Indonesia FDI. Every increase of Indonesia GDP per capita by 1 thousands US$/people then Indonesia FDI will increase by 1,42 percent, when cateris paribus. A high level of national income will increases public revenues, and further high-income societies will increase demand for goods and services. Then, corporate profits will be increase so that it will encourage more investment (Sukirno, 2006). The increase in Indonesia GDP per capita will increase the size of the domestic market which giving profit for investors.

4.4.2. Distance

Distance between capital of home country and capital of Indonesia has a negative and significant impact on Indonesia FDI. Every increase of distance between capital home country and capital of Indonesia by 1 KM then Indonesia FDI growth will decrease by 3,78 percent, when cateris paribus. This is due to investments related to bilateral trade the farther distance traveled then the greater the transport costs and time incurred.

4.4.3. Real Interest Rate

Home country real interest rate has negative and significant impact on Indonesia FDI. Every increase of home country real interest rate by 1 percent then Indonesia FDI will decrease by 0,05 percent, when cateris paribus. The real interest rate rises would cause the community prefer saving in banks rather than spend on goods and services. High interest rates will cause purchasing power decreased so that the lower sales of goods and services. Beside that, higher interest rates also illustrate the high risk that must be borne by the investors because the interest rate is the price of Loanable funds to be paid by the investor.

4.4.4. Inflation Ratio

The estimation results showed that at 5 percent significance level, inflation rate home country against Indonesia didn't significant impact on Indonesia FDI. This is due largely home country can suppress inflation below 10 percent. The positive coefficient indicates that when inflation rate ratios increase then the flow of FDI to Indonesia will increase. High inflation will increase the prices of goods, services and production input costs. This condition will show that output price competitiveness of goods and services lower. Inflation causes the purchasing power lower so that decreased trading activity and investor difficult to get a return and profit (Sukirno, 2006).

5. CONCLUSIONS

The gravity estimations reveal that Economic integration of European EU and AJCEP have investment creation on Indonesia FDI, Economic integration of ASEAN has investment diversion on Indonesia FDI. The economic integration of NAFTA, ACFTA and AKFTA haven't been able to increase Indonesia FDI. Other factors such as GDP per capita for both home country and Indonesia have positive impact on Indonesia FDI. The distance between capital of home country and capital of Indonesia and home country real interest rate have a negative impact on Indonesia FDI.

In order to increase FDI inflow to Indonesia, Indonesia government should strengthen cooperation with investors from European Union (EU) and Japan as well as continue to improve cooperation with other integration cooperation countries and does not just depend on ASEAN countries. Indonesia government also should improve the physical and social infrastructures to drive the productivity and economic efficiency. It will increase the GDP and also attract more investors. Low interest rate policy should be considered.

REFERENCES

Anderson, J.E. (1979). A Theoretical Foundation for Gravity Equation. American Economics Review

Baltagi. BH. (2008). Econometrics (4th edition). Jerman (DE): Springer.

Bergstrand, J.H. (1985). The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence. Review of Economics and Statistics, 67: 474-481.

Borensztein E,. Gregorio J,. & Lee J.W (1998). How does foreign direct investment affect economic growth. Journal of International Economics, 45, 157-168

Deardorff, A. (1995). Determinants of Billateral: FDI Does Model gravity in World Economy. The University of Chicago Press, Chicago.

Holzman, F. D (1974). Foreign Trade under Central Planning. Cambridge, MA: Harvard University

Ismail, N. W, Smith, P, & Kugler, M. (2009). The Effect of ASEAN Economic Integration on foreign Direct Investment. Journal of Economic Integration (pp. 385), 407.

Jovanovic, Miroslav N. (2006). The Economics of International Integration (1th ed). Cheltenham, UK and Northampton, MA, USA: Edward Elgar.

Ministry of Commerce of the Republic Indonesia. (2015). Towards an ASEAN economic community 2015 (1th ed), Jakarta.

Panayotou, T. (2000). Globalization and Environment. The Journal Environment and Development 1.Center for International Development 55: Harvard University.

Plummer,M.G., Cheong, David. (2008). Effect of Regional Integration on FDI: an Empirical Approach. Journal of Asian Economics, 19(5): 447-454.

Plummer, M. G. Cheong, David & Hamanaka. (2010). Methodology for effect assessment of free trade agreements (1th ed), Asian Development Bank.

Linneman, H. (1966). An Econometric Study of International Trade Flows. North Holland.

Ridwan. (2009). Dampak Integrasi Ekonomi Terhadap Investasi di Kawasan ASEAN: Analisis Model Gravitasi. Jurrnal Organisasi dan Manajemen, 5(2).

Sukirno, Sadono. (2006). Pengantar Teori Makroekonomi (2th ed.). Jakarta: PT RajaGrafindo Persada.

Tinbergen, J. (1962). Shaping The World Economy: Suggestions for an International Economic Policy. The Twentieth Century Fund. New York.

World Bank. (2014). Doing Business 2014, Understanding Regulation for Small and Medium-Size Enterprises. The World Bank, Washington D.C.

Analysis of the Effects of Foreign Currency Supply and Demand on Rupiah Exchange Rate

Piter Abdullah*
Bank Indonesia, Jakarta, Indonesia – piter@bi.go.id

Bayront Yudit Rumondor
Bank Indonesia, Jakarta, Indonesia – bayront@bi.go.id

Anggita Cinditya Mutiara Kusuma
Bank Indonesia, Jakarta, Indonesia – anggita_cmk@bi.go.id

Rahmat Eldhie Sya'banni
Bank Indonesia, Jakarta, Indonesia – rahmat_es@bi.go.id

Abstract

Indonesia's balance of payments (BOP) is one of the important indicators that is used to analyze the economy performance of a country. BOP data can be used to analyze the economic transactions between residents and nonresidents, also analyzing the foreign currency supply and demand in relation to the movement of rupiah's exchange rate. In addition to maintaining the balance of payments using the accrual basis record, Bank Indonesia also developed a statistical approach to balance of payments supply and demand on the cash basis. Data of BOP on accrual and cash basis has complex and specific characteristics so it needs conscientious in interpreting the data. The flow of foreign currency funds which recorded on a cash basis balance of payments is a potential foreign currency supply and demand and it is expected to affect the movement of rupiah. However, not all potential foreign currency supply and demand will become effective supply and demand, the flow of funds which is converted in domestic foreign exchange market. This research explains the components of cash basis BOP which included as potential foreign currency supply and demand, as well as analyzing the criteria of foreign currency supply and demand that affect rupiah's exchange rate. Moreover, this study also tries to explain the difference between potential foreign currency supply and demand and effective foreign currency supply and demand as an indicator of expectations on the rupiah. The result indicates that the effective foreign currency supply and demand variables significantly affect the value of rupiah, while the potential foreign currency supply and demand of is not proved as significant impact on the rupiah. In addition, the difference between potential and effective foreign currency supply and demand (supply-demand gap) also proved to be significant in affecting the value of rupiah.

Keywords: Foreign currency supply demand; Exchange rate; Rupiah; Balance of payment Indonesia.
JEL Classification: C20, F31

1. INTRODUCTION

In its capacity as a central bank, Bank Indonesia has one single purpose, to achieve and maintain stability of Rupiah. Stability of Rupiah contains two aspects, the stability of the currency's value towards goods and services, and the stability towards the currencies of other countries. The first aspect is reflected in inflation rate, while the second aspect is reflected in the development of the rupiah exchange rate against the currencies of other countries. The development of the rupiah exchange rate against the currencies of other countries is reflected in the movement of the exchange rate of rupiah in financial markets. In a fixed exchange rate system, the local currency is set permanently to some amount of foreign currencies.

While in floating exchange rate system, exchange rates may change at any time, depending on the number of supply and demand of foreign currencies relative to the domestic currency. There are some differences in the perception of the notion of foreign currency supply and demand that affect the value of rupiah. Supply and demand of foreign currency partially reflected on the recorded transactions in Indonesia's BOP. However, the data of BOP announced by the Central Statistics Agency (BPS) does not directly represent the foreign currency supply and demand that actually happens in the domestic foreign exchange market. Bank Indonesia then develop cash basis BOP to explain the flow of foreign currency funds better, but the data still can not described the real foreign exchange supply and demand. This is because not all transactions recorded on BOP and cash basis BOP proceed with the sale or purchase of foreign currency in the market or in other words become effective foreign currency supply and demand.

The high volatility of the rupiah has been allegedly derived from many potential foreign currency supply and demand that have not been traded in the domestic foreign exchange market, or does not become effective foreign exchange supply and demand. Thus, it takes a comprehensive analysis on factors that affect the movement of the rupiah exchange rate derived from foreign currency supply and demand, which are reflected prospectively in the cash basis balance of payments and in the effective domestic foreign exchange market.

Based on the description of the background and the formulation of research questions mentioned above, the purpose of this study are as follows: (i). Describe the criteria of foreign currency supply and demand which affect rupiah in the domestic foreign exchange market (ii). Defining components of BOP and cash basis BOP which included as the potential foreign currency supply and demand in Indonesia. (iii). Analyzing the difference between the effective and potential foreign currency supply and demand (supply-demand gap) in explaining the movements or the expectation of the rupiah. (iv). Identify policy options for the authorities in responding to the current economic conditions related to the exchange rate Rupiah.

2. LITERATURE REVIEW

The foreign exchange market, as the market in general, has two main strengths interacting, demand and supply. As the theory of demand and supply, Nugroho et al. (2014) explains that the exchange rate in a floating exchange rate system will be determined by the interaction between supply and demand. The market is in equilibrium when the interaction between the demand side and the supply side produce an equilibrium price on a certain quantity of demand and supply. In the context of the foreign exchange market, the commodity being traded is foreign currencies and the price is the exchange rate. As in other markets, excess demand for foreign currencies resulting in rising foreign currency prices so the rupiah is depreciated. Conversely, excess supply of foreign currency make the price of foreign currencies decrease so the rupiah is appreciated.

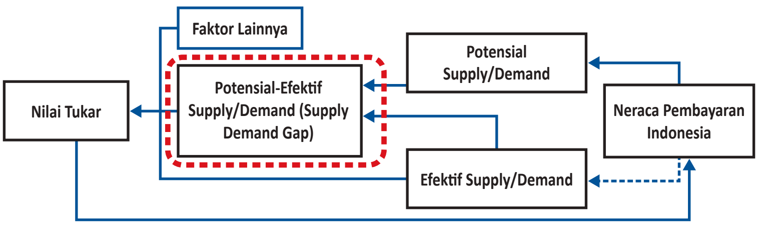
In this research, the demand and supply of foreign currency on the domestic foreign exchange market is an effective supply and demand, meaning supply and demand which has been realized in the form of foreign exchange conversion. As the studies that have been conducted by Sugeng et al. (2010), foreign currency supply or demand is distinguished from the standpoint of bank - as the forex market - based on the flow of foreign currency from foreign currency transactions conducted by the bank.

Sugeng et al. (2010) conducted a research of the interaction between supply and demand of foreign currency and its impact on rupiah. The research concluded that the interaction between demand and supply of foreign currency significantly affect rupiah, and the influence of demand and supply of foreign currency from foreign actor is more dominant than domestic actor. The results also show that the impact of exchange rate movements on the economy occurs only in the short term, the rupiah exchange rate movements significantly affect imports, and depreciation of the Rupiah had a greater impact than appreciation of rupiah.

Study of the rupiah with approach of supply and demand also made by Husman (2005). This study used a model of a composite (hybrid) that combines several approaches determination of the exchange rate which has been frequently used in the literature. The results showed that the variable demand and supply of foreign currency significantly influence the movement of the rupiah. In addition, the estimation results in these studies indicate that the rise in oil prices will lead to depreciation of the currency and that the risk factor is the most influential factor in the movement of rupiah.

3. METHODOLOGY

This study will use two approaches to explain the influence of supply and demand of foreign currency towards rupiah. The first approach is descriptive analysis that will explain the development of the foreign exchange market in Indonesia, describes the data of effective and potential foreign exchange supply and demand, and describes the components of cash basis BOP which classified as a potential foreign currency supply and demand. Then to support the explanation of the concept of effective and potential supply and demand, the second approach in this study used a composite model for the determination of the nominal value of rupiah. The framework of model used can be described as follows:

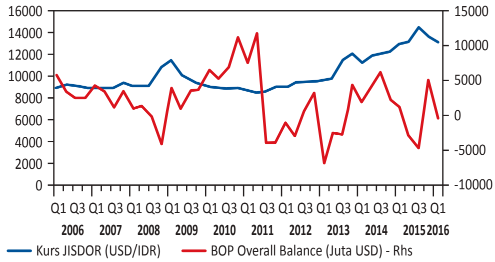


Econometric model used in this study is an Error Correction Model (ECM). This model was chosen to look at the long-term and short-term relationships of some variables to the determining the nominal value of rupiah. Stages in using this method is to first perform cointegration test between independent variables and dependent variable in the long-term equation. This test is used to determine the long-term equilibrium relationship between variables. The next stage is the analysis of short-term relationships between these variables.

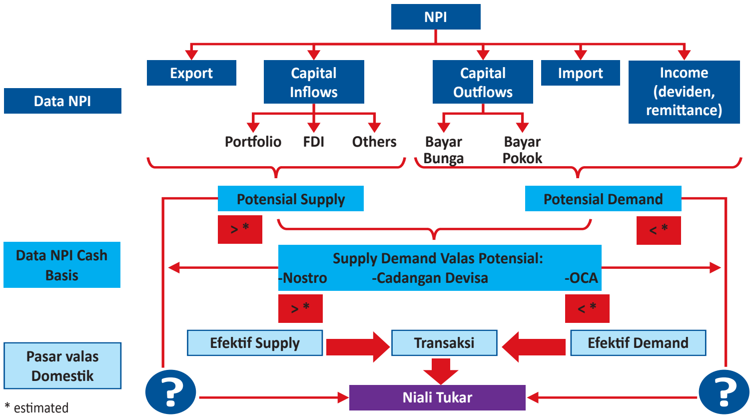
The data used in this research is secondary data related to supply and demand of foreign exchange in Indonesia include both potential and effective, as well as factors that affect the value of rupiah in the form of internal and external factors. Observations were made in the period June 2012 to May 2016.

4. RESULTS AND ANALYSIS

In theory, the flow of foreign currency supply and demand is recorded through economic transactions contained in Indonesia's balance of payments (BOP). As a result, the movement of rupiah exchange rate should be determined by the balance of payments. In the following figure, it appears that surplus balance of payments would normally followed by an increase in foreign currency supply in the domestic foreign exchange market so it will stimulate rupiah strengthened against US Dollar. On the other hand, deficit balance of payments is normally followed by increasing demand of foreign currency in domestic foreign exchange market so that rupiah weakened against US Dollar. The figure also shows that when the balance of payments is in surplus, the exchange rate of USD / IDR can be weakened. However, the data shows that in some periods the movement of rupiah is not in the same direction with the balance of payments.

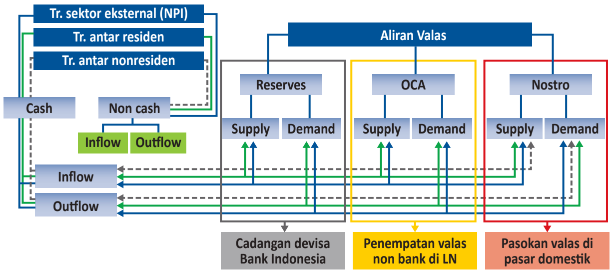


As the scheme demonstrated in following figure, it can be seen that the foreign currency supply and demand is recorded in the balance of payments. Foreign currency supply in BOP is derived from export transactions and capital inflows (portfolio investment, foreign direct investment and other investment). While foreign currency demand comes from import transactions, capital outflow (payment of debt principal and interest), as well as primary income (dividends and remittance).



However, not all supply and demand in BOP will be supply and demand of foreign currency has the potential to be transacted in the domestic foreign exchange market. Potential foreign currency supply and demand is recorded in cash basis BOP which is developed by the Department of Statistics - Bank Indonesia. This cash basis BOP record foreign currency movements on transactions between resident, resident with non-resident and between non-resident which are recorded in foreign currency of the nostro, OCA and reserves. Then, the transactions that took place in the domestic foreign exchange market is included as effective foreign currency supply and demand.

Potential currency supply and demand obtained from cash basis BOP represents the flow of foreign currency (supply and demand) between resident (R-R), resident-non-resident (R-NR) and between non-resident (NR-NR) recorded through nostro, OCA (Overseas Current account) and transfer of reserves (Cadev). The data used for nostro is taken from cash basis BOP (SKLLDI Bank), OCA from SKLLDI LBB, while reserves is taken from reserves table in BOP. Supply and demand of foreign currency (especially R-NR) was also compared with foreign exchange transactions in the BOP. Meanwhile, the supply and demand of foreign currency between non-resident is conducted through nostro with the source of data used is the data of nostro LLD Bank with adjustment. Transactions which are recorded on cash basis BOP is illustrated by following figure.



ECM models used in the composite model in the estimation of long-term exchange rate is:

$$s_t = a_0 + a_1 (P_t - P_t^*) + a_2 (i_t - i_t^*) - a_3 oil_t + a_4 tot_t + u_t \quad (6)$$

Short-term equation used in the composite model:

$$\Delta s_t = \beta_0 + \beta_1 \Delta(i - i^*)_t + \beta_2 \Delta tot_t + \beta_3 \Delta cds_t - \beta_4 \Delta poil_t - \beta_5 u_{t-1} + \beta_6 \Delta(pot - ef\,fa)_{t-1} - \beta_7 sdlna_{t-3} - \beta_8 sddna_t + \beta_9 pota_{t-2} - \beta_{10} dum2015 + e_t \quad (7)$$

The results of the long-term equation estimation show that rupiah exchange rate significantly influenced by fundamental factors such as price differential, interest rate differential, price of oil, and terms of trade. This is consistent with previous research, Husman (2005) that includes price differential, terms of trade and oil prices in the long-term equation. Meanwhile, in research on BEER, Economic Research Group

I (2010) analyzed fundamental economic factors that influence the real value of rupiah as mentioned above plus the ratio of non-tradable to tradable price as indicators of productivity and risk factors.

From the ECM estimation, the Error Correction coefficient is negative and significant indicating that the running ECM model is quite valid. Coefficient of error correction is -0.174 which shows the speed of adjustment toward long-term equilibrium is 17.4% per month.

In previous research, Husman (2005) restrict the price differential variable equal to one which in other words, the assumption of purchasing power parity (PPP) is fulfilled in the long term. Dornbusch (1976) states that the concept of PPP is one of the important pillar in the monetary model to explain the behavior of exchange rate. As for the inclusion of the PPP in the component factors that affecting long-term equilibrium is in line with research conducted by Kim (1990) and Baillie and Selover (1987). One example that is quite popular and often uses the basic concept of PPP is the preparation of the Big Mac Index. However, in reality the PPP condition is difficult to achieve due to several issues, such as transaction costs (Keynes: 1923). Meanwhile Taylor and Taylor (2004) mention several factors that lead to difficulty to achieve PPP conditions conditions such as tarif cost, transport costs and quotas. In line with expectations and consistent with research Kilian and Taylor (2003) and Coakley et al. (2004), the empirical results show a positive relationship between price differential and rupiah. Each 1% increase in difference between the domestic and foreign price index is predicted to weaken exchange rate of rupiah by 0.08% (ceteris paribus).

Estimates of interest rate differential coefficient shows that in both short and long term have a positive relationship with nominal value of rupiah. 1% increase in interest rate differentials between domestic and overseas are expected to cause rupiah depreciated by 3.40% in the long term (ceteris paribus). Meanwhile, in the short-term equation, each 1% change in the interest rate differential is predicted to cause a change in rupiah by 3.04% (ceteris paribus). This is consistent with Meese and Rogoff (1988) which states that a country's currency will depreciate with the widening interest rate differential between the interest rates of domestic and overseas. The wider interest rate differential between one country to another, the greater the expected depreciation of the country compared to other countries.

Oil price variable is included in the model to see specific effects of oil prices in the determination of exchange rate of rupiah. The result indicates that oil price is significant and negatively affect rupiah in both short and long term. In the long run every 1% increase in oil prices is estimated to cause an appreciation of Rupiah by 0.15% (ceteris paribus). While in the short term, every 1% change in the price of oil is estimated to lead to the appreciation of rupiah by 0.06% (ceteris paribus). The negative relationship between exchange rates and oil prices from these empirical results is consistent with expectations, especially if a country is an oil exporting countries. For the case of Indonesia, although this time as the oil importing countries, but Indonesia has also become exporters of other commodities such as coal and palm oil that has a high correlation with oil prices.

The result indicates the coefficient of variable tot is positive and significant in both short and long term. In the short term, every 1% increase in tot expected to cause the rate of change of exchange rate depreciated by 0.32% (ceteris paribus). This indicates that the substitution effect has greater influence than the income effect. The increase in export prices has greater influence in changing consumer behavior towards export goods from other countries compared to the effect of rising prices on the exporter earnings that should be able to add the supply of foreign currency in the domestic foreign exchange market and pushing rupiah to appreciate. This is consistent with Sahminan (2005) that states the substitution effect is greater than the income effect.

Based on estimation results, rupiah exchange rate is negatively and significantly affected by risk variable. Indicators of risk is represented by CDS premiums Indonesia which gives a negative coefficient, meaning that the higher the risk in Indonesia (rising CDS premiums) will make rupiah depreciated. This is consistent with Sahminan (2007) and Frankel (2007) who found that the risk of default of a country has a significant influence in determining the exchange rate.

Output estimates shows that every 1% increase in Indonesia CDS premium, is estimated to cause the rate of change of rupiah depreciated by 0.04% (ceteris paribus). As in previous research, the risk variable is one of the important variable in determining an exchange rate.

Margin of potential and effective foreign currency supply and demand variable is positively and significantly affect nominal value of rupiah. The estimation results show that every 1% increase in the difference between potential and effective foreign currency supply and demand of the previous period is expected to make the rate of change of rupiah depreciated by 0.01% (ceteris paribus). This is consistent with the expectations that the greater the difference between potential and effective foreign currency supply and demand, the greater the expected depreciation of rupiah.

Variable of foreign currency supply and demand by foreign actors are negatively and significantly affect nominal value of rupiah. The estimation results show that every 1% increase in the difference between foreign currency supply and demand with the previous three periods is expected to cause the rate of change of rupiah exchange rate appreciated by 0.02% (ceteris paribus). The greater the sales of foreign currency by foreign market participants in the domestic foreign exchange market, is expected to further push the appreciation of rupiah. In addition, the action of foreign actors tend to be followed by domestic players because they are more accustomed to do thorough analysis by using various methods and relevant information before do the transaction, so the decisions taken can be trusted. In general, effective foreign currency supply and demand affects the market or the level of rupiah. The result indicates that the variable of foreign currency supply and demand by domestic actors does not significantly affect rupiah. This is consistent with the opinion of Sugeng et al. (2010) that domestic actors tend to follow the action of foreign actors.

In addition, the estimation results indicate the variable of potential foreign currency supply and demand do not significantly affect rupiah. This can be explained by the fact that not all foreign currency in the domestic foreign exchange market will be converted by the owners and only affect the supply of foreign currency. As to see the behavior of all economic actors in Indonesia, in this research the potential foreign currency supply and demand use the data from nostro accounts, OCA accounts, and reserves accounts.

Furthermore, there is a potential development of further research on potential foreign currency supply and demand if you want to focus on research to certain accounts or a combination of existing accounts. In the short-term equation, we also used a dummy variable to capture the effect of the high volatility of rupiah at the end of 2015 along with the increasing speculation on the increase of the fed funds rate for the first time.

5. CONCLUSIONS

A number of conclusions can be drawn from the research as follows:

1. Effective foreign currency supply and demand is significantly influential in determining the nominal value of rupiah consistent with the price formation mechanism in the law of supply and demand, the price in the equilibrium is the result of supply and demand that has been transacted.
2. Meanwhile, the potential foreign currency supply and demand does not significantly affect rupiah. It can be caused by the nature of the foreign currency supply and demand which only influence liquidity / currency supply in the domestic foreign exchange market. Not all foreign currencies in the foreign exchange market will be converted by the owner. The supply of foreign currency will effectively influence rupiah when it has been converted. Moreover, potential foreign currency supply and demand variables and foreign currency supply and demand from domestic actors do not significantly affect rupiah.
3. Difference in potential and effective foreign currency supply and demand is significantly and positively affect the nominal exchange rate of rupiah. Difference in potential and effective foreign currency supply and demand can be used as an alternative to get the expectations of market participants towards rupiah exchange rate. The higher the expected depreciation of rupiah, the greater the difference in potential and effective foreign currency supply and demand.
4. In the long term, the variables that significantly affect rupiah are price differential, interest rate differential, oil prices and terms of trade. As in the short term, the variables that significantly affect rupiah are interest rate differential, terms of trade, risks, oil prices, the difference in effective and potential foreign currency supply and demand and also foreign currency demand and supply by foreign actors.

REFERENCES

Baillie, R. T., & Selover, D. (1987). Cointegration and models of exchange rate determination. *International Journal of Forecasting*, 3, 43-51

Budiman, Aida S., Hendarsah, N., Nugroho, M. N., dan Silviani, E. 2004. *Kajian terhadap Struktur Mikro Pasar Valas Antar Bank Domestik*. Buletin Ekonomi Moneter dan Perbankan, Juni 2004.

Coakley, Jerry., Fuerters, Ana M., Wood, Andrew. (2004). A New interpretation of the exchange rate– yield differential nexus. *International Journal of Finance and Economics*, 9, 201-218.

Corden, W. Max. 2002. *Too Sensational on the Choice of Exchange Rate Regime*. MIT Press, 2002.

Dornbusch, R. (1976). Expectations and exchange rate dynamics. *Journal of Political Economy*, 84, 1161-1176.

Frankel, Jeffrey. (2007). On the rand: Determinants of the south african exchange rate. *NBER Working Paper* #13050

Goeltom, M. dan Zoelverdi, D. 1998. *Manajemen Nilai Tukar di Indonesia dan Permasalahannya*. Buletin Ekonomi Moneter dan Perbankan, September 1998.

Husman, Jardine A. 2005. *Estimasi Nilai Tukar Rupiah Paska Krisis: Pendekatan Model Komposit*. Working Paper Bank Indonesia WP/07/2005.

International Monetary Fund. 2013. *Old Risks, New Challenges*. Global Financial Stability Report, April 2013.

Kandil, Magda. 2009. *Exchange Rate Fluctuations and the Balance of Payments: Channels of Interaction in Developing and Developed Countries*. Journal of Economic Integration 24(1), March 2009, 151-174.

Kelompok Riset Ekonomi I. (2010). Re-estimasi Model Nilai Tukar Behavioral Equilibrium Exchange Rate (BEER). *Catatan Riset-DKM*.

Keynes, John Maynard. (1923). A Tract on Monetary Reform. London: Macmillan. pp. 89-90, 91-92.

Kilian and Taylor. (2003). Why is it so difficult to beat the random walk forecast of exchange rates). *Journal of International Economics*, 60, 85-107.

Kim, Y. (1990). Purchasing power parity in the long run. *Journal of Money, Credit and Banking*, 22, 491-503.

Meese, R., and Rogoff, K. (1988). Was it real? The exchange rate-interest differential relation over the modern floating rate period. *Journal of Finance*, 43, 933-948.

Nugroho, M. Noor, Ibrahim, Winarno, T., Permata, M. I. 2014. *Dampak Pembalikan Modal dan Threshold Defisit Neraca Berjalan terhadap Nilai Tukar Rupiah*. Buletin Ekonomi Moneter dan Perbankan, Januari 2014.

Sahminan. (2005). Estimating equilibrium real exchange rates of the Rupiah. *Working Paper Bank Indonesia*, WP/08/2005.

Sahminan. 2007. Determinants of the Rupiah exchange rate. *Working Paper Bank Indonesia*, WP/07/2007.

Sercu, Piet dan Uppal, Raman. 2003. *Exchange Rate Volatility and International Trade: A General-Equilibrium Analysis*. European Economic Review 47 (2003) 429-441.

Simorangkir, Iskandar dan Suseno. 2004. *Sistem dan Kebijakan Nilai Tukar*. Pusat Pendidikan dan Studi Kebanksentralan (PPSK), Seri Kebanksentralan, Mei 2004.

Stockman, Alan C. 1980. *A Theory of Exchange Rate Determination*. The Journal of Political Economy, Vol. 88, No. 4 (Aug., 1980), pp. 673-698.

Sugeng., Nugrohro, M. Noor., Ibrahim., Yanfitri. 2010. *Pengaruh Dinamika Penawaran dan Permintaan Valas terhadap Nilai Tukar Rupiah dan Kinerja Perekonomian Indonesia*. Buletin Ekonomi Moneter dan Perbankan, Januari 2010.

Taylor, Alan M., & Taylor, Mark P. (2004). The purchasing power parity debate. *The Journal of Economic Perspectives*, 18, 135-158.

CPS14:
Macrofinancial Statistics (2)

Robust Single-Index Model with Adjusted Beta: A Case-Study in Indonesia Stock Exchange

Ezra Putranda Setiawan
Universitas Gadjah Mada, Yogyakarta, Indonesia – e2r4.ps@gmail.com

Dedi Rosadi *
Universitas Gadjah Mada, Yogyakarta, Indonesia – dedirosadi@gadjahmada.edu

Abstract

Single-index model is a simple and well-known technique to determine which stocks will be included in a portfolio and their proportion or weight. The important parameter used in this model is the 'beta' of each asset's return. The true 'beta' is not known, yet it must be estimated from historical data by some statistical method. In developing country, the financial market transaction sometimes considered to be non-synchronous and thin. Therefore, the estimated 'beta' must be adjusted in order to diminish the effect of these conditions. On the other hand, robust estimator is a kind of statistics estimator that possesses robustness properties, i.e. insensitivity to small deviations from the assumption about the probability distribution of the data. Previous study showed that the robust estimator that applied to the mean-variance model yield better-performed portfolio. In this paper, we investigate some well-known robust estimator, to estimate the adjusted beta. The robust-estimated adjusted beta then used to build the portfolio with Single-index model. Evaluation of these portfolios performance also carried out and compared each other.

Keywords: Portfolio optimization; Robust regression; Non-synchronous trading.

JEL classification: C13, C51, G11.

1. INTRODUCTION

Buying stocks in the financial market is a common way for both peoples and companies in order to get some benefit from their money. However, the investor must be aware to the large risk of financial loss when the stock prices declines. Therefore, making portfolio that consists of some stocks is a widely-used method to minimize the investment risk.

Markowitz's (1952) introduce a mathematical formulation to solve the problem of portfolio asset's selection. In his method, both the mean and variance of portfolio's return were used as the estimate of expected return and risk, respectively. Here, the portfolio's variance is calculated based on the variance of each assets and covariance between each pairs of them. When the number of assets is increased, the number of parameters that should be estimated also increased rapidly.

One of some approaches that proposed to simplify the mean-variance model has introduced by Sharpe (1963), that now well-known as single-index model. This model based on assumption that the dynamic movement of stock returns is correlated to the dynamic movement of market return. Sharpe introduced the 'beta' of an assets, measure the expected change in the asset's return given the change of the market return. By calculating the beta of each asset, the investor could determine which assets should be taken in his portfolio and how big should the investment weights. As explained by Elton *et al.* (2014), the beta could be calculated based on historical data by simple linear regression techniques, with some elective adjustment proposed by Blume (1975), Levy (1971), and Vasicek (1973). For a developing money market with non-synchronous trading, other adjustment method were introduced by Scholes and Williams (1977), Dimson (1979), and Fowler and Rorke (1983). All method aforementioned has been explained in Hartono (2009).

On the other hand, the least-square approach that used to estimate the asset's beta by means of linear regression techniques, could produce bad estimate when the data contains at least one outliers or unusual observations. A technique known as robust statistics could relieve this problem, by producing estimator that are insensitive to the presence of outliers. As noted by Maronna *et al.* (2006), robust regression could fit the data although they contains one or more unusual observations on the response or predictor variable. DeMiguel and Nogales (2009) showed that robust estimator could yield more stable portfolio than the traditional minimum variance portfolio.

In this paper, we examine the performance of single-index model with adjusted beta that estimated by robust statistical method. A brief introduction about single-index model and the beta are presented in the section 2. Section 3 will discuss mainly about robust estimator, with emphasis on estimator(s) that can be used in the single-index model. Empirical study with real data from Indonesia Stock Exchange (IDX) will deeply explained in the section 4. The last section is devoted for some conclusions.

2. SINGLE-INDEX MODEL

Following notations by Elton *et al.* (2014), based on the Single-Index model, the return on a stock i could be written as

$$R_i = a_i + \beta_i R_m \tag{1}$$

where

a_i is a random variable represent the component of security return that unexplained by market price movement

R_m is a random variable used to model the market rate of return

β_i is a constant that measures the expected change in R_i given a change in R_m

Moreover, the random variable a_i in the above model could be breaked into two components, namely expected value of a_i and random element of a_i. These element usually denoted by α_i and e_i, respectively, so the above formula (1) now appear as

$$R_i = \alpha_i + \beta_i R_m + e_i \tag{2}$$

Two important assumptions here is that e_i uncorrelated with R_m, and that $E(e_i) = 0$. The risk of an individual security could be represented as

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2 \tag{3}$$

where σ_{ei}^2 will converge to zero as the portfolio gets larger. It means that σ_{ei}^2 is a diversifiable risk, and the other terms represent the nondiversifiable risk.

2.1. Building a Portfolio based on Single-Index Model

When making a portfolio, a rational investor will prefer an asset with larger return than lower ones, especially that could give larger return than the riskless asset. From the equation (3) above, larger beta means larger non-diversifiable risk, so the investor will choose an asset with small beta. Therefore, in the Single Index Model, the available assets should be ranked by *excess return to beta*, i.e

$$ERB_i = \frac{\overline{R}_i - R_f}{\beta_i} \tag{4}$$

The ordered list of stock from the largest ERB_i represent the desirability of any stock's inclusion in a portfolio. So, the next step is determining the *cutoff* point C^* so that the assets with ERB_i larger than C^* will be included in the portfolio, and the others will be rejected. However, the cutoff point C^* depends on the number of stock included in the portfolio. Therefore, we must calculate C_i for each stocks, when the assets i assumed to be included in the optimal portfolio. Elton *et al.* (2004) gives the formula for computing C_i as follows:

$$C_i = \frac{\sigma_m^2 \sum_{j=1}^{i} \frac{\left(\overline{R}_j - R_f\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \sum_{j=1}^{i} \left(\frac{\beta_j^2}{\sigma_{ej}^2}\right)} \beta_j^2 \tag{5}$$

These C_i calculated here are candidates for cutoff point C^*. The cutoff point C^* determined equal to C_i when all assets that used to compute it have $ERB > C_i$ and the others have $ERB < C_i$. Last, we could determine the weight of each assets in the optimal portfolio as

$$X_i = \frac{Z_i}{\sum_{ERB_j > C^*} Z_j} \tag{6}$$

with

$$Z_i = \frac{\beta_i}{\sigma_{ei}^2}\left(ERB_i - C^*\right) \tag{7}$$

2.2. Estimating and Adjusting Beta

As mentioned earlier, the beta for an assets could be estimated from historical return of the assets and the market by simple linear regression. If someone has the return data from period $t = 1, 2, \ldots, T$, we get the Ordinary Least Squares (OLS) estimates of beta, namely

$$\beta_i = \frac{\sigma_{im}}{\sigma_m^2} = \frac{\sum_{t=1}^{T}\left[\left(R_{it} - \bar{R}_i\right)\left(R_{mt} - \bar{R}_m\right)\right]}{\sum_{t=1}^{T}\left(R_{mt} - \bar{R}_m\right)^2} \tag{8}$$

Blume (1975) show that the historical beta of an assets should be adjusted in order to preserve the beta property that moving closer to 1. He suggests to estimate the linear regression coefficient from a pair of beta of same stocks that observed in two subsequent period. Another approach by Vasicek (1973) was computing the weighted average of beta, where the weight is calculated based on the standard error of estimating beta. Klemkosky and Martin (1975) in Elton *et al.* (2014) show that both the Blume and Vasicek's approach yield better estimate of beta than historical beta.

On the other side, some research suggest that stock market in developing country are *thin market,* since there are non-synchronous trading (Hartono, 2009). In the thin market, the historical beta values are biased, therefore they must be adjusted by other method than Vasicek and Blume approach above. Some available method here are Scholes and William's (1975), Dimson (1979), and Fowler and Rorke (1983). Research by Ariff and Johnson (1990) and Hartono and Surianto (1999) show that these adjustment work well in Singapore and Indonesia stock market, respectively (Hartono, 2009). On the other side, Bartholdy and Riding (1994) show that no improvement achieved by both Dimson and Scholes and William's beta adjustment in New Zealand stock market. Until now, it is not known which of these adjustment methods is better than the other.

3. ROBUST REGRESSION ESTIMATOR

Many research has shown that ordinary least squares (OLS) estimate of beta as shown in equation (8) above is seriously affected by just an outlier or extreme data point. In addition, a leverage point also could alter the estimated regression coefficient, yielding a regression equation that not fit to the bulk of sample data. Robust regression consist of several estimation technique that will give better estimates of regression coefficients, i.e. that follow the majority data and not affected by the presence of outlier(s) and/or leverage points. Here, we discuss some robust regression techniques. Details of theory could be found in textbook on robust statistics, for example Maronna *et al.* (2006).

Tukey M-estimate. A M-estimate of simple linear regression could be seen as a generalization of OLS estimator. With this estimator, the beta of portfolio is a solution of

$$\min \sum_{t=1}^{T} \rho\left(\frac{R_{it} - \left(\alpha_i + \beta_i R_{mt}\right)}{s_i}\right) \tag{9}$$

where s is a scale estimator and ρ is a real-valued function that symmetric and strictly increasing in positive real number domain. A well-known ρ function that could be used here is the Tukey's biweight function, i.e.

$$\rho(u)=\begin{cases}\frac{k}{6}\left(1-\left[1-\frac{u^2}{k^2}\right]^3\right) & |u|\le k\\ \frac{k^2}{6} & |u|>k\end{cases} \tag{10}$$

Computation of the estimator value could be done by iteratively weighted least square (IWLS) procedure, as described in Maronna *et al.* (2006).

Least Trimmed Square. While M-estimates get their robustness by giving different weight for each residual, least trimmed square only use several residual to estimate the coefficient. The beta estimator by this method is calculated to minimize the sum of square of half smallest residual i.e.

$$\min\sum_{t=1}^{m}E_{(t)}^2 \tag{11}$$

where $m=\lfloor 0.5n\rfloor+1$ and $E_{(t)}$ represent the ordered residual $E_{it}=R_{it}-(\alpha_i+\beta_i R_{mt})$. In other word, the least trimmed square ignore the extreme observation with large residual. Comparing to the M-estimator for regression, least trimmed square regression has higher breakdown point.

MM-estimate. MM-estimator is an improvement of high-breakdown estimator to get another estimator with high efficiency. The MM-estimator for linear regression computed as follows.

1. Using a high-breakdown regression method (i.e. least trimmed square), estimate the regression coefficient and obtain the residual.
2. Calculate a M-estimates of scale (with breakdown equal to 0.5) with function ρ_0 based on the residual obtained in the first step above, namely s_n.
3. The MM-estimate of linear regression beta is calculated such that

$$\min\sum_{t=1}^{m}\rho_1\left(\frac{R_{it}-(\alpha_i+\beta_i R_{mt})}{s_n}\right) \tag{12}$$

Here, the ρ_1 must be symmetric, continouous differentiable, and $\rho_1(u)\le\rho_0(u)$.

Computation of ordinary least squares regression and the three robust regression method above could be done by software **R** using the command **lm** (library base), **rlm** (library MASS), **ltsreg** (library robustbase), and **lmrob** (library robustbase), respectively.

4. EMPIRICAL STUDY

In this empirical study, we choose 20 stocks in Indonesia Stock Exchange. All of these assets are belong to LQ45 index, means that they have high liquidity and/or large market capitalization. For each stocks, we compute the *daily return* for 120 market days. Full result of Jarque-Bera normality *goodness of fit* test, annual return, and beta estimate for each stock are provided in Table 1 below.

Table 1. Jarque-Bera (JB) Test, Annual Return, Beta Estimate and The Standard Error (s.e) for 20 Stock

Stock Name	JB Test*	A year Return	OLS estimate		Tukey - M estimate		LTS estimate		MM estimate	
			beta	s.e	beta	s.e	beta	s.e	beta	s.e
ASRI	250.9	0.2086	1.0391	0.02415	1.0067	0.02415	0.8757	0.02419	1.0067	0.02415
ADHI	46.8	-0.1309	0.7101	0.01818	0.8068	0.01820	0.6824	0.01819	0.8071	0.01820
BBNI	8.30	0.0846	1.0385	0.01476	1.0241	0.01476	0.9493	0.01478	1.0247	0.01476
BBRI	19.2	0.0766	0.9332	0.01374	0.9098	0.01374	0.7595	0.01382	0.91	0.01374
BBCA	329.9	0.1659	0.5201	0.00837	0.4111	0.00842	0.3156	0.00855	0.4198	0.00842
ICBP	22.12	0.2238	1.0998	0.01445	1.1522	0.01445	0.8603	0.01459	1.1605	0.01446
INDF	25.32	0.1858	1.2229	0.01495	1.1775	0.01495	0.8624	0.01526	1.1742	0.01496
INTP	33.82	-0.1371	1.1547	0.01795	1.0738	0.01797	0.8177	0.01818	1.0853	0.01796

Table 1. Jarque-Bera (JB) Test, Annual Return, Beta Estimate and The Standard Error (s.e) for 20 Stock

Stock Name	JB Test*	A year Return	OLS estimate		Tukey - M estimate		LTS estimate		MM estimate	
			beta	s.e	beta	s.e	beta	s.e	beta	s.e
KLBF	88.45	0.1678	0.8807	0.01592	0.7922	0.01594	0.7118	0.01598	0.8064	0.01593
LPPF	10.86	0.0205	0.8515	0.02053	0.7834	0.02054	0.5517	0.02069	0.7912	0.02054
MNCN	54.50	-0.0854	0.6892	0.02425	0.9219	0.02434	0.8413	0.02429	0.922	0.02434
PGAS	155.6	0.0988	1.0879	0.02718	0.9412	0.02721	0.8742	0.02724	0.9456	0.02721
PTBA	34.86	0.4318	1.262	0.02796	1.2586	0.02796	1.1671	0.02797	1.2614	0.02796
PTPP	34.50	0.0977	0.9339	0.01957	0.9176	0.01957	0.6978	0.01967	0.9174	0.01957
PWON	2.86	0.3388	1.2379	0.02344	1.2631	0.02344	1.1917	0.02345	1.2633	0.02344
SMGR	36.90	-0.0074	1.2849	0.01862	1.2079	0.01864	1.0566	0.01873	1.2149	0.01863
SMRA	3.93	0.105	1.1483	0.02219	1.2548	0.02222	1.0072	0.02223	1.2609	0.02222
TLKM	6.71	0.258	1.2642	0.01416	1.2974	0.01416	0.7662	0.01478	1.2969	0.01416
UNTR	9.49	0.1473	0.9931	0.02333	1.0084	0.02333	0.9958	0.02333	1.0083	0.02333
WIKA	247.6	0.0703	1.2109	0.02200	0.9435	0.02212	0.842	0.02222	0.9633	0.02210

*In confidence level 95 %, JB statistics > 5.99 show that the data doesn't follow normal distribution.

Besides the historical beta, we also calculate the adjusted beta to diminish the bias, i.e. by Blume's approach and Vasicek's approach (Elton *et al.,* 2014). We also calculate the beta to diminish the effect of non-synchronous trading in thin market, i.e. by (1) Scholes and William's method, (2) Dimson method, and (3) Fowler and Rorke method. The major improvement in this research is that we use four robust regression mentioned above to calculate the beta with each adjustment procedure. As members of LQ-45 index, we assume that these assets have high liquidity and traded actively in the stock market, so we only use the one period adjustment. Table 2 below summarized the result.

Table 2. Aritmatic Mean of Estimated Beta for 20 Stocks, Calculated by Different Regression Estimator and Adjustment

Regression Method	Calculation Methods					
	Historical	Blume	Vasicek	Scholes and Williams	Dimson	Fowler and Rorke
OLS	1.0281	1.0973	1.1075	1.0974	1.0769	1.0823
Tukey -M	1.0076	1.0591	1.0756	1.0392	1.0405	1.0441
LTS	0.8413	0.8212	0.8223	0.7901	1.0336	1.0367
MM	1.0120	1.0610	1.0766	1.0507	1.0544	1.0599

From the table 2 above, we know that each regression method yield different beta value: the least trimmed square (LTS) yield lowest historical beta compared to the OLS, LTS, and MM estimator. These result are similar when the adjustment are made. However, both Dimson and Fowler and Rorke's method seems reduce the difference between each beta from different estimator.

Next, we examine the weight and performance of portfolio that calculated by Single Index Model based on different beta estimator and adjustment. Here, we assume that no short selling are made, and the risk-free rate is equal to 6% per year.

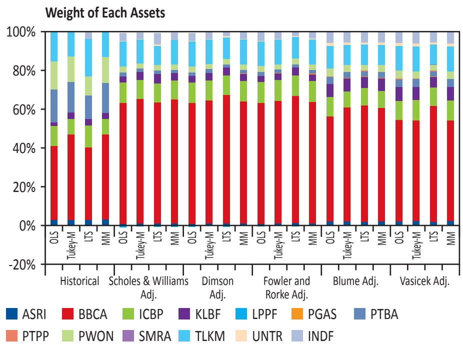


Figure 1. Portfolio Resulted by Single Index Model, Computed by Different Beta Estimator and Beta Adjustment

The number of assets in each portfolio resulted in figure 1 above are distributed from 7 assets only (in 3 portfolios) until 12 assets (2 portfolios). As represent on the figure, most of portfolio consist of some dominant asset, i.e. BBCA and below them there are ICBP and TLKM here, which are available at all portfolio cases. Although, we formulate the portfolio without short selling, actually, in some portfolio (except the Historical, Blume adjusted, and Vasicek adjusted) there is a stock that get negative weight, since they had negative beta (negative correlation with the movement of market return) that implied to negative excess-return to beta (ERB). We still include this stock in the portfolio.

Table 3. Total Return, Risk, and Sharpe Ratio (S.R.) Obtained in a Month for All Portfolios Calculated with Different Beta Estimator and Beta Adjustment

Historical Data Only

Estimator	Return	Risk	S.R.
OLS	0.03798	0.01376	2.4820
Tukey-M	0.03708	0.01326	2.5076
LTS	0.03662	0.01305	2.5124
MM	0.03729	0.01330	2.5155

Blume Adjusted

Estimator	Return	Risk	S.R.
OLS	0.03455	0.01193	2.5735
Tukey-M	0.03520	0.01191	2.6331
LTS	0.03552	0.01187	2.6703
MM	0.03528	0.01192	2.6391

Vasicek Adjusted

Estimator	Return	Risk	S.R.
OLS	0.03439	0.01194	2.5597
Tukey-M	0.03385	0.01192	2.5187
LTS	0.03530	0.01186	2.6517
MM	0.03403	0.01192	2.5324

Scholes and William's adjusted (1 period)

Estimator	Return	Risk	S.R.
OLS	0.03877	0.01186	2.9461
Tukey-M	0.03745	0.01175	2.8608
LTS	0.03822	0.01195	2.8772
MM	0.03769	0.01179	2.8725

Dimson adjusted (1 period)

Estimator	Return	Risk	S.R.
OLS	0.03870	0.01187	2.9372
Tukey-M	0.03714	0.01176	2.8317
LTS	0.03641	0.01173	2.7780
MM	0.03720	0.01179	2.8280

Fowler and Rorke adjusted (1 period)

Estimator	Return	Risk	S.R.
OLS	0.03851	0.01187	2.9204
Tukey-M	0.03682	0.01177	2.8025
LTS	0.03608	0.01171	2.7537
MM	0.03702	0.01179	2.8141

When these betas were used to build a long-only portfolio by Single Index Model, table 3 show that the adjustment method could generate better portfolios, with larger Sharpe Ratio than the historical method only. In the historical data and beta-corrected by Blume's technique, portfolio with robust beta estimator are perform better than the classical, but the result is very different when we use Scholes and William's adjustment, Dimson adjustment, and Fowler and Rorke adjustment. With these preceding adjustment, the portfolio resulted from robust beta estimators work poorer than the classical one. It means that the adjustment for thin market is more important than the usage of robust estimator to determine the asset's beta in Indonesia Stock Exchange.

5. CONCLUSIONS

A case study about the usage of robust regression estimator to calculate the beta portfolio has been done by us. It appear that these robust estimator do not give large improvement on the performance of portfolio that calculated by Single Index Model. Instead, the adjustment method is more important to be considered when estimating asset's beta in Indonesia Stock Exchange. Another research could be done to grasp the effect of robust regression estimator in multi-index or multifactor model and prove the relations with different data sets from different stock market.

REFERENCES

Blume, M. 1975. Betas and Their Regressions Tendencies. *The Journal of Finance.* 10. 785-795.

Bartholdy, J. and Riding, A. 1994. Thin trading and the estimation of betas: The efficacy of alternative techniques. *Journal of Financial Research*. 17(2). 241-254.

DeMiguel, V. and Nogales, F.J. 2009. Portfolio Selection with Robust Estimation. *Operational Research*. 53(3). 560-577.

Dimson, E. 1979. Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics.* 7. 197-226.

Elton, E.J., Gruber, M.J., Brown, S.J., and Goetzmann, W.N. 2014. *Modern Portfolio Theory and Investment Analysis.* Ninth edition. John Wiley and Sons, New Jersey.

Fowler, D.J. and Rorke, C.H. 1983. The risk measurement when shares are subjected to infrequent trading. *Journal of Financial Economics.* 12. 279-289.

Hartono, J. 2009. *Teori Portofolio dan Analisis Investasi.* Edisi keenam. BPFE UGM, Yogyakarta.

Kaszuba, B. 2013. Empirical comparison of robust portfolios' investment effects. *The Review of Finance and Banking.* 85(1). 047-061.

Markowitz, H.M. 1952. Portfolio selection. *The Journal of Finance.* 7. 77-91.

Maronna, R., Martin, D., Yohai, V. 2006. *Robust Statistics Theory and Methods.* John Wiley and Sons, New Jersey.

Scholes, M. and Williams, J. 1977. Estimating betas from non-synchronous trading. *Journal of Financial Economics.* 5. 309-327.

Sharpe, W. 1963. A simplified model for portfolio analysis. *Management Science.* 9. 277-393.

Vasicek, O. 1973. A note of using cross-sectional information in Bayesian estimation of security betas. *Journal of Finance.* 8. 1233-1239.

Household Investment Forecast in The Stock Exchange of Thailand

Gorn Tepvorachai
The Bank of Thailand, Bangkok, Thailand – gornt@bot.or.th

Somsajee Siksamat*
The Bank of Thailand, Bangkok, Thailand – somsachp@bot.or.th

Yuwawan Rattakul Boonyaleephan
The Bank of Thailand, Bangkok, Thailand – yuwawanr@bot.or.th

Abstract

The investment outlook of households is an important factor for domestic investment, leading to the expansion of domestic economy and possibly resulting in the improvement of household well-being. Therefore, the investment perspective of households can be used as an indicator of consumer confidence in the domestic economic outlook. This study creates a measure that reflects the household investment view through the actual investment in the Stock Exchange of Thailand (SET), forecasted one month ahead. The one month ahead security-by-security net household transaction is forecasted based on the Stock Exchange of Thailand's internal security-by-security factors: stock price, stock trading volume, outstanding value by investors' portfolio, price change effect by investors' portfolio, and unit change effect (net transaction) by investors' portfolio. The data used in this study is from secondary data source, collected monthly from January 2009 to July 2016 total of 91 months from the SET and the authors' calculation. The net household transaction forecast is modelled by two comparison techniques: moving multiple regression (MMR) and moving artificial neural network (MANN). This study finds that MANN forecast produces lesser mean absolute error (MAE), root mean square error (RMSE), and mean absolute percent error (MAPE) statistics than the MMR, but higher directional accuracy (DA test) statistic than the MMR. Both MANN and MMR techniques produces significantly different forecasts with Diebold Mariano (MA test) statistic at 95% confidence level. This finding confirms the MANN forecast is more efficient and more accurate than the MMR forecast. The MANN forecast suggests the household investment is positively affected by stock price and price change effect within household's portfolio. On the other hand, the MANN forecast suggests the household investment is adversely affected by the outstanding portfolio value of the non-financial investors, financial investors, and public investors.

Keywords: Investment outlook; Unit change; Security-by-security; Moving artificial neural network.
JEL Classification: C53; E17; G17

1. INTRODUCTION

The perspective of the household sector investor can be used as an indicator of consumer confidence in the economic outlook of the country, encouraging investment, promoting economic growth, and improving the well-being of households in the country.

This study proposes an indicator that reflects the view of household sector investor, forecasted one month ahead, which is a measure of household investor perspective. The measure is calculated using a model based on the data collected from the Stock Exchange of Thailand (SET) and the calculation from the Bank of Thailand (BOT). This indicator is linked to stock price, stock trading volume, outstanding value by investors' portfolio, price change effect by investors' portfolio, and unit change effect (net transaction) by investors' portfolio.

The objectives of this study include:

1. Construct an indicator that reflects the perspective of household sector investor through the net transaction of household investor forecasted one month ahead.
2. Analyze factors that affect the net transaction of household investor forecasted one month ahead.

The hypothesis of this study include:

1. Net transaction of household investor forecasted one month ahead can be effectively predicted by the following factors: security-by-security (stock) price, security-by-security (stock) trading volume, security-by-security outstanding value by investors' portfolio, security-by-security price change effect by investors' portfolio, and security-by-security unit change effect (net transaction) by investors' portfolio.
2. Net transaction of household investor forecasted one month ahead changes in the same direction with stock trading volume, security-by-security unit change effect (net transaction) by investors' portfolio, and price change effect by investors' portfolio.
3. Net transaction of household investor forecasted one month ahead changes in the opposite direction to stock price and outstanding value by investors' portfolio.

The boundaries of this study include:

1. Use monthly information at the end of each month from September 2013 to July 2016 from the Stock Exchange of Thailand (SET), the Bank of Thailand (BOT) and authors' calculation including security-by-security (stock) price, security-by-security (stock) trading volume, security-by-security outstanding value by investors' portfolio, security-by-security price change effect by investors' portfolio, and security-by-security unit change effect (net transaction) by investors' portfolio.
2. Compare the prediction results using moving multiple regression (MMR) and moving artificial neural network (MANN).

The benefits of this study is that analysts can effectively identify factors that affect the investment outlook of household sector investor through the net transaction of the household investor forecasted one month ahead.

2. LITERATURE REVIEW

The economic factor forecasting is based on the relationship between quantitative factors by using a function to convert the causal factors to the factor of interest. Some forecasting methods use explicit functions to indicate the relationship between these factors. (Atikankul, 2012) compares forecasting methods on the Stock Exchange of Thailand (SET) index to find a suitable forecasting model and studies the relationship between the SET index and market indices in other countries in order to find correlations with other indices using four forecasting methods, including simple exponential smoothing, Holt's method, Box-Jenkins' method, and linear regression. Another research by (Khanthavit, 2009) studies the prediction on the size and direction of changes in the securities price in a volatile market, using Bayesian Averaging of Classical Estimates (BACE) as the stock price predictor. This study finds that BACE generates profits to investors at a higher level.

While other SET index forecasting techniques use black-box-liked functions which do not explicitly indicate the relationship between these factors such as artificial neural network (ANN). (Kaewmart & Chancharat, 2013) (Pantuwon, 2014)

(Pantuwon, 2014) studies a mixed model that incorporates an ARIMA (Box-Jenkins method) model and an ANN model, which combines the capabilities of a single model. The mixed model reduces the limitations of a single model, improves the efficiency in describing relationships among data, and improves the forecasting accuracy on the Stock Exchange of Thailand index. This study finds that the mixed model (AR-NN) has the ability to forecast out-of-sample SET50 index with higher forecasting accuracy than single model (ARIMA). (Kaewmart & Chancharat, 2013) studies an ANN model to predict the stock market index of nine countries. The study finds that the ANN model yields an efficient and high accuracy prediction.

Therefore, in this study, the authors choose to use artificial neural network for our prediction method, since it can potentially provide higher forecasting accuracy than linear prediction method (Pantuwon, 2014).

3. METHODOLOGY AND DATA

In this study "household investment forecast in the Stock Exchange of Thailand," the authors conduct the study following the guideline below.

1. Using bottom-up approach, projecting security-by-security household investment and collecting the projected investments into the market-wide household investment (net transaction of household portfolio).
2. Using quantitative approach, analyzing only quantitative data, including security-by-security (stock) price, security-by-security (stock) trading volume, security-by-security outstanding value by investors' portfolio, security-by-security price change effect by investors' portfolio, and security-by-security unit change effect (net transaction) by investors' portfolio.
3. Using technical approach, utilizing quantitative factors above to describe the movement of the net transaction of household investor forecasted one month ahead.

This study uses secondary data from the Stock Exchange of Thailand, the Bank of Thailand, and author's calculations. These quantitative data are collected at the end of every month between July 2013 and July 2016, total of 37 months. The data collection includes the following items:

1. Security-by-security price from the Stock Exchange of Thailand (stock price; IDX)
2. Security-by-security trading volume (stock turnover) from the Stock Exchange of Thailand (turnover volume; VLM)
3. Security-by-security outstanding value by investors' portfolio[1] from the Bank of Thailand (value; VL)
4. Security-by-security unit change effect (net transaction) by investors' portfolio from the authors' calculation (unit change; UC)
5. Security-by-security price change effect by investors' portfolio from the authors' calculation (price change; PC)

This study screens securities on the Stock Exchange of Thailand to a total of 293 securities in order to reduce the propensity of securities that are rarely traded or not traded at all with the following guideline:

1. Common stocks (excludes preferred stocks, warrants, derivative warrants, ETFs, unit trusts, transferable subscription right, property funds, REITs)
2. The Stock Exchange of Thailand (excludes Market of Alternative Investment)
3. Security is not under rehabilitation.
4. Security has the minimum turnover of 1,000 baht per month (calculated from the accumulative probability of beta distribution, alpha=2, beta=5 at 99% confidence level).
5. Security must meet all of the above requirements in every month during this study period.

This study creates a modeled indicator that reflects the perspective of household sector investor through the net transaction of household investor forecasted one month ahead. The model utilizes the above information from the current month back to 2 previous months, a total of 3 months. The authors believe that the Stock Exchange of Thailand is continuously changing and news affecting the market in short term not exceeding 3 months. The authors define forecasting model for the net transaction of household investor forecasted one month ahead as shown in Equation 1.

$$UC_{t+1}^{HH} = f\left(IDX_{t,t-1,t-2}, VLM_{t,t-1,t-2}, VL_{t,t-1,t-2}, UC_{t,t-1,t-2}, PC_{t,t-1,t-2}\right) \qquad \text{Equation 1}$$

The data or variables used in this study are as follows:

UC_{t+1}^{HH} is the dependent variable measuring the net transaction of household investor forecasted one month ahead (at the end of month $t + 1$)

f(·) is the forecasting model predicting the net transaction of household investor forecasted one month ahead. This study will compare predictions of two models: moving multiple regression (MMR) and moving artificial neural network (MANN).

1 Investors are classified into 5 groups: general government (GG), financial corporation (FC), non-financial corporation (NFC), household (HH), and the rest of the world (ROW) by the System of National Accounts 2008.

The data or independent variables used in this study are as follows:

$IDX_{t,t-1,t-2}$ is the security-by-security price at the end of months t, t - 1 and t - 2 measuring the price of a stock. When the stock price rises, the minimum transaction size for buying or selling a stock for the household investor increases. Household investor needs more capital to buy the same stock. As a result, the net transaction of household investor reduced. Thus, the stock price is correlated in the opposite direction with the net transaction of household investor.

$VLM_{t,t-1,t-2}$ is the security-by-security trading volume at the end of months t, t - 1 and t - 2 measuring the turnover of all purchases made in one month for one stock. When the turnover of a stock increases, household investor tends to increase its interest in the stock and thinks the probability that the price of the stock will rise. As a result, the net transaction of household investor increases. Thus, the stock turnover is correlated in the same direction as the net transaction of household investor.

$VL_{t,t-1,t-2}$ is the security-by-security outstanding value by investors' portfolio at the end of months t, t - 1 and t - 2 which consists of 5 investor groups: general governments (GG; $VL^{GG}_{t,t-1,t-2}$), financial corporations (FC; $VL^{FC}_{t,t-1,t-2}$), non-financial corporations (NFC; $VL^{NFC}_{t,t-1,t-2}$), households (HH; $VL^{HH}_{t,t-1,t-2}$), and the rest of the world (ROW; $VL^{ROW}_{t,t-1,t-2}$) by the System of National Accounts 2008. It measures the outstanding value of a stock in an investors' portfolio. When the stock accumulation in the household portfolio increases, the household investor tends to invest less in that stock and considers diversifying the risk by investing in another stock instead. As a result, the net transaction on that stock of household investor reduced. Thus, the stock outstanding is correlated in the opposite direction with the net transaction of household investor.

$UC_{t,t-1,t-2}$ is the security-by-security unit change effect (net transaction) by investors' portfolio at the end of months t - 1 and t - 2, measuring the net transaction of all buying transactions minus selling transactions made in one month for one stock in the investors' portfolio. When the net transaction of a stock in the household's portfolio increases, household investor tends to increase interest in that stock and thinks the probability that the price of the stock will rise. As a result, the net transaction of household investor increases. Thus, the net transaction of a stock on an investors' portfolio is correlated in the same direction as the net transaction of household investor.

$PC_{t,t-1,t-2}$ is the security-by-security price change effect by investors' portfolio at the end of months t - 1 and t - 2, measuring the changes in the value of a stock from price movement that occurs in one month in an investors' portfolio. When the value of a stock in household's portfolio increases from price movement only, household investor tends to increase interest in that stock and thinks the probability that the price of the stock will rise. As a result, the net transaction of household investor increases. Thus, the price change of a stock on an investors' portfolio is correlated in the same direction as the net transaction of household investor.

This study analyzes quantitative information, which tests the relationships between the factors that affect net transaction of household investor forecasted one month ahead, from January 2014 to July 2016. Additionally, this research studies the magnitudes in the relationships by computing the relation coefficient for each independent factor. Then, the authors will summarize and describe the relationships between the independent factors and the net transaction of household investor forecasted one month ahead.

This study evaluates the efficacy of the two forecasting models: moving multiple regression and moving artificial neural network. The authors compare the predictions of the models with actual net transaction of household portfolio in three ways:

1. Model accuracy in forecasting the level of net transaction is measured by the following statistics. The model with smaller statistic has higher level forecasting accuracy.
   a. Mean absolute error (MAE)
   b. Root mean square error (RMSE)
   c. Mean absolute percent error (MAPE)
2. Model accuracy in forecasting the direction of net transaction (directional accuracy; DA test) is measured by comparing the rate of forecasts showing a plus and a minus. The model with greater statistic has higher directional forecasting accuracy.

3. Comparing predictive accuracy (Diebold Mariano; DM test) is measured by the errors of predictions from models with the actual net transaction of household investors and computed the predictive accuracy with a loss function.

4. FORECASTING RESULTS

The preliminary statistical analysis of the factors related to net transaction of household investor for all the stocks from April 2009 to June 2016 is analyzed. The preliminary statistical analysis shows that security-by-security price (*IDX*) has the mean of 24.96 THB per share and the median of 5.80 THB per share, which is vastly different from the mean, the minimum of 0.00 THB per share, the maximum of 538.00 THB per share, and the standard deviation of 56.99 THB per share. This shows that the security-by-security price factor is not normally distributed. The distribution can also be found in the security-by-security trading volume (*VLM*) and the security-by-security outstanding value by investors' portfolio (VL^{GG}, VL^{FC}, VL^{NFC}, VL^{HH}, VL^{ROW}). For the security-by-security unit change effect (net transaction) by investors' portfolio (UC^{GG}, UC^{FC}, UC^{NFC}, UC^{HH}, UC^{ROW}) and the security-by-security outstanding value by investors' portfolio (PC^{GG}, PC^{FC}, PC^{NFC}, PC^{HH}, PC^{ROW}) have better data distribution than the other factors mentioned above. However, all factors are not normally distributed, as tested in normality tests with Shapiro-Wilk, D'Agostino-Pearson and Jarque-Bera tests.

The test of correlations among the factors (Multicollinearity) uses the correlation matrix to compute the correlations among the factor pairs. The correlation matrix shows that there are two factor pairs that have a correlation of 0.8 and higher: the security-by-security unit change effect (net transaction) of household investors' portfolio (UC^{HH}) and the security-by-security outstanding value of household investors' portfolio (PC^{HH}) pair with the correlation of -0.93 and the security-by-security unit change effect (net transaction) of non-financial corporation investors' portfolio (PC^{NFC}) and the security-by-security outstanding value of non-financial corporation investors' portfolio (PC^{NFC}) pair with the correlation of -0.80. These pairs introduce the problem of multicollinearity. Therefore, the authors choose to use only the security-by-security unit change effect (net transaction) of household investors' portfolio and non-financial corporations' portfolio, since these factors reflects the investment perspective of household and non-financial corporation investors. This applies only to the moving multiple regression (MMR), since the moving artificial neural network (MANN) can mitigate the multicollinearity problem.

From the preliminary statistical analysis of the factors related to net transaction of household investor above, the authors propose a moving multiple regression showing the relationships of the 15 factors, tested multicollinearity, with the net transaction of household investor forecasted one month ahead, as shown in Equation 2.

$$\begin{aligned} UC_{t+1}^{HH} = {} & a_0 + a_1 IDX_t + a_2 IDX_{t-1} + a_3 IDX_{t-2} \\ & + a_4 VLM_t + a_5 VLM_{t-1} + a_6 VLM_{t-2} \\ & + a_{7,8,9,10,11} VL_t + a_{12,13,14,15,16} VL_{t-1} + a_{17,18,19,20,21} VL_{t-2} \\ & + a_{22,23,24,25,26} UC_t + a_{27,28,29,30,31} UC_{t-1} + a_{32,33,37,35,36} UC_{t-2} \\ & + a_{37,38,39} PC_t + a_{40,41,42} PC_{t-1} + a_{43,44,45} PC_{t-2} \end{aligned} \qquad \text{Equation 2}$$

The relationship coefficient of all factors is calculated by moving multiple regression, moving in the previous three-month period, with the pooled regression technique covering all securities. The moving multiple regression is tested for model reliability with F-statistic test, R^2, and Adjusted R^2.

From the model reliability test, we can conclude that using the F-statistic test to check the model reliability the calculated F value is in the range from 6.6787 to 18.5303, which is statistically significant at 0 percent to 5 percent. As a result, we cannot reject the null hypothesis (H_0) that no factor has influence the dynamics of net transaction of household investors forecasted one month ahead at the significance level of 95%. Additionally, we accept the alternative hypothesis (H_A) that there exists at least a factor affecting the change in net transaction of household investor forecasted one month ahead at the significance level of 95%.

The moving multiple regression forecasts the net transaction of household investors from Equation 2 is shown in Figure 1. The actual net transaction of household investor (blue line) between January 2014 to July 2016 compares with the net transaction of household investor forecasted one month ahead (red line).

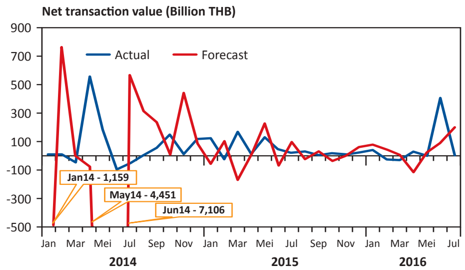


Figure 1. The Forecast of Net Transaction of Household Investor Using Moving Multiple Regression in Equation 1

Moving artificial neural network shows the relationship of the 17 factors with the net transaction of household investor forecasted one month ahead, as shown in Equation 3.

$$UC_{t+1}^{HH} = ANN(IDX_t, IDX_{t-1}, IDX_{t-2}, VLM_t, VLM_{t-1}, VLM_{t-2}, VL_t, VL_{t-1}, VL_{t-2}, UC_t, UC_{t-1}, UC_{t-2}, PC_t, PC_{t-1}, PC_{t-2}) \quad \text{Equation 3}$$

The moving artificial neural network forecasts the net transaction of household investors from Equation 3 is shown in Figure 2. The actual net transaction of household investor (blue line) between January 2014 to July 2016 compares with the net transaction of household investor forecasted one month ahead (red line).

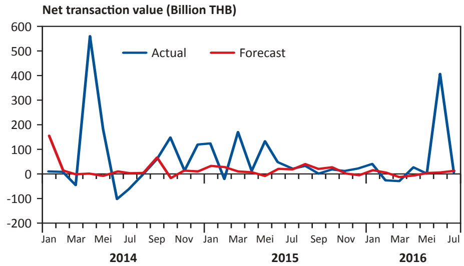


Figure 2. The Forecast of Net Transaction of Household Investor Using The Moving Artificial Neural Network in Equation 1

The performance evaluation of the forecasting models of moving multiple regression (Figure 1) and the forecasting model of the moving artificial neural network (Figure 2) are shown in Table 1.

Table 1. The Performance Evaluation of The Forecast From The Moving Multiple Regression (Figure 1) and the Forecast from The Moving Artificial Neural Network (Figure 2)
[* is 90% confidence level; ** is 95% confidence level; * is 99% confidence level]**

Performance Statistics	Moving multiple regression	Moving artificial neural network	DM Test (Probability) Significance Level
MAE	535	80	-4.738 (0.0%) ***
RMSE	1,405	144	-4.738 (0.0%) ***
MAPE	12,153%	583%	-4.738 (0.0%) ***
DA	30%	63%	

5. CONCLUSIONS

This study of the household sector investment behavior in the equities market reflects the net transaction of household investor forecasted one month ahead with the following factors, including security-by-security (stock) price (*IDX*), security-by-security (stock) trading volume (*VLM*), security-by-security outstanding value by investors' portfolio (*VL*), security-by-security unit change effect (net transaction) by investors' portfolio (*UC*), and security-by-security price change effect by investors' portfolio (*PC*). This research studies how these factors affect the net transaction of household investor forecasted one month ahead, by modeling with moving multiple regression and moving artificial neural network. This study covers monthly data from July 2013 to July 2016, a total of 37 months.

The moving multiple regression cues that the net transaction of household investor for July 2016 is correlated in the same direction as the stock price, the stock outstanding value of financial corporation, non-financial corporation, and the rest of the world investors' portfolio, the net transaction of general government and household investors' portfolio, and the stock price change effect of general government investors' portfolio. On another hand, the net transaction of household investor is correlated in the opposite direction with the stock trading volume, the stock outstanding value of general government and household investors' portfolio, the net transaction of financial corporation, non-financial corporation, and the rest of the world investors' portfolio, and the stock price change effect of financial corporation and the rest of the world investors' portfolio. However, the influential factors show no significant changes in the net transaction of household investor forecasted one month ahead.

The moving artificial neural network cues that the net transaction of household investor for July 2016 is correlated in the same direction as the stock price, the stock trading volume, the stock outstanding value of non-financial corporation and household investors' portfolio, and the stock price change effect of non-financial corporation, household, and the rest of the world investors' portfolio. On another hand, the net transaction of household investor is correlated in the opposite direction with the stock outstanding value of each investors' portfolio, the net transaction of general government, financial corporation, and the rest of the world investors' portfolio, and the stock price change effect of general government and financial corporation investors' portfolio. The stock outstanding value of non-financial corporation investors' portfolio is the most influential factor to the net transaction of household investment forecasted one month ahead. However, the stock trading volume factor shows no significant changes in the net transaction of household investor forecasted one month ahead.

The performance evaluation in predicting the net transaction of household investors' portfolio one month ahead using the forecasting models: the moving multiple regression and the moving artificial neural network is shown in Table 1. Table 1 shows that the moving artificial neural network has lower MAE, RMSE, MAPE statistics. We conclude that the moving artificial neural network has higher level forecasting accuracy. Table 1 also shows that the moving artificial neural network has higher DA statistic. We conclude that the moving artificial neural network has higher directional forecasting accuracy. Lastly, the moving multiple regression and the moving artificial neural network have the DM statistic for MAE, RMSE, MAPE at -4.738 at the 99% confidence level. We can conclude that the prediction by the moving multiple regression is statistically significantly different from the prediction by the moving artificial neural network. Finally, we can conclude that the moving artificial neural network is a more powerful forecasting model than the moving multiple regression.

REFERENCES

Atikankul, Y., 2012. *Comparison of statistical techniques for forecasting the Stock Exchange of Thailand index,* Bangkok: Faculty of Science and Technology, Rajamangala University of Technology Phra Nakorn.

Kaewmart, W. & Chancharat, S., 2013. Artificial neural network forecasting for stock indices. *KKU Res J HS (GS)*, January-April, pp. 108-118.

Khanthavit, A., 2009. *Thai stock forecasting in volatile market,* Bangkok: Faculty of Commerce and Accountancy, Thammasat University.

Pantuwon, P., 2014. *Evaluation of hybrid modeling approach for forecasting stock index of Thailand,* Bangkok: School of Development Economics, National Institute of Development Administration (NIDA).

Application of Fuzzy Time Series Model to Forecast Indonesia Stock Exchange (IDX) Composite

Tri Wijayanti Septiarini*
Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Pattani, Thailand – triwijayantiseptiarini@gmail.com

Salang Musikasuwan
Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Pattani, Thailand– salang.m@psu.ac.th

Abstract

The aim of this research is to propose the fuzzy time series model for forecasting the Indonesia Stock Exchange (IDX) composite. The Mamdani inference method has been employed into the model by using centroid defuzzification. After the model has been implemented and tested, the MAPE and MSE were used to evaluate the performance of the fuzzy model. The result showed that the fuzzy time series model is a good tool for forecasting with the value of MAPE for training and testing data are 3.12% and 1.48%, respectively. Finally, it can be recommended that to improve the performance of the model, the optimization technique should be employed into the proposed model.

Keywords: Forecasting model, Fuzzy time series, IDX composite.

1. INTRODUCTION

Indonesia Stock Exchange (IDX) Composite is an index of all stock listed on the Indonesia Stock Exchange. Indonesia Stock Exchange is a stock exchange based in Jakarta, Indonesia. It was previously known as Jakarta Stock Exchange (JSX) and its name changed in 2007 after merging with Surabaya Stock Exchange (SSX). Indonesia Investments' Stock Market Update section contains a daily analysis regarding the performance of the Indonesia Stock Exchange. Indonesia Stock Exchange (IDX) is highly affected by foreign investor that can threat the economy. Domestic investor tends to invest in risk-free asset such as deposit in the bank since they are not familiar with the stock market and anxious about the risk (risk-averse type of investor). Therefore, it is important to educate domestic investor to involve in the stock exchange. Indonesia Stock Exchange is highly affected by foreign investor due to the small number of domestic investor. In 2009, it is approximately 300,000 investors (only about 0.1% from total population) which is fewer than other Asian countries such as Singapore (about1.26 million people or 30% from total population), and Malaysia (about 3 million people or 12.8% from total population).

The prediction is an approximate (measurement) process about the amount of something in the future based on past data that be analyzed scientifically, especially using statistical methods. The prediction is important in financial field. The investor can use prediction result to decide policy in the future. After the data predicted is obtained, everyone who has interest in this case, can take action to reduce significant losses. For example, the multi-national company can be determined short-term financing decisions, short-term investment decisions, capital budgeting decisions, long-term financing decisions and judgments that all decisions profit is influenced by changes in currency exchange rates.

In order to stabilize the economy and the IDX is not highly affected by foreign investors, Indonesia has to escalate the number of its domestic investors by educating them. For all this time, domestic investors in Indonesia are risk-averse investors. Analyze data used mathematical model is one way to educate domestic investors. The using of appropriate methods to predict have several influence factors, that are time, the pattern of data, the relationship between the previous data and the accuracy of prediction. This research proposed a fuzzy time series model to forecast Indonesia Stock Exchange (IDX) composite in order to make market participants are easier to decide for do all economic activity and the accuracy of fuzzy time series model in predicting IDX composite will be evaluated.

Fuzzy time series method is a dynamic process that used linguistic values as observations. This method has been proposed by proposed by Song, Q., &Chissom, B.S. (1993) to predict the number of enrollment in a university. Chen (1996) developed Song method which is further developed by Yu (2005). Again, Song method is developed by Lee, M.H. &Suhartono (2012) talking about the seasonal data and Suhartono, &Lee, M.H. (2011) proposed forecasting seasonal method and trend data. A'yun *et al.* (2015) applied weighted fuzzy time series to forecast Trans Jogja's passengers. Rubio, A. *et al.* (2016) proposed forecasting portfolio return using weighted fuzzy time series methods. And Septiarini, T.W. *et al.* (2016) applied wavelet fuzzy model to forecast the exchange rate IDR of USD. The data that used in this research is daily data of IDX composite.

2. FUZZY TIME SERIES

Fuzzy time series is the development of fuzzy system using time series data. The basic step of forecasting process in fuzzy time series are the following:

1. Define the universal set
2. Define fuzzy set
3. Determine the membership function of fuzzy set
4. Determine the relation of fuzzy
5. Defuzzification and forecasting

Let us briefy review some prelimenry definitions.

Definition 2.1 Zadeh (1965)

Let $U=\{u_1, u_2,\ldots,u_m\}$ the universe of discourse. A fuzzy set A of the universe of discourse U can be defined by its membership function, $\mu_A: U \rightarrow [0,1]$, $\mu_A(u)$ being the degree of membership of the element $u \in U$ in the fuzzy set A.

Definition 2.2 Song, Q., &Chissom, B.S. (1993)

Let $Y(t)$ be a subset of real number, for t = ..., 0, 1, 2, ... being the universe of discourse in which the fuzzy sets $A_i(t)$ defined, i = 1,2,...,m Let $F(t)$ be a collection of fuzzy sets $A_i(t)$, i = 1,2, ..., m then $F(t)$ is called fuzzy time series on $Y(t)$.

Definition 2.3 Chen (1996)

Let $F(t)$ be a fuzzy time series. If for any time, $F(t) = F(t - 1)$ and $F(t)$ only has finite elements, then $F(t)$ is called a time-invariant fuzzy time series. Otherwise, it is called a time-invariant fuzzy time series.

According to Definition 2.3, the data on this research have to stationary data.

Definition 2.4 Song, Q., &Chissom, B.S. (1993)

Let $F(t - 1) = A_i$ and $F(t) = A_k$, for t = 1,2, ..., N and i, k = 1,2, ..., m The relationship between two consecutive observations, $F(t - 1)$ and $F(t)$, is denoted by $F(t-1) \rightarrow F(t)$ or by $A_i \rightarrow A_k$, where A_i is called the left-hand side and A_k the right hand side of the fuzzy logical relationship (FLR)

3. THE FORECASTING OF IDX WITH FUZZY TIME SERIES

The processes for forecasting IDX using fuzzy time series are given as follows :

Step 1. Determine stationary data

The data used in this research is obtained from « www.yahoofinance.com » in the period of December, 7th 2015 to September, 21st 2016. The number of data is 200 daily data. The data will be divided into 150 training data and 50 testing data. The data that use for forecasting have to be stationary. In order to obtain stationary data, the data will be differentiated. The plot of all original IDX data is shown in Fig. 1., while Fig. 2. show the plot of the diferentiation for IDX data which used in fuzzy time series model.

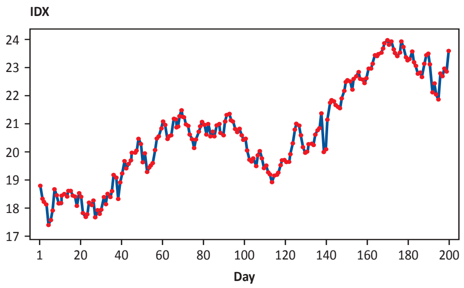


Figure 1. The Plot of IDX Data

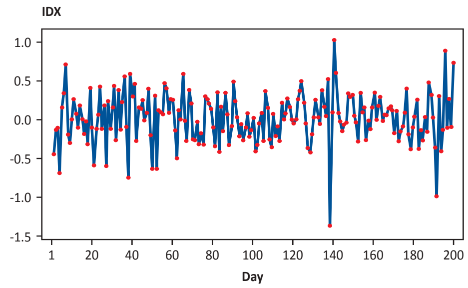


Figure 2. The Plot of IDX Data After Being Differentiated

Step 2. Define the universal set of discourse U

Based on stationary data, the data is obtained with the smallest and the largest in the range of -1.37 and 1.029, respectively. So, the universal of discourse is U=[-1.7127, 1.0290].

Step 3. Determine the membership function of fuzzy set

This model uses triangular membership function for input and output.

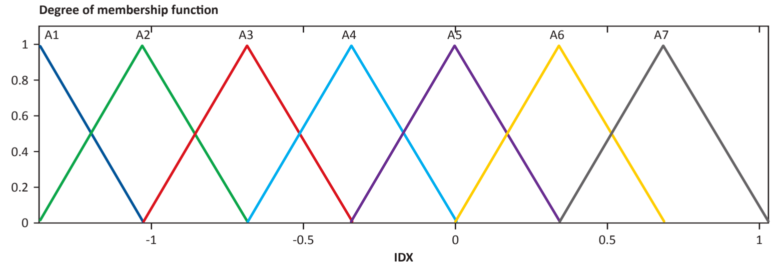


Figure 3. Triangular Membership Function of Input and Output

From Fig 3, we can see that there are 7 fuzzy sets, such are A1, A2, A3, A4, A5, A6, and A7. They all have the same length. This method is refered from Chen (1996) Fuzzy Time Series method.

Step 4. Determine the relation of fuzzy logic (rules)

The relation of fuzzy logic is establish by training data and conducted based on fuzzy logic, $F(t-1) \rightarrow F(t)$ or by $A_i \rightarrow A_k$, this refered from Definition 2.4. In this research, there are 17 rules that had been elimininated in similar case as shown in Table 1.

Table 1. Fuzzy Rules

No	A_i	A_k
1	A1	A5
2	A1	A6
3	A3	A5
4	A3	A6
5	A4	A3
6	A4	A4
7	A4	A5
8	A4	A6
9	A5	A1
10	A5	A3
11	A5	A4
12	A5	A5
13	A5	A6
14	A6	A1
15	A6	A4
16	A6	A5
17	A6	A6

Step 5. Defuzzification and forecasting

We apply defuzzfication process using Matlab R2013a. After we obtained the result of defuzzification of stationary data, the result have to converse to the original data form for forecasting result of IDX. We can compare the value of original and forecasting data.

4. RESULTS

Once the fuzzy time series model has been implemented, the performance of the model has been evaluated with both training and testing data of IDX composite. Table 2 shows the value of predicting comparing with real value of the testing data. We can compare the value of original and forecasting data in plot time series, as shown in Fig. 4.1. and Fig. 2.

Table 2. The Forecasting Result of Testing Data

No	Original Data	Forecasting Data	No	Original Data	Forecasting Data
1	22.55	22.264	26	23.52	23.095
2	22.51	22.225	27	23.92	23.195
3	22.23	22.185	28	23.74	23.699
4	22.58	22.004	29	23.36	23.421
5	22.68	22.358	30	23.26	23.189
6	22.84	22.355	31	23.30	22.935
7	22.58	22.511	32	23.56	22.975
8	22.57	22.337	33	23.19	23.292
9	22.45	22.245	34	23.05	23.019
10	22.61	22.124	35	22.79	22.723
11	22.96	22.281	36	22.82	22.547
12	22.96	22.738	37	22.66	22.495
13	23.14	22.634	38	23.14	22.331
14	23.43	22.816	39	23.46	22.917
15	23.42	23.178	40	23.48	23.134
16	23.48	23.095	41	23.12	23.155
17	23.54	23.155	42	22.13	22.76
18	23.69	23.215	43	22.43	22.303
19	23.86	23.362	44	22.02	22.184
20	23.98	23.53	45	21.88	21.849
21	23.81	23.654	46	22.78	21.553
22	23.92	23.479	47	22.68	22.609
23	23.65	23.594	48	22.95	22.355
24	23.50	23.416	49	22.86	22.687
25	23.42	23.172	50	23.59	22.535

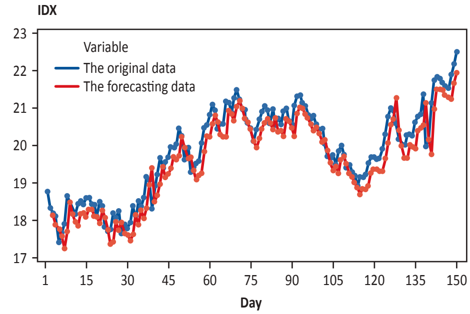


Figure 4. The Plot of Original Training Data and Forecasting Result

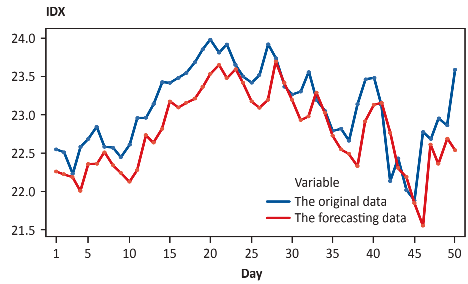


Figure 5. The Plot of Original Testing Data and Forecasting Result

In Table 3, the comparison value of MAPE and MSE for training and testing data that is used to analyze fitted fuzzy time series in IDX composite. It is shown that MAPE and MSE on the testing data (1.48%, 18.9329) is lower than in the trainging data (3.12%, 478.3445).

Table 3. The Value of MAPE and MSE

Training Data		Testing Data	
MAPE	MSE	MAPE	MSE
3.12%	478.3445	1.48%	18.9329

5. CONCLUSIONS

By using fuzzy time series model to predict the short term value of IDX, the predicting result obtained a level of accuracy MAPE of 3.12% for training data and 1.48% for testing data. It means that the forecasting model is acceptable to forecast the future. Fuzzy time series is a good forecast model since the value of MAPE is less than 5%.

The result of the forecasting model in this research is based on the data time series of the IDX composite only. To improve the accuracy of prediction, then further research would consider the other factors that influence the IDX composite such as the changes in a country's financial policy, interest rates, political changes in the country, and the others unpredictable factors. Finally, it can be suggested that by employing the optimization techniques can also improve the performance of the model.

REFERENCES

A'yun, Kurrotul, Abadi, A.M., &Saptaningtyas, F.Y. (2015). Application of weighted fuzzy time series model to forecast Trans Jogja's passengers. *International Journal of Applied Physics and Mathematics*, 5(2), 76-85.

Chen,S.M.(1996).Forecastingenrollmentsbasedonfuzzytimeseries.*FuzzySetsandSystems*,81(3),311–319.

Lee, M.H. &Suhartono. (2012). A weighted fuzzy time series for forecasting seasonal data. *Journal of Quality Measurement and Analysis*, 8(1), 85-95.

Rubio, A., Bermudez, J.D., &Vercher, E. (2016). Forecasting portfolio return using weighted fuzzy time series methods. *International Journal of Approximate Reasoning*, 75, 1-12.

Septiarini, T.W., Abadi, A.M., &Taufik, M.R. (2016). Application of wavelet fuzzy model to forecast the exchange rate IDR of USD. *International Journal of Modeling and Optimization*, 6(1), 66-70.

Suhartono, &Lee, M.H. (2011). A hybrid approach based on winter's model and weighted fuzzy time series for forecasting trend and seasonal data.*Journal of Mathematics and Statistics*, 7(3), 177-183.

Song, Q., &Chissom, B.S. (1993). Forecasting enrollment with fuzzy time series-part I. *Fuzzy Sets System*, 54, 1-9.

Yu, H.K. (2005). Weighted fuzzy time series models for TAEIX forecasting. *Phisica A: Statistical Mechanics and Its Applications*, 349(3-4), 609-624.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.

Forecasting of Dow Jones Industrial Average by Using Wavelet Fuzzy Time Series and ARIMA

Muhamad Rifki Taufik[1], Lim Apiradee[2], Phatrawan Tongkumchum[3], Nureen Dureh[4]
[1,2,3,4] Research Methodology, Math and Computer Science Department
Faculty of Science and Technology, Prince of Songkla University, Thailand. –
[1] m.rifki.taufik@gmail.com

Abstract

The Dow Jones Industrial Average (DJIA) is one of the oldest, single most-watched indices in the world and includes companies such as General Electric Company, the Walt Disney Company, Exxon Mobil Corporation and Microsoft Corporation. DJIA had strong affect for stock market in the world. This study conducted to forecast DJIA by using Wavelet Fuzzy Time Series and Autoregressive Integrated Moving Average (ARIMA). The data were obtained from 1/29/1985 until 12/14/2016. The result was comparing between both methods by calculating the MAPE which measured the accuracy of prediction and ARIMA had 2.6% and Wavelet Fuzzy Time Series had 1.4% of MAPE.

Keywords: ARIMA, DJIA, Fuzzy Time Series, Wavelet.

1. INTRODUCTION

In the globalization, stock market had strong role for fundamental policy and economic development. Inflation and increasing of exchange rate could be caused by stock market movement. Numerous factors influence stock market performance, including political events, general economic conditions, and trader expectations (Abbasi et al 2015). Based on Clements, economics fundamentals had recently become so important (Clements et al, 2009). One of the well-known stock market and the oldest one was Dow Jones.

In 1884 Charles Dow, one of the founders of Dow Jones & Co, created the first stock market index which published by Barron's and The Wall Street Journal. The index used price weights instead of conceptually superior market valuation weights. The Dow Jones Industrial Average (DJIA) was becoming the most quoted stock market index in the world. The index changes were believed be representative of entire stock market index (Shoven & Sialm, 2000). Forecasting on time series data apparently has important part.

Predicting financial market changes is an important issue in time series analysis, receiving an increasing attention in last two decades. Financial time series prediction is an important subject for many financial analysts and researchers as accurate forecasting of different financial applications play a key role in investment decision making (Grigoryan, 2015). A time series is a set of well-defined data items collected at successive points at uniform time intervals (Mondal, Shit, & Goswami, 2014). Time series analysis is an important part in statistics, which analyzes data set to study the characteristics of the data and helps in predicting future values of the series based on the characteristics.

Nowadays forecasting being important part of researchers could improve the domain field of existing predictive models. Not only Investment decisions and planning ability but also develop effective strategy about their daily and further effort (Adebiyi et al, 2014). Investors wish that they hold the forecasting method which can guarantee easy profiting and minimize risk on their stock market investment. That is the reasons why forecasting of stock price becomes an important topic in finance and economics over year to get better predictive models.

2. LITERATURE REVIEW

Lately, some predictive models have been presented especially in stock market movement. Nowadays researchers are interpreting and summarizing the models in different way to find out better result (Granger, 1992). The better prediction apparently is still improving which the best predictive models is not developed yet.

If the time series analysis model is divided into two categories; linear and nonlinear model, then in many cases it will be difficult to determine whether a time series model belongs to a linear model or a nonlinear model and then it is also difficult to determine which model should be used in the study. In practice, there is few time series with pure linear or nonlinear feature. There is no single model that can adapt to all situations and solve the problem (Jia, Wei, Wang, & Yang, 2015).

Autoregressive Integrated Moving Average (ARIMA) models have been used in time series prediction of stock market. ARIMA are from statistical models perspectives. ARIMA models are known to be robust and efficient in financial time series prediction. Jaret (2011) demonstrated the usefulness of ARIMA-Intervention time series analysis as both an analytical and forecast tool. Pual et al (2013) also examined empirically the best ARIMA model for forecasting. Hence, ARIMA (2, 1, and 2) is found as the best model for forecasting the Square Pharmaceuticals Limited (SPL) data series.

A general ARIMA formulation is selected to model the price data. Jadhav et al (2015) did a selection which was carried out by careful inspection of the main characteristics of the hourly price series. In most of the competitive electricity markets this series presents: high frequency, nonconstant mean and variance, and multiple seasonality (corresponding to daily and weekly periodicity, respectively), among others.

The second model is combination between Wavelet Transformation and Fuzzy Time Series Analysis. Since the 2000s, wavelet decomposition has been combined with time series models as a preprocessing method. Wavelet decomposition (or wavelet transform) decomposes time series data into approximation and detail components, so that different forecasting models can be applied to each component (Jin & Kin, 2015). This property can improve the performance of forecasting. The validity of this approach has been proved in various studies.

Since fuzzy time series model was proposed by Song and Chissom in 1993, there are many forecasting models have been developed to deal with the forecasting problems (Qiu, Zhang, & Ping, 2015). Qiu et al did a test of the effectiveness of the Fuzzy model, the proposed method is demonstrated on the procedure of forecasting enrollments as well as its experiment on forecasting the close price of Shanghai Stock Exchange Composite Index. Empirical analyses show that the proposed method gets a higher average forecasting accuracy rate than the existing methods.

Rostamy et al (2013) did a research with the major purpose was to predict the total stock market index of Tehran Stock Exchange, using a combined method of Wavelet transforms, Fuzzy genetics, and neural network in order to predict the active participations of finance market as well as macro decision makers.

ARIMA and WFTS are interesting models based on author. Both performance give acceptable predictive models. This paper showed the comparison between two models in current period. The accuracy of prediction had been measured by Mean Absolute Percentage Error (MAPE).

3. DATA MANAGEMENT AND METHODOLOGY

Data management

Dow Jones Industrial Average (DJIA) data had been obtained from yahoo finance websites. The data structure is as time series data which consisted of 5 work-days from January 29th 1985 until December 14th 2016. The data provided some type of prices such as open, high, low, close, volume and adjusted price. The data that been chosen is adjusted closing price column. An adjusted closing price is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open.

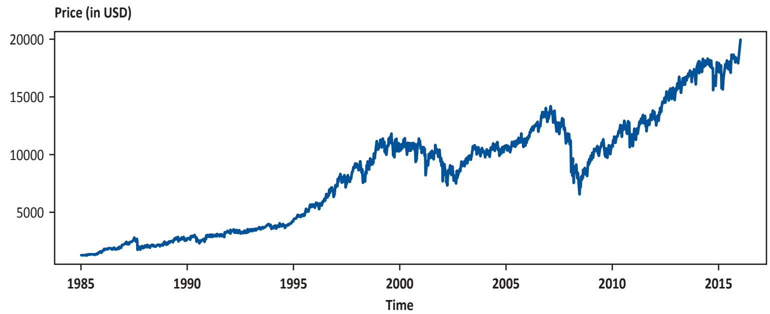


Figure 1. DJIA Time Series Plot

Autoregressive Integrated Moving Average

ARIMA is a basic linear forecasting model, which uses a lagged series. Because of its simplicity and good performance, ARIMA has been applied to many time series analyses (Jin & Kin, 2015). ARIMA model is derived by general modification of an autoregressive moving average (ARMA) model. Based on Mondal et al (2014) ARIMA models are generally used to analyze time series data for better understanding and forecasting. This model was calculated using R i386 3.2.1 version.

Initially, the appropriate ARIMA model has to be identified for the particular datasets and the parameters should have smallest possible values such that it can analyze the data properly and forecast accordingly. This model type is classified as ARIMA (p,d,q) where p denotes the partial autocorrelation parts of the data set, d refers to differencing part of the data set and q denotes autocorrelation parts of the data set and p, d, q is all nonnegative integers. ARIMA (p,d,q) forecast 30 days based on previous data pattern.

Wavelet

The performance of the wavelet transformation was superior in all cases, whereas the application of detail components in the forecasting was only able to yield a small improvement in forecasting performance (Jin & Kin, 2015). Therefore, forecasting with only an approximation component would be acceptable, in consideration of forecasting efficiency.

Discrete Wavelet Decomposition is a preprocessing method that projects a time series onto a collection of orthonormal basis functions. This transformation is applied to DJIA time series data to obtain further information from the time-domain original data. First After applying DWD to the data, data was analyzed signals by decomposing them into various frequencies or levels. In simple, the wavelet transformation is used to transform a function with the time domain into the frequency domain. This study conducted Wavelet Transformation used Haar Wavelet. 8000 data as training data was used to decompose in some levels. Number of level we got from this formula:

$$level_{\max} = \frac{\ln\left(\frac{data\ size}{filter\ size - 1}\right)}{\ln 2}$$

Wavelet Fuzzy Time Series

This model combined between Wavelet Transformation and Fuzzy Time Series on decomposition part. The purpose of wavelet fuzzy models was to predict the data of the components Discrete Wavelet subseries (DWS) obtained by using DWT on original data. Each DWS form of time series data and had a different effect and frequency with the original data. This model was calculated using matlab R2013a version. Training data as around 8000 observation forecast 30 days further data as testing data. DWS corresponding selected as input into this model. The authors use this model based on the research of Septiarini et al (2016).

MAPE (Mean Absolute Percentage Error)

The MAPE (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error. This method is useful for evaluating the accuracy of prediction. MAPE indicates how big error in the prediction comparing with real value. MAPE also has ability to evaluate the precision of techniques or models as percentage of mean absolut error. MAPE is calculated as following formula :

$$\frac{1}{n}\Sigma\frac{|Actual - Forecast|}{|Actual|} * 100\%$$

4. RESULT AND DISCUSSION

4.1. ARIMA

Differencing a Time series

ARIMA models were defined for stationary time series. Therefore from the Figure 1, the data started off with a non-stationary time series, time series data first needed to 'difference' the time series until the data was obtained a stationary time series. Then the data had to difference the time series d times to obtain a stationary series, then we got an ARIMA (p,d,q) model, where d is the order of differencing used. DJIA times series plot of adjusted price consist of 8038 records was showed in Figure 1. The figure identify was that this time series looked like random fluctuation in over time. Then, the time series of first differences as Figure 2 to be statuinary in mean and variance, and so an ARIMA (p, 1, q) model was probably appropriate for the DJI data. The next step was to figure out the values of p and q for the ARIMA model.

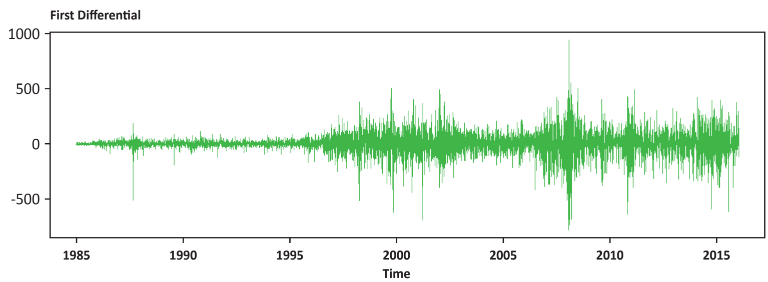


Figure 2. First Order Differencing

Selecting a Candidate of ARIMA Model

After transforming the DJIA data to be stationary time series by differencing one time, the next step is to select the appropriate ARIMA model, which means finding the value of most appropriate values of p and q. The data was examined the correlogram and partial correlogram. As Figure 3, plot of correlogram that the autocorrelation at lag 1, 5, 12, 15 and lag 18 (-0.054, 0.039, 0.036 -0.044, and -0.034) exceeded the significance bounds. There were 5 lags out of significant border.

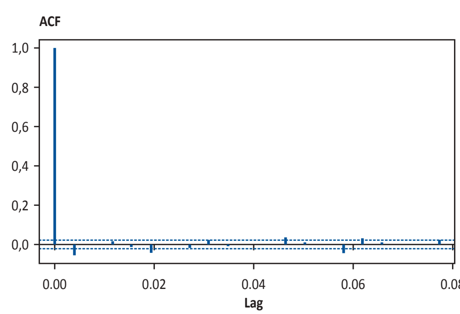


Figure 3. ACF

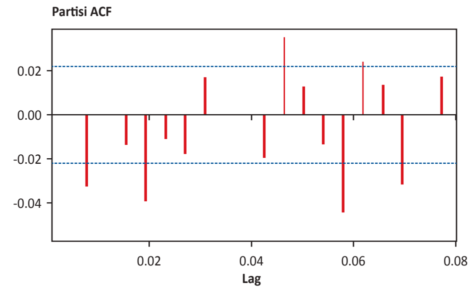


Figure 4. Partial ACF

Then as Figure 4, the partial correlogram shows that the partial autocorrelations at lags 1, 2, 5, 10, 12, 15, 16 and 18 exceeded the significance bounds, and those were negative. Then we got 8 lags which outside the bounds. Probably the appropriate ARIMA model was (8, 1, 5).

4.2. Wavelet

Wavelet Level

The number of training data was 8000 records. Then the data was filtered into two part as decomposition and approximation. After calculating the formula, then the rounding levels were consist of 12 levels.

Figure 5. Wavelet Decomposition

The data had been separated into 12 levels called Discrete Wavelet Decomposition. Every level was measure the association using Pearson Correlation Coefficient, then two strongest correlation which more than 0.4 were D12 and A12. Both series had been combined to be Discrete Wavelet Series (DWS).

From DWS, 8000 new records put in Fuzzy Inference System as training data. Universal set was in interval Max and min values of Discrete Wavelet Series. Referred on Chang's Method, universal set had to been separated as seven membership function using Gaussian membership function.

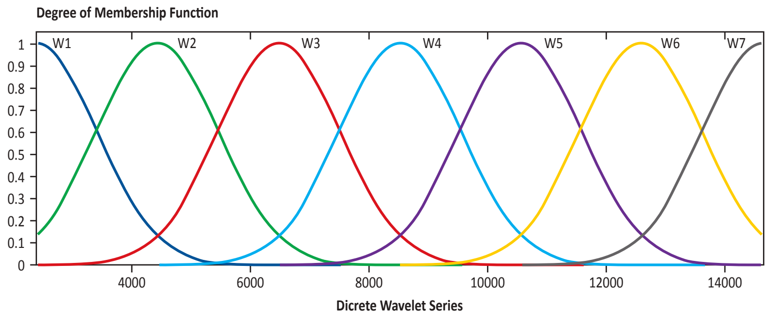


Figure 6. Seven Membership Functions

Mamdani method was used on this Fuzzy Inference System (FIS). FIS needed rules which the rules had been created from DWS, then created seven rules. Last the defuzzification forecast the 30 records based on rules then the obtained data re-transform to be actual data as predicted data.

4.3. Comparison

Predicted data as forecasting 30 days record of Dow Jones Industrial Average showed both model had small MAPE. MAPE for ARIMA was 2.6% and MAPE for WFTS was 1.4%. In this case, ARIMA and WFTS could predict next 30 days record with acceptable performance.

5. CONCLUSIONS

Autoregressive Integrated Moving Average (ARIMA) and Wavelet Fuzzy Time Series (WFTS) were acceptable predictive model. ARIMA model had bigger error than WFTS, but based on prediction the ARIMA model might be not appropriate to predict long term period of time series data because the forecasting mostly constant which would give bigger error for longer time. WFTS had good pattern track, it gave chance to predict longer period of time.

ACKNOWLEDGMENTS

This study was supported by the Higher Education Research Promotion and The Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission. We would acknowledge the Department of Mathematics and Computer Science of Prince of Songkla University for providing the plotform for this study.

REFERENCES

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. *16th International Conference on Computer Modelling and Simulation* (pp. 105-111). UK: Computer Society.

Clements, M. P., Milas, C., & Dijk, D. v. (2009). Forecasting returns and risk in financial markets using linear and nonlinear models. *Science Direct: International Journal of Forecasting*, 215–217.

Granger, C. W. (1992). Forecasting stock market prices: Lessons for forecasters. *International Journal of Forecasting*, 3-13.

Grigoryan, H. (2015). Stock Market Prediction using Artificial Neural Networks: Case Study of TAL1T, Nasdaq OMX Baltic Stock. *Database Systems Journal vol. VI*, 14-23.

Jadhav, S., Kakade, S., Utpat, K., & Deshpande, H. (2015). Indian Share Market Forecasting with ARIMA Model. *International Journal of Advanced Research in Computer and Communication Engineering*, 334-336.

Jia, C., Wei, L., Wang, H., & Yang, J. (2015). A Hybrid Model Based on Wavelet Decomposition-Reconstruction in Track Irregularity State Forecasting. *Hindawi: Mathematical Problems in Engineering*, 1-13.

Jin, J., & Kin, J. (2015). Forecasting Natural Gas Prices Using Wavelets, Time Series, and Artifcial Neural Networks. *Plos*, 1-13.

Mondal, P., Shit, L., & Goswami, S. (2014). STUDY OF EFFECTIVENESS OF TIME SERIES MODELING (ARIMA) IN FORECASTING STOCK PRICES. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 13-29.

Rostamy, A. A., Abbasi, N. M., Aghaei, M. A., & Fard, M. M. (2013). Forecasting Stock Market Using Wavelet Transforms and Neural Networks: An integrated system based on Fuzzy Genetic algorithm (Case study of price index of Tehran Stock Exchange). *International Journal of Finance, Accounting and Economics Studies*, 83-94.

Septiarini, T.W, Abadi A.M & Taufik M.R. 2016. Apprication of Wavelet Fuzzy Model to Forecast the Exchange Rate IDR to USD. *International Journal of Modeling and Optimization,* 66-70.

Shoven, J. B., & Sialm, C. (2000). The Dow Jones Industrial Average: The Impact of Fixing its Flaws. *THE JOURNAL OF WEALTH MANAGEMENT*, 9-18.

This page intentionally left blank

CPS15: Statistical Theory & Methodology (2)

Outlier Detection Framework for Extreme Fluctuation Time Series Data

Brilian Surya Budi
Badan Pusat Statitik, Sorong, Indonesia – brilian.surya@bps.go.id

Abstract

Outlier is an indication of abnormality appears in a data set. If the data set is obtained from the results of observations, then every possible outlier leads us to abnormal phenomenon. One of the implementations is price fluctuation monitoring towards chili commodity in Province of DKI Jakarta. Empirically, the price of chili commodity provides rapid response towards shocks (especially from the aspect of supply such as the problems of crop and distribution) that makes the price fluctuations of chili commodity change rapidly (volatile) and extremely. This present problem is whether it can be considered outlier when the price of chili commodity reaches its extreme fluctuation peak since false alarm makes a policy less accurate. The focus of this paper is to discuss a framework to detect outlier of data time series that have extreme fluctuation by conducting case studies towards the prices of chili commodity in Province of DKI Jakarta.

Keywords: Outlier Detection; Time series; Extreme fluctuation.

1. INTRODUCTION

At the present time, we are entering a big data era, where the data can be obtained from various sources with a large number. With the available of numerous data, researchers are more freely to do an extensive data exploration. This moment is accompanied by the growing popularity of data mining as a tool to analyze large data. Data mining is a powerful knowledge discovery tool useful for modeling relationships and discovering hidden pattern in large database (E. Kough, 2005). One of many topics often being raised on data mining is outlier detection. Outlier detection can be called as anomaly detection, event detection, novelty detection, deviant discovery, change point detection, fault detection, intrusion detection or misuse detection. Theoretically, the definition of each other terms is different. However, in general term, the purpose is to identify an abnormality of an observation or a set of observation compared with the majority of other observations in a set of data. There are many definitions of outlier, nevertheless there is no a definition that can be accepted universally. According to Barnett and Lewis (1994), the letter definition of an outlier is an observation (or subject of observation) which appear to be inconsistent with the reminder of that set of the data. Therefore, an outlier itself is an observation's result which significantly different with the majority of observation in a data set.

The main cause possibilities of outlier in a data set are the occurrence of error and event. The error that occurred can be caused by human mistakes, mechanism mistakes and so forth that cause abnormality or deviation which affects the observation's results to be significantly different with other observations. While an outlier that caused by the occurrence of event can be identified from an unexpected pattern that unusually happened. It can lead us to a phenomenon that can become valuable information which needs a further investigation of the causes of the event.

The implementation of outlier detection has been widely used. As an example, which often to be discussed is Fraud detection on credit cards. The abnormal purchasing or using pattern of credit cards which unusual from previous patterns can shows a deviation in the using of credit cards. The deviation can indicates the fraud on credit cards, thus, the banks can take an action to prevent or to warn the occurrence of fraud. This approach can helps to detect others commercial fraud such as the using of mobile phones, insurance claims and other commercial transactions.

The focus on this paper is how to detect outlier in the fluctuation prices of chili in the Province of DKI Jakarta. The price of chili in the Province of DKI Jakarta is highly fluctuated. In the second semester of 2016, the price of chili increased from IDR 33,588 on 1st July 2016 to IDR 70,372 on 18th October

2016. There are various factors that affect the increasing price of chili commodity. Other than the nature of chili which is nonperishable, according to the working paper of BI, the problem of supply factors are the main factor that affecting the fluctuation price of food commodities, whereas the demand side is less volatile. Supply factors in question include the distribution issues of food commodities which were impeded due to the weather; the commodities are not entering the harvest season yet while the stock has been depleted, crop failures in some areas, and many more. The use of outlier detection on chili price may become a warning that the increasing of the price is abnormal and requires special treatment from the government to normalize the price.

2. TYPE OF TIME SERIES DATA

Before further discussion regarding the outlier detection, we will review the types of time series data first. According to Cheboli (2010), there are two characteristic types of time series data, which are periodically and synchronously. Thus, there are four different combinations from those characteristic keys as follows:

- Periodic and Synchronous

  This type of time series data has a constant period of time and in conformity for temporary (every series started from the same time instance).
- Aperiodic and Synchronous

  This type of time series data has no clear period of time, however, in conformity for temporary (every series of data started from the same time instance).
- Periodic and Asynchronous

  This type of time series data has a clear period of time, however, in temporary is not in conformity (not started from the same time instance).
- Aperiodic and Asynchronous

  This type of time series data has no clear period of time and in temporary is not in conformity (not started from the same time instance).

From the four types of data described above, the suitable model for series data of the chili's price is the Aperiodic and synchronous type. Because the fluctuation on the price of the commodity of chili has no clear period of time, but almost in the same time for every year there is a price increase such as on Eid Fitr and at the end of year. These happen because on Eid Fitr there is an increase of demand from the society and at the end of the year it's more likely to the supply of chili commodity (production decrease and distribution issues due to the weather).

3. DETECTING OUTLIER ON TIME SERIES DATA

There are lot of theories have been developed in data mining to detect outlier on time series data with various approaches. One of the many techniques that often being used in univariate time series data is the Median Absolute Deviation (MAD). This method detects observations which considered being a possible outlier outside cut off values of the calculation results. The MAD itself basically sees the value that has been deviated far away from the center of data distribution. Median was used in this method as the center of data distribution. The advantage of median as the center of data distribution compared with mean is that median is not sensitive with the value of outliner. It's because the mean was formed from the observation values, so that, if an outlier occur on the data, the center of data distribution tends to the outlier (sensitive to the outlier value). While the median is obtained from the sequence of the center of data distribution, so if the outlier occurs on the center of data distribution, the data will not tend to the outlier.

The formula of MAD is as follows:

$$MAD = b\, M_i\left(\left|x_i - M_j\left(x_j\right)\right|\right)$$

where x_j is the n original observation and M_I is the median of the series. Usually b = 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outlier (Roussew & Criux, 1993). From the formula above, can be described be the following algorithm.

Algorithm 1: Finds the MAD value

- Input : Univariate time series data
- Output : Median Absolute Deviation value
    1. Sorting ascending of data
    2. Finding Median (M) value from dataset
    3. **for** each x_i where i ≤ n
    4. **looping (1) :**
    5. absolute (xi – M)
    6. **end loop (1)**
    7. Sorting ascending the series of absolute value
    8. Finding Median from the series of absolute value
    9. Median from the series of absolute value multiplied by 1.4826

After the MAD value was obtained, we must define the rejection criterion of a value. As for the mean and standard deviation, it is necessary to define a level of decision: This remains the unavoidable subjective aspect of the decision. Depending on the stringency of the researcher's criteria, which should be defined and justified by the researcher, Miller (1991) proposes the values of 3 (very conservative), 2.5 (moderately conservative) or even 2 (poorly conservative). So, the following formula was used.

$$M - 2.5{*}MAD \leq x_j \leq M + 2.5{*}MAD$$

Cut off value that was obtained will be used as upper and lower normal limit. Outside upper and lower limit is the possible outlier within the dataset. The value bellow M–2.5*MAD or the value above M+2.5*MAD can be estimated as possible outlier.

4. EXPERIMENT

This experiment used the data of red chili pepper commodity in the Province of DKI Jakarta since 1st May 2016 to 15th January 2017. The source of data was obtained from http://infopangan.jakarta.go.id/. The following is the graph of the fluctuation of chili price in the Province of DKI Jakarta since 1st May 2016 to 15th January 2017.

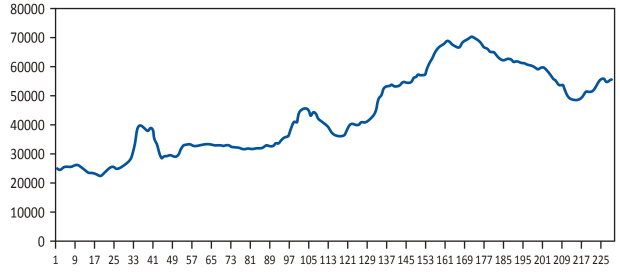


Figure 1. Price Fluctuation of Chili Comodity in DKI Jakarta Province1st May 2016 to 15th January 2017

From the data above, it can be concluded that there was a significant increase on the price of chili commodity in the Province of DKI Jakarta which is IDR 22,930 on May 18th 2016 to 70,372 on 18th October 2016. From the data, the median value is 40907 and MAD value of 17411.65 was obtained. Therefore, the rejection criterion value of upper limit is 75730.3 and the lower limit is 6083.7. Thus, from the data set, there is no possible outlier. So we can conclude that although the data is volatile, not necessarily the data with extreme value is an outlier

5. CONCLUSIONS

Outlier detection is one topic in data mining is widely implemented. One implementation is the outlier detection in chili commodity prices in the province of DKI Jakarta. The function of outlier detection at the price of chili is to see whether the price fluctuations that are considered extreme outliers is present or not. Because if these fluctuations simply calculated by personal judgment may be extreme price rise is not an outlier as the results in the experiment. Median Absolute Deviation (MAD) as a method that is used as a tool to detect outliers have ease in calculation and are not sensitive to outliers. So this method is easy to use on very large data

REFERENCES

Aggarwal, C.C. (2015). *Data Mining "The Textbook"*. New York. Springer.

Barnett V. and Lewis T. (1994). *Outlier Detection 3rd edition.* New York. John Wiley

Cheboli, Deepthi. (2010). *Anomaly detection of time series*. Diss. University of Minnesota

Kliest, Caroline. (2015). Time Series Data Mining Methods: A review. Berlin. Humboldt-Universitat zu Berlin

E. Keogh., J. Lin, and A. Fu. (2005). *Hot SAX: Efficiency finding the most unusual time series sub-sequence*. In Proc. IEEE Int. Conf. Data Mining, Houston, TX, USA, pp 226-223

Gagare, Selochana M. and Prof Natikar, S.B. (2015). *An Improved Framework For Outlier Periodic Pattern Detection in Time Series using Walmat Transaction Data*.

Leys, Christoper. (2013). Detecting outliers: *Do not use standard deviation around the mean, use absolute deviation around the median*. Journal of Experimental Social Psychology.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. The Quarterly Journal of Experimental Psychology, 43(4), 907–912,

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation.Journal of the American Statistical Association, 88(424), 1273–1283,

On the Efficiency of the Cochrane-Orcutt and Prais-Winsten Regression for AR (1) Model

Jessa Jane D. Langoyan*
University of the Philippines Cebu, Cebu Philippines - jdlangoyan@up.edu.ph

Nelia S. Ereno*
University of the Philippines Cebu, Cebu Philippines – nsereno2@up.edu.ph

Abstract

The Ordinary Least Squares Regression method cannot be used in AR (1) time series model due to the existence of serial correlations among residuals. It is the Prais-Winsten and the Cochrane-Orcutt found to be the appropriate methods to use. However, it is the Prais-Winsten which gives more efficient estimates than the Cochrane-Orcutt based on the computed Root Mean Square Error. A simulation, through the statistical software R, reveals that the Prais-Winsten method is best for all econometric time series data. This paper recommends that other explanatory variables be explored in establishing the fitted time series model. Since Prais-Winsten and Cochrane-Orcutt methods can only be applied to AR (1) models, other Feasible Generalized Least Squares methods be identified and then determine as to what other time series models they can be applied. It is also recommended that their performance in terms of efficiency or accuracy or other properties of a good estimator be investigated.

Keywords: Serial correlation; Regression time series models; Feasible Generalized Least Squares.

1. INTRODUCTION

Serial correlation among residuals has been a serious problem in the analysis of time series data [1]. Various factors such as omitted variables, ignoring nonlinearities and measurement errors can produce residuals that are correlated with each other. Hence, there is a need to rectify the autocorrelation of the residuals caused by these factors.

The most widely used regression method is the Ordinary Least Squares Regression (OLSR). If the statistical tests find serial correlation though no lagged dependent variables, then OLS is consistent but inefficient and the standard errors are inappropriate [2]. Thus, the Feasible Generalized Least Squares (FGLS) is better to use. There are two procedures under the FGLS which is the concern of this paper: the Prais-Winsten (PW) and the Cochrane-Orcutt (CO) [3].

The Prais-Winsten (PW) and Cochrane-Orcutt (CO) are computational iterative procedures suit for AR(1) time series model with serially correlated residuals. The Prais-Winsten procedure is essentially identical as Cochrane–Orcutt except in the sense that the former keeps the first observation.[4]

This paper focused to: 1) determine if AR(1) model fits the collected data; 2) identify the Feasible Generalized Least Squares method that solve the most efficient estimates to AR (1) model; and, 3) simulate data that confirms the best method for all econometric time series data.

2. THEORETICAL FRAMEWORK AND METHODOLOGY

The Generalized Least Squares (GLS)

Suppose an AR(1) model has a lagged dependent variable

$$Y_t = \beta_0 X_t + e_t$$

where

$$e_t = v_t + \rho e_{t-1}.$$

Then,

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + e_{t-1}. \qquad \text{Equation (1)}$$

It is equivalent to

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho e_{t-1}. \qquad \text{Equation (2)}$$

Subtracting Equations (1) and (2),

$$(Y_{t-1} - \rho Y_{t-1}) = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + (e_t - \rho e_{t-1}).$$

Since $e_t - \rho e_{t-1} = v_t$, then

$$(Y_{t-1} - \rho Y_{t-1}) = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + v_t.$$

which can be expressed as

$$Y^* = \beta_0^* + \beta_1 X_t^* + v_t.$$

The Cochrane-Orcutt Procedure

First, estimate the model by using the OLS and get the residuals. Then, regress the residuals using the equation

$$\hat{e}_t = \hat{e}_{t-1} + v_t.$$

The results are the coefficient, $\hat{\rho}$. Next, use these coefficients to develop the models

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1} \quad \text{and} \quad X_t^* = X_t - \hat{\rho} X_{t-1}.$$

Take note that the first observation using the CO procedure is lost. Regressing *t* on, the constant β0 (1 – ρ) is obtained. Using these constant, the new residuals are generated. Regressing these new residuals, new coefficients ρ are calculated. Repeat the procedure until the convergence of the residual is achieved.

The Prais-Winsten Procedure

The Prais-Winsten procedure is essentially the same for Cochrane-Orcutt except that it keeps the first observation and it uses the following transformation:

$$\hat{Y}_0 = Y_0(\sqrt{1-\hat{\rho}^2}) \text{and } \hat{X}_0 = X_0(\sqrt{1-\hat{\rho}^2}).$$

Data Connection

The monthly cost of steel , volume of the production (in metric tons), cost of electricity, oil, air and water consumptions for 41 months were requested from the Global Steel Corporation, Iligan City, Lanao del Norte, Philippines

3. RESULTS AND DISCUSSION

Let p, d and q stand for autoregressive, differencing and for moving average, respectively. Using the Ordinary Least Squares method, the ARIMA Models, its Probability Value (p-value), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are obtained and displayed in Table 1.

Table 1. The ARIMA Models, Its Probability Value, Akaike Information Criterion and Bayesian Information Criterion

BIC	AIC	P-Value	ARIMA(p,d,q)
297.54	282.1178	0.0000	ARIMA(1,0,0)
307.7961	290.6604	0.0000	ARIMA(1,0,1)
309.3414	293.9192	0.0000	ARIMA(0,0,1)
307.0741	289.9383	0.0000	ARIMA(2,0,0)
300.4866	283.3509	0.0000	ARIMA(2,0,1)
307.5453	286.9824	0.0000	ARIMA(2,0,2)
304.7002	289.278	0.0000	ARIMA(0,0,2)
301.145	284.0093	0.0000	ARIMA(1,0,2)
305.9769	287.1276	0.0000	ARIMA(3,0,0)
309.5447	288.9818	0.0000	ARIMA(3,0,1)

All ARIMA models in Table 1 have the same p-value of 0.000. But the model with the smallest value of AIC and BIC is the ARIMA (1, 0, 0) or the AR (1) model; thus, AR (1) is the best model for the given time series data.

To verify the assumptions of stationarity of the data and the non-existence of white noise process, Figure 1 which displays the trend of the cost of steel over time was investigated.

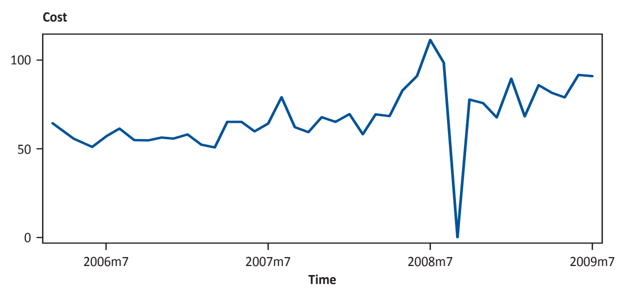


Figure 1. The Monthly Cost of Steel

It can be observed in Figure 1 that the costs of the steel were fluctuating and thus, there was no trend in the monthly cost of steel which implies that there was a stationarity of the data. This result can be verified in Table 2.

Table 2. Dickey-Fuller Test for Stationarity

P-Value	Test Statistic
0.0001	-4.732

Based on Dickey-Fuller test for stationarity with p-value of .0001, the data set is stationary at 5% level of significance.

For the existence of white noise in the data, Table 3 displays the Portmanteau white noise test results.

Table 3. Portmanteau White Noise Test

P-Value	Test Statistic
0.5284	16.9230

Since p-value is 0.5284, at 5% level of significance, there is no white noise process. The computed Durbin-Watson test statistic value of 0.9717408 (less than 1) verified that residuals were positively correlated and thus, there was really no white noise process. Therefore, the Ordinary Least Squares method cannot be applied in the given time series data. It was either the Prais-Winsten Regression or Cochrane-Orcutt

Regression that can be applied since they are the methods that can be used best in an AR (1) model with residuals that are correlated.

Table 4 displays the comparative results obtained using the Cochrane-Orcutt and Prais- Winsten Regression procedures.

Table 4. Estimates of the Independent Variables in the Two Methods

RMSE	IndependentVariables							
	Constant	Water	Air	Electricity	Oil	Volume		
7.281	-504.84	-1.47992	1.36573	0.15457	0.19344	-0.00007	Coefficient	Cochrane Orcutt Regression
	175.87	2.71146	1.28260	0.37237	0.24344	0.00013	Standard Error	
	-2.87	-0.55	1.06	4.15	0.79	-0.56	t-value	
	0.007*	0.589	0.295	0.000*	0.433	0.58	p-value	
7.220	-440.73	-0.82696	1.00638	0.15402	0.24051	-0.00007	Coefficient	Prais-Winsten Regression
	140.99	2.54440	1.17216	0.03713	0.22921	0.00013	Standard Error	
	-3.13	-0.33	0.86	4.15	1.05	-0.57	t-value	
	0.004*	0.747	0.397	0.000*	0.301	0.575	p-value	

*significant at α = 0.05

Predictors of the models considered were volume, oil, electricity, air and water. However, based on the t-values or p-values of both methods, the only independent variable that was significant was electricity. Looking at the standard errors in Table 4, the PW method yields the lowest standard errors of the estimates as compared to CO method. Also, the Root Mean Square Error (RMSE) of the PW method was smaller than the CO method. Thus, PW method is more efficient than CO method. The fitted model using PW method is given by

$$\hat{Y} = -440.7294 \ + \ 0.1540198\, X_{electricity}$$

The model explains that for every one Philippine peso increase of electricity, there is about 15 cents increase in the cost of steel. Simulation results generated using the statistical software R are shown in Table 5.

Table 5. Simulation Results of Prais-Winsten and Cochrane-Orcutt Methods

Prais-Winsten				Cochrane-Orcutt			
1000	Sum Weights	1000	N	1000	Sum Weights	1000	N
4999.681	Sum Observations	5.000	Mean	500.990	Sum Observations	5.001	Mean
0.0196	Variance	0.140	Std Deviation	0.023	Variance	0.150	Std Deviation
-0.171	Kurtosis	0.022	Skewness	-0.200	Kurtosis	0.023	Skewness
19.539	Corrected SS	25016	Uncorrected SS	22.485	Corrected SS	25032	Uncorrected SS
0.004	Std Error Mean	2.797	Coeff of Variation	0.005	Std Error Mean	3.000	Coeff of Variation

With 1000 iterations, the coefficient of variation of PW method (2.797) was found out to be smaller than that of CO method (3.000). It can be concluded that the Prais-Winsten method is best for all econometric time series data.

4. CONCLUSION AND RECOMMENDATIONS

Ordinary least squares method is not appropriate for AR (1) model for any time series data due to residuals that are correlated. In this case, the Prais-Winsten and the Cochrane-Orcutt regression time series methods can be used. Based on the obtained Root Mean Square Errors, the Prais-Winsten is more efficient than the Cochrane-Orcutt. Simulation results also reveal that the Prais-Winsten method is best for all econometric time series data.

The fitted model using the PW method given by

$$\hat{Y} = -440.7294 + 0.1540198\, X_{electricity}$$

shows that only electricity is the significant predictor for the cost of steel. There might be other significant explanatory variables to the response variable. It is recommended to identify these explanatory variables.

PW and CO methods can only be applied to AR (1) models. It is interesting to explore other time series regression methods that are applicable to other time series models and moreover investigate their performance.

REFERENCES

Stock, James H., and Mark W. (2011). Introduction to Econometrics. 3rd ed. Boston:Pearson Education/ Addison Wesley.

Greene, W. H. (2012). Econometric Analysis. 7th ed. Upper Saddle River, N. J.: Pearson Education.

Wooldridge, Jeffrey M. (2009). Introductory Econometrics: A Modern Approach. 4th ed. Mason, O.: South-Western Cengage Learning.

S.A. Van De Geer (2005). Least Squares Estimation. Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd, Chichester. ISBN-13:978-0-470-86080-9, ISBN-10:0-470-86080-4.

Kariya, T. and Kurata, H. (2004). Generalized Least Squares. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

On the Total Least-Squares Approach to a Multivariate Errors-In-Variables model

Burkhard Schaffrin[1)*], Kyle Snow[1)2)] and Xing Fang[1)3)]

[1)] Geodetic Science Program, School of Earth Sciences, Ohio State University - aschaffrin@earthlink.net

[2)] Topcon Positioning Systems, Inc., Columbus, Ohio, USA - ksnow6378137@gmail.com

[3)] School of Geodesy and Geomatics, Wuhan University, Wuhan, PR China – xfang@sgg.whu.edu.cn

Abstract

Since the work of Golub and van Loan in the early 1980's, the Total Least-Squares (TLS) approach to the Errors-In-Variables (EIV) model has become quite popular for estimating model parameters in the presence of random measurement errors for both dependent and independent data variables, including the case addressed by Schaffrin, Snow and Neitzel when the data dispersion matrices may be singular. In another case, Schaffrin and Felus applied a specific multivariate version of the EIV-Model for the adjustment of an overdetermined affine transformation. Here, we shall try to provide a general version of the TLS approach that can be used in any Multivariate EIV-Model, even when the dispersion matrices turn out to be singular. The applicability will be shown in an example from geodetic science.

Keywords: Parameter estimation, Overdetermined systems, Random measurement errors, Singular dispersion matrices.

1. INTRODUCTION

After Golub & van Loan (1980) published their now famous paper where they introduced the Total Least-Squares (TLS) approach to estimate the parameters within an Errors-In-Variables (EIV) model, the interest in this technique has been ever increasing, soon resulting in a first textbook by van Huffel & Vandewalle (1991). For a long time, however, it was thought that the TLS technique was limited to element-wise weighted data; see, e.g., Markovsky et al. (2006). Eventually, Schaffrin & Wieser (2008) would come up with a TLS algorithm that allows fairly general dispersion matrices to be used. Even more general forms of such data dispersion matrices were accommodated by Fang (2011) and Mahboub (2012); see also Schaffrin et al. (2012). They had, however, to be nonsingular.

The TLS adjustment of EIV-Models with singular covariance matrices was finally addressed by Schaffrin et al. (2014) for the standard case, resp. by Jazaeri et al. (2014) when multiple (linear and quadratic) constraints are to be considered at the same time; see also Snow (2012).

Meanwhile, Schaffrin & Felus (2008) as well as Schaffrin & Wieser (2009) had made use of a multivariate version of the EIV-Model to estimate the parameters of an affine transformation, both in the homoscedastic and the heteroscedastic case. Similar interest in certain aspects of the multivariate EIV-Model was shown by Kukush & van Huffel (2006) and by Shklyar (2010) regarding consistency, by Cheng & Kukush (2004), and Kukush & Polekha (2006) regarding goodness-of-fit tests, and by Kukush & Tsaregorodtsev (2016) regarding asymptotic normality.

Here, we shall take an algorithmic view, essentially developing three different algorithms to find the estimated parameters, the respective residuals, and the estimated variance component within the standard formulation of the Multivariate EIV-Model. Emphasis is put on the possibility of allowing some of the dispersion matrices to be singular. Thus, in *Section 1*, all the formulas necessary for the computation of the full TLS solution within a Multivariate EIV-Model will be presented, followed by the design of three different algorithms that will be tested on a numerical example from geodetic science in *Section 2*, namely a 2-D affine transformation. Some conclusions will then be drawn and an outlook given.

2. THE MULTIVARIATE EIV-MODEL: DEFINITION, FORMULAS, AND THREE ALGOMITHMS

2.1. The model definition

Let the Multivariate Errors-In-Variables Model be defined by

$$\underbrace{Y - E_Y}_{n\,x\,p} = (\underbrace{A - E_A}_{n\,x\,m}) \cdot \underset{m\,x\,p}{\Xi} \qquad p \le \text{rank }(A) = m < n, \tag{1.1a}$$

$$\begin{pmatrix} e_Y := \text{vec}\,E_Y \\ e_A := \text{vec}\,E_A \end{pmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \underset{p\,x\,p}{\Sigma_Y} & 0 \\ 0 & \underset{m\,x\,m}{\Sigma_A} \end{bmatrix} \otimes \underset{n\,x\,n}{Q} \right) \tag{1.1b}$$

Where Y and A denote (random) data matrices,

E_Y and E_A the corresponding matrices of observational noise, and

Ξ the matrix of (unknown) parameters.

In addition, 'vec' transforms a matrix into a vector by stacking one column underneath the previous one; the symbol $\otimes$ denotes the 'Kronecker-Zehfuss' product of matrices, defined by

$$G \otimes H := \left[g_{ij} . H \right] \text{ if } G = \left[g_{ij} \right] \tag{1.2}$$

Furthermore, Σ_Y and Σ_A denote matrices of variance components and Q a symmetric positive definite cofactor matrix.

For simplicity, it is assumed that, Σ_Y and Σ_A are being split into

$$\Sigma Y = \sigma_0^2\, Q_Y \quad \text{and} \quad \Sigma A = \sigma_0^2\, Q_A \tag{1.3}$$

where σ_0^2 denotes a common (unknown) variance components, while Q_Y and Q_A are given matrices of suitable size.

2.2. Formula development for TLS adjustment

Assuming invertibility of Σ_Y and Σ_A, the (weighted) TLS target function would read:

$$tr\left(\Sigma_Y^{-1} E_Y^T Q^{-1} E_Y\right) + tr\left(\Sigma_A^{-1} E_A^T Q^{-1} E_A\right) = \min_{E_y, E_A} \text{subject to model } (1.1a - b) \tag{1.4}$$

and may be translated into vectorized form by setting up the equivalent Lagrange function

$$\begin{aligned} \Phi(e_Y, e_A, \xi := \text{vec}\,\Xi, \lambda := \text{vec}\,\Lambda) &:= e_Y^T (\Sigma_Y \otimes Q)^{-1} e_Y + e_A^T (\Sigma_A \otimes Q)^{-1} e_A \\ &+ 2\lambda^T \left[\text{vec}\,Y - (I_p \otimes A)\xi - e_Y + (I_p \otimes E_A)\xi \right] = \text{stationary.} \end{aligned} \tag{1.5}$$

where Λ denotes a n x p matrix of Lagrange multipliers. The respective *Euler-Lagrange conditions* are then obtained as

$$\frac{1}{2}\frac{\partial \Phi}{\partial e_Y} = (\Sigma_Y \otimes Q)^{-1} \hat{e}_Y - \hat{\lambda} \doteq 0 \qquad \Rightarrow \hat{E}_Y = Q\hat{\Lambda}\Sigma_Y \tag{1.6a}$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial e_A} = (\Sigma_A \otimes Q)^{-1} \hat{e}_A + (\hat{\Xi} \otimes I_n) \hat{\lambda} \doteq 0 \Rightarrow \hat{E}_A = -Q\hat{\Lambda}\hat{\Xi}^T \Sigma_A \tag{1.6b}$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial \xi} = \left[I_p \otimes (\hat{E}_A - A)^T \right] \cdot \hat{\lambda} \doteq 0 \qquad \Rightarrow A^T \hat{\Lambda} = \hat{E}_A^T \hat{\Lambda} \tag{1.6c}$$

$$\frac{1}{2}\frac{\partial \Phi}{\partial \lambda} = \text{vec}\,Y - (I_p \otimes A)\hat{\xi} - \hat{e}_Y + (I_p \otimes \hat{E}_A)\hat{\xi} \doteq 0$$

$$\Rightarrow Q^{-1}(Y - A\hat{\Xi}) = Q^{-1}\hat{E}_Y - Q^{-1}\hat{E}_A \cdot \hat{\Xi} = \hat{\Lambda} \cdot (\Sigma_Y + \hat{\Xi}^T \Sigma_A \hat{\Xi}) \tag{1.6d}$$

resulting in the solution

$$\boxed{\hat{\Lambda} = Q^{-1}(Y - A\hat{\Xi})\cdot(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})^{-1}} \tag{1.7}$$

and in the relationship

$$\boxed{\hat{E}_Y\Sigma_Y^{-1}\hat{\Xi}^T + \hat{E}_A\Sigma_A^{-1} = Q\hat{\Lambda}\hat{\Xi}^T - Q\hat{\Lambda}\hat{\Xi}^T = 0} \tag{1.8}$$

Note that, for (1.7), only the matrix combination $(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})$ needs to be invertible, not the original matrices Σ_Y and Σ_A individually.

Now, by combining (1.7) with (1.6 b - c), we obtain

$$A^T\hat{\Lambda} = A^TQ^{-1}(Y - A\hat{\Xi})(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})^{-1} = -\hat{E}_A^T\hat{\Lambda} = -\Sigma_A\hat{\Xi}(\hat{\Lambda}^TQ\hat{\Lambda}) \tag{1.9a}$$

or

$$A^TQ^{-1}(Y - A\hat{\Xi}) = -\Sigma_A\hat{\Xi}\cdot(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})^{-1}\cdot(Y - A\hat{\Xi})^TQ^{-1}(Y - A\hat{\Xi}). \tag{1.9b}$$

which allows the decomposition into

$$\boxed{A^TQ^{-1}Y = (A^TQ^{-1}A)\cdot\hat{\Xi} - \Sigma_A\hat{\Xi}\cdot\hat{N}} \tag{1.10a}$$

with

$$\boxed{\hat{N} = (\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})^{-1}\cdot(Y - A\hat{\Xi})^TQ^{-1}(Y - A\hat{\Xi})} \tag{1.10b}$$

Obviously, the formulas (1.10a-b) coincide with those from Schaffrin & Wieser (2009) as well as, in simplified form, from Kukush & Tsaregorodtsev (2016). The matrix N is of special interest as it may give us the information how to estimate the variance component in (1.3) via

$$\underset{\sim}{\hat{N}} = (\hat{\Lambda}^TQ\hat{\Lambda})\cdot(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi}) = \Sigma_Y^{-1}\hat{E}_Y^TQ^{-1}\hat{E}_Y - \hat{\Lambda}\cdot\hat{E}_A\hat{\Xi} \tag{1.10c}$$

$$\Rightarrow \boxed{\mathrm{tr}\hat{N} = \mathrm{tr}(\Sigma_Y^{-1}\hat{E}_Y^TQ^{-1}\hat{E}_Y) - \mathrm{tr}(\hat{\Xi}\hat{\Lambda}\hat{E}_A) = \tilde{e}_Y^T(\Sigma_Y\otimes Q)^{-1}\tilde{e}_Y + \tilde{e}_A^T(\Sigma_A\otimes Q)^{-1}\tilde{e}_A = \hat{\sigma}_0^2\cdot(n-m)p.} \tag{1.11}$$

see also Schaffrin & Felus (2008).

2.3. Three algorithms

In the following subsection, the formulas (1.10a-b) along with (1.6a-b), (1.7) and (1.11) will be cast into *two algorithms* that differ in the way (1.10a) is solved for ˆΞ.

Algorithms Ia /Ib:

Start : $\hat{N}^{(0)} := 0 \Rightarrow \hat{\Xi}^{(0)} = (A^TQ^{-1}A)^{-1}A^TQ^{-1}Y;$

Iteration (j=1,2,3,…) :

$$\hat{N}^{(j)} = [\Sigma_Y + (\hat{\Xi}^{(j-1)})^T\Sigma_A\hat{\Xi}^{(j-1)}]^{-1}\cdot(Y - A\cdot\hat{\Xi}^{(j-1)})^TQ^{-1}(Y - A\cdot\hat{\Xi}^{(j-1)}),$$

$$(a)\ \hat{\Xi}^{(j)} = (A^TQ^{-1}A)^{-1}(A^TQ^{-1}Y + \Sigma_A\hat{\Xi}^{(j-1)}\hat{N}^{(j)}),\ or;$$

$$(b)\ \mathrm{vec}\hat{\Xi}^{(j)} = [I_p\otimes(A^TQ^{-1}A) - \hat{N}^{(j)}\otimes\Sigma_A]^{-1}\cdot\mathrm{vec}(A^TQ^{-1}Y);$$

Stop :

$$\|\hat{\Xi}^{(j+1)} - \hat{\Xi}^{(j)}\| < \delta \text{ and } \|\hat{N}^{(j+1)} - \hat{N}^{(j)}\| < \delta \text{ for a chosen } \delta.$$

After convergence:

$$\hat{\Lambda} = Q^{-1}(Y - A\hat{\Xi})\cdot(\Sigma_Y + \hat{\Xi}^T\Sigma_A\hat{\Xi})^{-1},$$

$$\hat{E}_Y = Q\hat{\Lambda}\Sigma_Y,\quad \hat{E}_A = -Q\hat{\Lambda}\hat{\Xi}^T\Sigma_A,\ \hat{\sigma}_0^2 = \mathrm{tr}\hat{N}/[(n-m)p].$$

For the *third algorithm*, formulas (1.10a-b) are modified via (1.6b) to obtain

$$\hat{\Lambda}'Q^{-1}(Y - A\hat{\Xi})(\Sigma_0 + \hat{\Xi}'\Sigma_A\hat{\Xi})^{-1} = -\Sigma_A\hat{\Xi}\cdot\hat{N}(\Sigma_0 + \hat{\Xi}'\Sigma_A\hat{\Xi})^{-1} = -\Sigma_A\hat{\Xi}(\hat{\Lambda}'Q\hat{\Lambda}) = \tilde{E}_A'\hat{\Lambda} \qquad (1.12)$$

and consequently

$$(A - \tilde{E}_A)'Q^{-1}(Y - A\hat{\Xi}) = 0, \qquad (1.13a)$$

respectively, as suggested by Fang (2011).

$$\boxed{(A - \tilde{E}_A)'Q^{-1}(A - \tilde{E}_A)\hat{\Xi} = (A - \tilde{E}_A)'Q^{-1}(Y - \tilde{E}_A\hat{\Xi})} \qquad (1.13b)$$

Algorithm II *:*

Start : $\tilde{E}_A^{(0)} = 0 \Rightarrow \hat{\Xi}^{(0)} = (A'Q^{-1}A)^{-1}A'Q^{-1}Y,$

Iteration : $(j = 1, 2, 3, \ldots):$

$$\hat{\Lambda}^{(j)} = Q^{-1}(Y - A\hat{\Xi}^{(j-1)})[\Sigma_0 + (\hat{\Xi}^{(j-1)})'\Sigma_A\hat{\Xi}^{(j-1)}]^{-1}, \quad \tilde{E}_A^{(j)} = -Q\hat{\Lambda}^{(j)}(\hat{\Xi}^{(j-1)})'\Sigma_A,$$

$$\hat{\Xi}^{(j)} = [(A - \tilde{E}_A^{(j)})'Q^{-1}(A - \tilde{E}_A^{(j)})]^{-1}(A - \tilde{E}_A^{(j)})'Q^{-1}(Y - \tilde{E}_A^{(j)}\hat{\Xi}^{(j-1)});$$

Stop $\left\|\hat{\Xi}^{(j+1)} - \hat{\Xi}^{(j)}\right\| < \delta$ and $\left\|\tilde{E}_A^{(j+1)} - \tilde{E}_A^{(j)}\right\| < \delta$ for a chosen δ.

After convergence : $\tilde{E}_A = -Q\hat{\Lambda}\Sigma_A,\ \hat{N} = (\hat{\Lambda}'Q^{-1}\hat{\Lambda})(\Sigma_0 + \hat{\Xi}'\Sigma_A\hat{\Xi}) \Rightarrow \hat{\sigma}_0^2 = \mathrm{tr}\hat{N}/[(n-m)p]$

The performance of these three algorithms will be tested on a geodetic example in *Section 2*.

3. A GEODETIC EXAMPLE. THE AFFINE TRANSFORMATION

In order to validate Algorithm *Ia*, we estimated the parameters of a 2D affine transformation. Such a transformation maps coordinate pairs ($x1$;$x2$) from a source coordinate system to coordinate pairs ($y1$;$y2$) in a target coordinate system by applying two rotations, two translations, and two multiplicative scale factors to the source coordinates. Such affine transformations are encountered in a variety of fields in science and engineering, including computer science.

Following Schaffrin & Felus (2008), an affine transformation may be modeled as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \approx \begin{bmatrix} s_1\cos\beta & -s_2\sin(\beta+\varepsilon) \\ s_1\sin\beta & s_2\cos(\beta+\varepsilon) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \xi_{31} \\ \xi_{32} \end{bmatrix} \qquad (1.14)$$

where S_1 and S_2 arescale factors; β and ε are rotation angles; and t_1 and t_2 are translation parameters. Here, the coordinates are considered to have been observed or estimated so that they contain random errors, hence the approximately-equals sign in the model. The above assignments allow four physical functions to be replaced by mathematical parameters in order to yield a linear system in the unknown mathematical parameters.

In order to uniquely determine the six parameters of the model, at least three common points must by measured in both the source and target coordinate systems. However, if one desires to minimize random errors in the total least-squares sense, a redundant set of n points, $n > 3$, ought to be measured In this case, (2.1) may be rearranged and written in the following more condensed form, which conforms to the Multivariate EIV Model presented in (1.1a) with p = 2 , m = 3, and n=5 :

$$\underset{n\times 2}{Y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} \approx \begin{bmatrix} x_{11} & x_{12} & 1 \\ x_{21} & x_{22} & 1 \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & 1 \end{bmatrix} \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \\ \xi_{31} & \xi_{32} \end{bmatrix} = \underset{n\times 3}{A}\,\underset{3\times 2}{\Xi}. \qquad (1.15)$$

The measured coordinates used in the experiment are listed in Table 1.

Table 1. Coordinates Measured in Source and Target Systems in Units of Meters

Point No	$x_{n}[m]$	$x_{12}[m]$	$x_{n}[m]$	$x_{12}[m]$
1	453.8001	137.6099	400.0040	100.0072
2	521.2865	350.7972	500.0019	299.9994
3	406.8728	433.9247	399.9925	399.9933
4	110.5545	386.9880	100.0059	400.0022
5	157.4861	90.6802	99.9956	99.9978

The cofactor matrices Q_Y and Q_A introduced in (1.3) have the following numerical values:

$$\mathbf{Q}_Y = \begin{bmatrix} 1.05 & 0.05 \\ 0.05 & 1.00 \end{bmatrix}, \quad \mathbf{Q}_A = \begin{bmatrix} 0.95 & -0.05 & 0 \\ -0.05 & 1.00 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Furthermore, the upper-triangular elements of the 5×5 symmetrical cofactor matrix 10^5. Q are are given row-wise starting at each respective diagonal component, with semi-colons separating the rows, as follows:

0.679448757180059 −0.020674926525156 −0.017523719870004 −0.045154752590708 −0.025230754063775; 0.642931803954814 −0.019705123004066 −0.014037421948659 −0.009875708917601; 0.622939970049753 −0.020512909482907 −0.045522446479986; 0.507997251772178 −0.011733918486581; 0.513926261159557.

Algorithm Ia was used to estimate the unknown affine transformation parameters using as convergence criterion $\delta = 10^{-12}$. The algorithm converged in three iterations. The estimated parameters are listed in Table 2.

Table 2. Estimated Parameters

Sclae factors 1 and 2:	$\hat{s}_1$ = 0.999965608225, $\hat{s}_2$ = 0.999972884565
Rotation angles 1 and 2-1:	$\hat{\beta}$ = -9.000372465815, $\hat{\varepsilon}$ = 0.000476690348 degrees
Translations 1 and 2:	$\hat{t}_1$ = -69.725693231691, $\hat{t}_2$ = 35.078281633661 m

The observation residuals turned out to be

$$\left[\hat{\mathbf{E}}_Y \middle| \hat{\mathbf{E}}_A\right] = \left[\begin{array}{cc|ccc} 3.5464960 & 3.7777653 & -2.2661657 & -3.9442791 & 0 \\ 1.3592426 & -0.0278862 & -1.2269760 & -0.0466320 & 0 \\ -5.0346477 & -3.4351442 & 3.6853079 & 3.6855061 & 0 \\ 3.3664970 & 2.7629091 & -2.3510697 & -2.9269394 & 0 \\ -2.9450533 & -2.8199360 & 1.9588855 & 2.9605106 & 0 \end{array}\right] \text{mm}$$

where the column of zeros corresponds to the column of ones in the matrix A of (3.2). The total sum of squared residuals (TSSR) turned out to be TSSR = tr **N** = 30.12076549108 leading to the estimated variance component σ_0^2= tr**N** / [2 . (5-3)] = 7.530191372771.

The numerical values shown above contain more significant digits than what the accuracy of the data warrants. This was done in case others want to check the results of their own algorithm implementations against ours. We also used Algorithms Ib and II with the example data set. The latter yielded precise agreement with Algorithm Ia, while the former showed some disparities in the parameter estimates ranging from 10^{-8} to 10^{-12} which seem surprisingly high. Further investigations are needed to determine the source of this disparity.

4. CONCLUSIONS

In this contribution we have developed three different algorithms to find the estimated parameters, the respective residuals, and the estimated variance component within the standard formulation of the Multivariate EIV-Model. All the formulas necessary for the computation of the full TLS solution have been presented. A numerical example from geodetic science, namely a 2-D affine transformation, was presented to validate the algorithms, with complete listing of the results such that interested readers can compare the numerical results of their own algorithms with ours. In work to follow, we plan to investigate the feasibility of estimating two separate variance components associated with $\Sigma^{\mathbf{Y}}$ and $\Sigma^{\mathbf{A}}$, or even estimating those matrices in full when only the cofactor matrix $\mathbf{Q}$ is known.

REFERENCES

Cheng, C.L. & Kukush, A. (2004). A goodness-of-fit test for a polynomial errors-in-variables model[J]. *Ukrainian Mathematical Journal*, 2004, 56(4): 641-661.

Fang, X. (2011). Weighted total least-squares solutions for applications in geodesy[M]. Ph.D dissertation, Report Nr 294, Fachrichtung Geodäsie und Geoinformatik der Leibniz-Univ., Hannover/ Germany, 2011.

Golub, G. H. & Van Loan, C. F. (1980). An analysis of the total least squares problem[J]. *SIAM Numer Anal*, 17(6): 883-893. Kukush, A. & Polekha, M. (2006). A goodness-of-fit test for a multivariate errors-in-variables model[J]. *Theory of Stochastic Processes*, 2006, 12(3–4), 63–74.

Jazaeri, S., Schaffrin, B. & Snow, K. (2014). On Weighted Total Least-Squares adjustment with multiple constraints and singular dispersion matrices[J]. *Z. fuer Vermessungswesen*, 2014, 139, 229-240.

Kukush, A. & Tsaregorodtsev, Y. (2016). Asymptotic normality of total least squares estimator in a multivariate errors-in- variables model AXB, *Modern Stochastics: Theory and Applications*, 2016, (3): 47–57.

Kukush, A. & Van Huffel, S. (2004). Consistency of element-wise weighted total least-squares estimator in a multivariate errors-in-variables model AXB[J]. *Metrika*, 2004, 59(1): 75-97.

Mahboub, V. (2012). On weighted total least-squares for geodetic transformations. *J of Geodesy*, 2012, 86(5): 359-367

Markovsky, I. Rastello, M. L. Premoli, A. Kukush, A. & Van Huffel, S. (2006). The element-wise weighted total least-squares problem. C*omputational Statistics & Data Analysis*, 2006, 50(1): 181-209.

Schaffrin, B. & Felus, Y. A. (2008). On the multivariate total least-squares approach to empirical coordinate transformations. Three algorithms. *Journal of Geodesy*, 2012, *82*(6): 373-383.

Schaffrin, B., Neitzel, F., Uzun, S. & Mahboub, V. (2012). Modifying Cadzow's algorithm to generate the optimal TLS-solution for the structured EIV-Model of a similarity transformation. *Journal of Geodetic Science*, 2012, *2*(2), 98-106.

Schaffrin, B. Snow, K. & Neitzel, F. (2014). On the Errors-In-Variables model with singular dispersion matrices. *Journal of Geodetic Science*, 2014, 4: 28-36.

Schaffrin, B. & Wieser, A. (2008). On weighted TLS adjustment for linear regression. *J of Geodesy*, 2008, 82(7): 415-421. Schaffrin, B. & Wieser, A. (2009). Empirical affine reference frame transformations by weighted multivariate TLS adjustment. In *Geodetic Reference Frames* (IAG –Symp No 134, pp. 213-218). Springer Berlin/Heidelberg, 2009.

Shklyar, S. (2011). Conditions for the consistency of the total least-squares estimator in an errors-in-variables linear regression model. *Theory of Probability and Mathematical Statistics*, 2010, *83*: 175-190.

Snow, K. (2012). Topics in Total Least-Squares Adjustment within the Errors-In-Variables Model: Singular Cofactor Matrices and Prior Information. PhD dissertation, Report No 502, OSU, 2012

Van Huffel, S. & Vandewalle, J. (1991). *The total least squares problem: Computational aspects and analysis*. Society for Industrial and Applied Mathematics, Philadelphia/PA, 1991.

CPS16:
Health & Social Statistics (3)

The Quadratic Effect of Life Expectancy on Economic Growth

Nazirul Hazim A Khalim*
Monash University Malaysia, Kuala Lumpur, Malaysia – nazirul.hazim@monash.edu

Abstract

Rapid increases in life expectancy are causing a significant shift in the global age structure which has raised concerns about a future slowing of economic growth as rising old-age dependency ratios translate into growing tax burdens and as political pressure may lead to a crowding out of public investment spending in favour of social spending on the elderly, with negative effects for capital accumulation and productivity growth. The evidences linking life expectancy and economic growth have been contradictory and heterogeneous. While most studies assert a linear relationship, some have proposed it to be non-linear. Using panel data from 164 countries over 24 years, this study attempts to consolidate contradictory prior evidence by proposing a non-linear quadratic effect of life expectancy on economic growth. It found statistically significant quadratic effect with a maximum turning point at 61 years of life expectancy, beyond which a rise in life expectancy impedes economic growth. This turning point was found to be consistent with the retirement ages of the countries studied.

Keywords: Life expectancy; Economic growth; Non-linear; Panel data.

1. INTRODUCTION

Rapid increases in life expectancy are causing a significant shift in the global age structure. Recent estimates suggest that the number of people over the age of 60 is projected to reach 1 billion by 2020 and almost 2 billion by 2050, representing 22% of the world's population. Moreover, the proportion of individuals aged 80 or over is projected to rise from 1% to 4% of the global population between today and 2050 (UN, 2009). This development has raised concerns about a future slowing of economic growth as rising old-age dependency ratios translate into growing tax burdens and as political pressure may lead to a crowding out of public investment spending in favour of social spending on the elderly, with negative effects for capital accumulation and productivity growth (Kunze, 2014).

The evidences linking life expectancy and economic growth have been contradictory and heterogeneous. While most studies assert a linear relationship, some have proposed it to be non-linear. The latter were motivated by the contradictory prior results and therefore attempted to probe further into the effect on life expectancy on economic growth across countries, over time, differing development levels, age cohorts, and health vulnerability factors (Chay & Greenstone, 1999; Cutler, Huang, & Lleras-Muney, 2016; Dadgar & Norström, 2016; Davis, Laden, Hart, Garshick, & Smith, 2010; Karanikolos, Heino, McKee, Stuckler, & Legido-Quigley, 2016; Noland & Zhou, 2017; Ruhm, 2000). By using panel data from 164 countries over 24 years, this study attempts to consolidate contradictory prior evidence by proposing a non-linear quadratic effect of life expectancy on economic growth.

2. REVIEW OF RELATED STUDIES

A substantial body of empirical literature has investigated the relation between life expectancy and economic growth (see e.g. Acemoglu and Johnson, 2007 and Lorentzen et al., 2008). However, empirical evidence remains inconclusive and the findings of recent studies suggest that there may be non-linearities in the pattern of life expectancy and growth.

Life expectancy affects growth through various channels. First, it raises the savings rate and thereby increases the rate of physical capital accumulation. Second, it lowers investments into children's education as old-age consumption becomes relatively more important. Third, it reduces the amount of bequests parents devote to their children which in turn slows down physical capital accumulation. Fourth, life expectancy affects the size of public education expenditures through political adjustments in the income tax rate. The latter effect turns out to be growth enhancing as tax rates are increasing functions of life expectancy and the politically chosen levels are always below their corresponding growth maximizing size. The balance of these effects determines the relation between life expectancy and economic growth.

However, empirical evidence on the life expectancy–growth relationship turns out to be mixed. Barro and Sala-I-Martin, 1995 and Lorentzen et al., 2008 find in growth regressions that an increase in longevity is associated with higher growth rates. Kalemli-Ozcan, Ryder, & Weil (2000) probed into the role of increased life expectancy in raising human capital investment during the process of economic growth and found that longevity produces economically significant increases in schooling and consumption which in turn spurs economic growth.

On the other hand, Acemoglu & Johnson (2007) found no evidence of a positive growth effect. They found that life expectancy has a small effect on total GDP both initially and over a 40-year horizon. Consequently, they also found no evidence that large exogenous increase in life expectancy leads to a significant increase in per capita economic growth.

The findings by Kelley & Schmidt, 1995 and An & Jeon (2006) point to the existence of a non-linear relationship between aging and economic growth. They suggest increasing life expectancy may be good for growth when starting from a low level and bad for growth when starting from a high level. A non-linear relationship is further supported by Kunze (2014) which observed that growth rates increase substantially as the initial life expectancy rises from below 60–69. However, when the initial life expectancy rises to 70 and over growth rates decline even though their levels are still higher than those in countries with low life expectancy. De la Croix & Licandro (1999) found that the effect of life expectancy on growth is positive for economies with a relatively low life expectancy, but could be negative in more advanced economies. The positive effect of a longer life on growth could be offset by an increase in the average age of the workers. They explain this by positing that individuals have to choose the length of time devoted to schooling before starting to work and this depends positively on life expectancy.

Motivation

The demographic trends over the past 50 years – which can be summarized as showing a steady decline in fertility rates and increase in life expectancy – is thought to have a powerful impact on the rate of economic growth (An & Jeon, 2006). However, most empirical studies on the economic consequence of demographic change including Cutler et al. (1990), Bloom et al. (2000), Jones (2002), etc. find little cross-country evidence. Instead, there is a continuing debate over the demographic effects on economic growth. As discussed in Bloom et al. (2003), the debate involves three positions, such that demographic change (or population growth) restricts, promotes or is independent of economic growth, representing the "Pessimistic," "Optimistic," and "Neutralist" theories respectively.

This study is motivated by the seemingly contradictory empirical evidence in prior studies of the effect of life expectancy on growth and thus attempts a consolidation. This study proposes a non-linear quadratic (polynomial of order 2) effect of life expectancy on economic growth.

3. METHODOLOGY

Data

All data are sourced from the World Development Indicators Database (World Bank Group, 2016). The period under study is 1991 – 2014 with annual frequency (24 years). 164 countries are studied. The total panel (unbalanced) observations is 3643.

The following variables has been included in the final model. Definitions are provided by the World Development Indicators Database (World Bank Group, 2016):

Table 1. Variable Included in the Final Model

Dependent Variable	
GROWTH	*GDP growth (annual %)*: Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. Source: World Bank national accounts data, and OECD National Accounts data files.
GROWTHPC	*GDP per capita growth (annual %)*: Annual percentage growth rate of GDP per capita based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. GDP per capita is gross domestic product divided by midyear population. Source: World Bank national accounts data, and OECD National Accounts data files.

Independent Variables	
CPI	*Inflation, consumer prices (CPI, annual %):* Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. Source: International Monetary Fund, International Financial Statistics and data files.
FDI	*FDI net inflows (% of GDP), Foreign Direct Investment:* Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors, and is divided by GDP.
POP	*Total of Population:* Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates. Source: (1) United Nations Population Division. World Population Prospects, (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Population and Vital Statistics Report (various years), (5) U.S. Census Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Programme.
POP_GROWTH	*Population growth (annual %):* Annual population growth rate for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage. Source: (1) United Nations Population Division. World Population Prospects, (2) Census reports and other statistical publications from national statistical offices, (3) Eurostat: Demographic Statistics, (4) United Nations
UNEMP	*Unemployment, total (% of total labor force) (modeled ILO estimate):* Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Source: International Labour Organization, Key Indicators of the Labour Market database.
LIFE	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
LIFE^2	Squared of life expectancy at birth, total (years)
Y1998	Dummy variable, = 1 if the year is 1998, Asian Financial Crisis
Y2009	Dummy variable, = 1 if the year is 2009, The Subprime Crisis

Model 1

The base model (Model 1) is an unbalanced panel data with cross-section random effects of 16 countries spanning 23 years (355 observations). It can be written as follows:

Model 1 is a small sample unbalanced panel data with cross-section random effects of 16 countries spanning 23 years, 1991-2013 (355 observations).

$$growth_{it} = \beta_0 + \beta_1 cpi_{it} + \beta_2 fdi_{it} + \beta_3 unemp_{it} + \beta_4 y1998 + \beta_5 y2009 + \beta_6 pop_{it} + \beta_7 life_{it} + \beta_8 life_{it}^2 + \varepsilon_{it}$$

Model 2, 3, 4 and 5 are modifications to Model 1 with an extended sample unbalanced panel data with cross-section fixed effects of 164 countries spanning 24 years, 1991-2014 (3643). FDI was dropped due to lack of availability and Y1998 was dropped due to insignificance, most likely because the Asian Financial Crisis only affects mostly Asia Pacific economies. Model 3, 4 and 5 test the base model's robustness by switching in GDP Growth Per Capita and Life Expectancy for male and female.

Development of the Base Model

The Breusch-Pagan Lagrangian Multiplier (LM) Test for testing Pooled OLS Model against Random Effects Model (RE) rejected the null hypothesis of no random effects for cross-section, time, and both. The test's results are presented below.

Table 2. The Test Results of the Breusch-Pagan Lagrangian Multiplier Test for Pooled OLS Model Against Random Effects Model (RE)

Lagrange Multiplier Tests for Random Effects
Null hypotheses: No effects
Alternative hypotheses: Two-sided effects

	Cross-section	Test Hypothesis Time	Both
Breusch-Pagan	81.81984 (0.0000)	15.04652 (0.0001)	96.86637 (0.0000)

Rejection of the Breusch-Pagan Lagrangian Multiplier (LM) Test hypotheses therefore suggest that the regression should incorporate random effects for both cross-section and time. However, the final model only incorporates cross-section random effects. The period effects are captured by two dummy variables for years 1998 and 2009 which respectively captures the low (even negative) growth during the Asian Financial Crisis and the Subprime Crisis. The years are chosen based on the observation of the cross-section mean of GDP Growth (annual %) over time which shows substantial drops in years 1998 (growth at -0.6%) and 2009 (growth at 0.2%).

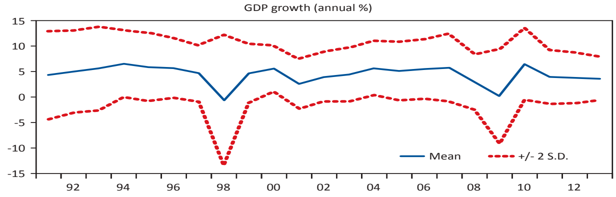


Figure 1. The Cross-section Mean of GDP Growth during the Asian Financial Crisis and the Subprime Crisis

The Hausman Test for testing Random Effects (RE) against Fixed Effect (FE) does not reject the null hypothesis of no fixed effects at 10% significance level. The test results presented below. Therefore, Random Effects (RE) is appropriate.

Table 3. The Test Result of the Hausman Test for Random Effect (RE) Against Fixed Effect (FE)

Correlated Random Effects-Hausman Test
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	11.380026	8	0.1811

4. RESULTS AND DISCUSSION

Model 1 (Base)

Table 4. The Cross-section Random Effect Regression Results

Dependent Variable: GROWTH
Method: Panel EGLS (Cross-section random effects)
Sample (adjusted): 1991 2013
Periods included: 23
Cross-sections included: 16
Total panel (unbalanced) observations: 355S
wamy and Arora estimator of component variances
White diagonal standarderrors & covariance (d.f. corrected)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
CPI	-0.182450	0.056121	-3.251003	0.0013
FDI	0.197289	0.037132	5.313176	0.0000
UNEMP	-0.292981	0.070713	-4.143251	0.0000
Y1998	-4.847554	1.336226	-3.627795	0.0003
Y2009	-4.113368	0.702688	-5.853760	0.0000
POP	2.76E-09	8.54E-10	3.232481	0.0013
LIFE	2.667280	0.753563	3.539560	0.0005
LIFE^2	-0.020216	0.005103	-3.961899	0.0001
C	-80.17673	27.76047	-2.888161	0.0041
		Effects Specification	S.D.	Rho
Cross-section random			1.001988	0.1571
Idiosyncratic random			2.320652	0.8429

Weighted Statistics

R-squared	0.431967	Mean dependent var	1.912813
Adjusted R-squared	0.418833	S.D. dependent var	3.064236
S.E. of regression	2.334997	Sum squared resid	1886.464
F-statistic	32.88993	Durbin-Watson stat	1.446903
Prob(F-statistic)	0.000000		

The panel cross-section random effects regression results show statistically significant p-values (at 5% level) for all independent variables. The model explains 41.9% of the variation in GDP growth based on the weighted *Adjusted* R^2. The F-statistics suggests significant overall power of the model. White diagonal (cross-section and time) robust standard errors & covariance are used to correct for heteroskedasticity and serial correlation, which are suggested to be present by the Durbin-Watson statistics being 1.44 which is not close to 2. The signs for the control variables are reasonable and theoretically sensible. Positive coefficients are found for FDI (foreign direct investment) and POP (total population) whereas negative coefficients are found for CPI (inflation), UNEMP (unemployment), Y1998 and Y2009.

The variable of interest, life expectancy is specified as a polynomial of degree 2 and both the linear and polynomial (quadratic) components (*Life* and $Life^2$) are statistically significant at 1%. The linear component has a positive sign whereas the quadratic component is negative. These signs suggest a quadratic relationship with a maximum turning point. From calculation, the turning occurs at 66 years of life expectancy. In other words, an increase in life expectancy promotes a rise in economic growth until it reaches 66 years old where the effect reverses whereby longer life expectancy becomes detrimental to economic growth. To provide perspective, the average retirement age of the countries studied is 62 for male and 61 for female, ranging between 58-66 for male and 55-66 for female. The average for OECD is 65 for male and 63.5 for female.

The consistency of the turning point age of 66 with the general retirement age across countries suggests a negative effect of an increase in ageing (specifically retired) population to economic growth past 66 years old. Controlling for population growth, an increase in life expectancy would likely increase ageing population (Zhang & Zhang, 2005). Population ageing tend to lower both labour force participation and savings rates, thereby decelerating future economic growth (Bloom, Canning, & Fink, 2010). On the other hand there also robust and consistent contrary evidence suggesting a positive effect of life expectancy on growth (Zhang & Zhang, 2005). This empirical relationship is supported by inconclusive evidence whereby rising life expectancy positively affects savings rates (Zhang & Zhang, 2005). Intuitively, when expecting a higher chance of surviving to retirement, rational individuals save more for old-age consumption.

Some studies have suggested differing effects of ageing population on economic growth between developed and developing countries. Bloom et al. (2010) suggest that OECD countries are likely to see modest declines in the rate of economic growth whereas ageing population in most non-OECD countries will not significantly impede their economic growth. The latter's experience (at least in the short run) is caused by the skewing of adults towards the older ages being offset by shrinking share of (non-working) young people due to declining fertility rates (Zhang & Zhang, 2005). The shrinking share of young people causes the labour-force-to-population ratios to rise and in turn promotes growth per capita (Bloom et al., 2010). Based on the findings of Bloom et al. (2010), this study attempted to identify if the effect of life expectancy on growth differs for OECD and non-OECD countries with the use of an indicator variable (OECD=1) but found no statistically significant evidence of such distinction.

Summary of Results for Models 2, 3, 4 and 5. Sample of 164 countries over 24 years (1991-2014).

Table 5. Summary of result for Model 2,3,4 and 5

Dep. Var: Model	GROWTHPC 2	GROWTH 3	GROWTHPC 4	GROWTHPC 5
CPI	-0.000517 [-2.1848]*	-0.000526 [-3.2240]**	-0.000518 [-3.2377]**	-0.000517 [-3.2378]**
POP	0.182333 [4.0225]**			0.165163 [3.4030]**
POP_GROWTH	-0.509149 [-3.2344]**	0.536561 [3.8169]**	-0.518137 [-3.7416]**	-0.485338 [-3.5463]**
UNEMP	-0.178754 [-3.2907]**	-0.204740 [-4.5313]**	-0.202869 [-4.5739]**	-0.177225 [-3.9704]**
Y2009	-4.362529 [-10.6131]**	-4.256496 [-9.2600]**	-4.233025 [-9.3955]**	-4.376476 [-9.7009]**
LIFE	1.319759 [6.3021]**	1.103048 [5.5300]**		
LIFE^2	-0.010573 [-6.4225]**	-0.008167 [-4.9098]**		
C	-45.992152 [-5.8718]**	-31.200826 [-5.2564]**	-30.409847 [-5.1433]**	-42.769549 [-6.3651]**

Dep. Var: Model	GROWTHPC 2	GROWTH 3	GROWTHPC 4	GROWTHPC 5
LIFE_FEMALE			1.049067 [5.3447]**	
LIFE_FEMALE^2			-0.007533 [-4.6978]**	
LIFE_MALE				1.259373 [6.3683]**
LIFE_MALE^2				-0.010250 [-5.9266]**
Observations:	3643	3643	3643	3643
R-squared:	0.1864	0.2022	0.1828	0.1860
F-statistic:	4.6786	5.2098	4.5970	4.6666

T-statistics in [brackets]. * indicates significance at 5%, ** at 1%.

Modifications to Model 1, as presented by Models 2-5 maintain the statistical significance and coefficient signs of the variables. Life Expectancy for Male and Female show significance as well. One interesting difference is that the larger sample size show that the turning point of life expectancy to be 61 years old, lower than the 66 years old found in Model 1. This results put the turning point closer to the average retirement age around the world.

This study's finding of a statistically significant quadratic relationship with a maximum turning point between life expectancy and growth consolidates and harmonizes the seemingly contradictory findings within growth and longevity literature. In other words, this study suggests that prior studies are correct but only up to a certain point. Such diminishing effect has been suggested by some studies as well (Zhang & Zhang, 2005).

Interestingly, when only a linear component of life expectancy was included in the model, the coefficient is statistically insignificant. Hence lending further evidence for a non-linear relationship. Some studies has also suggested a non-linear relationship between life expectancy and economic growth (Kunze, 2014). This study builds on those suggestions and provide longitudinal empirical evidence across countries over many years. Further, this study's model takes into account other factors influencing growth and found the signs to be consistent with prevailing theory and the coefficients statistically significant.

The closest equivalent to these findings from prior evidence was presented by An & Jeon (2006). They found that demographic changes appear to first increase and then decrease economic growth. They named this relationship the Demographic U Hypothesis (Curve) — an inverted U-shape relationship between demographic change and economic growth

5. CONCLUSIONS

Based on these empirical findings, three subjects can be addressed for the further study. First, these findings will enable us to make a more accurate projection of the effect of aging on growth for the countries which recently began to experience the effects of an aging population. Second, while the impact of demographic change itself plays a significant role in economic performance, also very important is how quickly the change occurs. Korea is now confronting the fastest population aging in the world, while Japan has already suffered from a long-term economic recession due to a fast aging rate. Hence, as demographic change occurs more rapidly, we can expect the bigger impact on the economic growth. In addition to studying the demographic impact on economic growth, it is also meaningful to address the speed of the demographic changes which vary across countries depending on the timing of baby booms and busts, and the extent to which fertility rate and mortality rate decline in a given period. Third, we need a theoretical approach to the demographic U hypothesis, which might support the empirical finding of the demographic U curve. The theoretical investigation should also be applied to the channels of the demographic impact on the economic growth.

Limitations

Due to lack of available data, this study failed to cover any aspect of investment and government expenditure and therefore its effect on economic growth. Given the composition of aggregate demand, these components might prove crucial. However, investment may have been partly captured by the variation in unemployment, a variable which is included in the model. Higher investment ought to correspond with lower unemployment.

This study also failed to capture the possible evolution of the U-shape over a longer period of time. While the findings show an inverted U-shape with a maximum turning point, a non-inverted U-shape may have materialized over the longer run and within different subsets of periods, in addition to cross country variations. A casual observation of the individual country's life expectancy vs growth scatter plots seems to point towards this pattern.

REFERENCES

Acemoglu, D., & Johnson, S. (2007). Disease and Development: The Effect of Life Expectancy on Economic Growth. *Journal of Political Economy*, *115*, 925–985.

An, C.-B., & Jeon, S.-H. (2006). Demographic change and economic growth: An inverted-U shape relationship. *Economics Letters*, *92*, 447–454.

Bloom, D. E., Canning, D., & Fink, G. (2010). Implications of population ageing for economic growth. *Oxford Review of Economic Policy*, *26*, 583–612.

Chay, K. Y., & Greenstone, M. (1999). *The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w7442

Cutler, D. M., Huang, W., & Lleras-Muney, A. (2016). *Economic Conditions and Mortality: Evidence from 200 Years of Data* (Working Paper No. 22690). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w22690

Dadgar, I., & Norström, T. (2016). Short-term and long-term effects of GDP on traffic deaths in 18 OECD countries, 1960–2011. *Journal of Epidemiology and Community Health*, jech-2015-207138.

Davis, M. E., Laden, F., Hart, J. E., Garshick, E., & Smith, T. J. (2010). Economic Activity and Trends in Ambient Air Pollution. *Environmental Health Perspectives*, *118*, 614–619.

de la Croix, D., & Licandro, O. (1999). Life expectancy and endogenous growth. *Economics Letters*, *65*, 255–263.

Kalemli-Ozcan, S., Ryder, H. E., & Weil, D. N. (2000). Mortality decline, human capital investment, and economic growth. *Journal of Development Economics*, *62*, 1–23.

Karanikolos, M., Heino, P., McKee, M., Stuckler, D., & Legido-Quigley, H. (2016). Effects of the Global Financial Crisis on Health in High-Income Oecd Countries A Narrative Review. *International Journal of Health Services*, *46*, 208–240.

Kunze, L. (2014). Life expectancy and economic growth. *Journal of Macroeconomics*, *39*, *Part A*, 54–65.

Noland, R. B., & Zhou, Y. (2017). Has the great recession and its aftermath reduced traffic fatalities? *Accident Analysis & Prevention*, *98*, 130–138.

Ruhm, C. J. (2000). Are Recessions Good for Your Health? *The Quarterly Journal of Economics*, *115*, 617–650.

World Bank Group. (2016). The World Bank DataBank [Database]. Retrieved November 21, 2016, from http://databank.worldbank.org/data/home.aspx

Zhang, J., & Zhang, J. (2005). The Effect of Life Expectancy on Fertility, Saving, Schooling and Economic Growth: Theory and Evidence*. *Scandinavian Journal of Economics*, *107*, 45–66.

On Modeling Transport Accident Deaths in Thailand based on Poisson distributions

Nuntaporn Klinjun
Office of Disease Prevention and Control, Songkhla, Thailand. (e-mail: pan.klinjun@gmail.com)
Apiradee Lim
Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani, Thailand. (e-mail: api_45@hotmai.com)
Wandee Wanishsakpong
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand. (e-mail: w.wandee43@gmail.com)
Khairil Anwar Notodiputro
Department of Statistics, Faculty of Mathematics and Science, Bogor Agricultural University, Bogor, Indonesia. (e-mail: khairilnotodiputro@gmail.com)

Abstract

This study aimed to investigate patterns of transport accident deaths in Thailand from year 2000 to 2009. The death certificate data were obtained from the Bureau of Policy and Strategy, Ministry of Public Health in Thailand. A generalized linear model (GLM) based on the Poisson distribution was used for identifying the patterns of transport accident deaths with respect to location, gender-age group and year. The transport accident deaths were found to be the lowest in Bangkok followed by the three south-most provinces and the most of the North-east provinces. The overall transport accident deaths during the past ten years have decreased by 30%. This was happened especially among Thai males, in which the peaked was in age group 20-29 years. The results from this study have provided useful information for the relevant agencies to continuously reduce deaths caused by transport accidents.

Keywords: Generalized Linear Model, Poisson distribution, Transport accident deaths

1. INTRODUCTION

Injuries and deaths from transport accidents are a major problem in every country around the world. WHO's Global Burden of Disease Project in 2004 reported that over 1.27 million people die each year and between 20 and 50 million are injured in road traffic accidents globally. The number of these deaths was predicted to increase from the 2004 level by 2.4 million by year 2030 (World Health Organization, 2009a). A survey of the status of road safety from 178 countries in 2008 reported that low-income and middle-income countries have road traffic fatality rates of 21.5 and 19.5 per 100,000 population respectively, compared with high-income countries 10.3 per 100,000 population (World Health Organization, 2009a). Ninety-one percent of road accident mortalities occurred in low-income and middle-income countries, especially in the South-East Asian region, from a rapidly increasing number of road accidents (World Health Organization, 2009a). In 2004, 306,000 people were killed on the roads of all countries in this region and it was the tenth leading cause of death, responsible for 2% of all causes of mortality and 18% of injury-related mortality (World Health Organization, 2009b).

Transport accidents have a direct impact on human physical and mental health, quality of life and property. Moreover, they have substantial impact on both the household income and the national economy, including loss of productivity, expenses of prolonged medical care, and funeral costs or rehabilitation (Nantulya and Reich, 2003). In addition, transport accidents cause a burden to health care systems globally, especially in low-income and middle-income countries. Many deaths and injuries from transport accidents are preventable (Anjuman et al, 2007; World Health Organization, 2009a).

A survey in 10 of the 11 countries of the WHO South-East Asia region in 2007 reported that Thailand had the highest mortality rate of 25.4 per 100,000 population (World Health Organization, 2009b), and over 80% of those killed on roads in Thailand in 2007 were vulnerable road users (pedestrians, motorcyclists and cyclists) (World Health Organization, 2009a). A study of traffic accident costs in Thailand in 2004 estimated that the economic loss due to traffic accidents was 204,050 million Baht (approximately 5,101 million US$) or between 3.9 and 4.7 million Baht per fatality (Department of Highways and Faculty of Engineering, Prince of Songkla University, 2007). In 2007-2009, the incidence of transport accident deaths in Thailand showed that transit travelers in the northern region are at greater risk of death than travelers in other regions (Thairoads Foundation and Thailand Accident Research Center, 2554). In 2004, road accidents were the second highest cause of premature mortality and disability-adjusted life years (DALYs), especially for males aged 15-29 years (Bundhamcharoen et al, 2011).

The Thai government has emphasized road safety and has tried to reduce transport accident mortalities since 2004 (Tanaboriboon and Satiennam, 2005). The current target is to reduce the road accident death rate from 14.15 per 100,000 population in 2012 to be less than 10 per 100,000 population in 2020 (Thairoads Foundation and Thailand Accident Research Center, 2011). The aim of this study was to investigate the mortality patterns from transport accidents by gender-age group, location, and year in Thailand from year 2000 to 2009.

2. MATERIALS AND METHODS

2.1. Data sources and management

This study used data from death certificates during years 2000 to 2009, available from the Thai Ministry of Public Health. These data are classified by gender-age group, year, place of death and cause of death using the International Classification of Diseases, Tenth Revision (ICD-10) code for principal diagnosis.

Death certificate data are a main source for death data from injuries (Suriyawongpaisal, 2003). However the reporting system has a problem with incompleteness (Porapakkham et al, 2010; Bundhamcharoen et al, 2011). Approximately 15% of hospital deaths and more than 50% of deaths that occurred outside hospitals were reported in year 2005 with ill-defined causes (Pattaraarchachai et al, 2010; Polprasert et al, 2010). In 2008, the age-standardized all-cause of mortality rate in Thailand was 934.7 deaths per 100,000 population which was ranked 78th among 193 countries. Correspondingly the age-standardized death rate from road traffic accidents in Thailand was 42.9 per 100,000 population (World Health Organization, 2011). Thai population data separated by gender, age group, year, and district were obtained from the 2000 population and housing census, National Statistical Office (2002) and were used as denominators for calculating mortality rates and mean of transport accidents in Thailand.

We grouped contiguous ages into 10-year age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 and 80 years and older. A total of 926 contiguous districts were reduced to 235 "super-districts" with approximate average populations of 200,000 persons per super-district (Odton, 2010). Death Certificates use WHO codes and include transport accidents classified to ICD-10 codes V01-V99. For Thailand, deaths recorded as ICD-10 code R00-R99 were 37.8% of all deaths for the period 2000-2009 (37.9% in 2009), ranging from 10.8% for males aged 10-19 years to 75.5% for females aged 80 years and older. However, some data were not specific about cause of death. Code within R00-R99 from chapter XVIII of ICD-10 include "senility" (R54) (Porapakkham et al, 2010) but the high usage of R00-R99 code indicates the possibility of imprecision.

2.2. Statistical Modeling

We fitted a statistical model to the transport accident death rates data in Thailand for all age groups for years 2000 to 2009. The model is a simple generalized linear model based on the Poisson distribution. This gives the model

$$\log(\lambda_{ijt}/P_{ijt}) = \mu + \alpha_i + \beta_j + \gamma_t.$$

For this model λ_{ijt} is the mean of the Poisson distribution giving the number of transport accident deaths for a specified super-district (i) (i=1,2,3,…,235), gender-age group (j) (j=100, 110,…, 180, 200,…, 280) and year (t) (t=2000, 2001, 2002,…, 2009). P_{ijt} is the corresponding population at risk in 100,000s and the terms α_i, β_j and γ_t represent super-district, gender-age group and year, respectively. μ is a constant encapsulating the overall incidence. The model thus has 42,300 cells (235*18*10) corresponding to 235 super-districts, 18 gender-age group combinations and a 10 year period. Using sum contrasts

(Venables and Ripley, 2002; Tongkumchum and McNeil, 2009), we obtain adjusted mortality rates and corresponding confidence intervals for comparing them. Transport accident death rates in each super-district were estimated using these confidence intervals to separate super-districts into three groups, and with this information we used corresponding colors: red, green and light blue to show above the mean, crossing the mean and below the mean categories, respectively. Statistical modeling and graphical presentation were carried out by using R statistical software (R Development Core Team, 2011).

3. RESULTS

Figure 1 shows a bar chart of super-districts. Parameters with confidence intervals were divided into three groups according to whether the confidence interval is entirely above the mean, crossing the mean, or below the mean of transport accident mortality rate. Each bar of the graph represents super-districts from code 1 to 235. This graph shows that super-districts around but not in Bangkok, the Northern region, and most of the Southern region had higher transport accident mortality than the average transport accident rate. In the past 10 years, the highest transport accident mortality occurred in the central region around Bangkok such as Rayong, Phrachin Buri, Chonburi, followed by the Northern region, such as Chiang Mai, Kamphaeng Phet, Lamphun, and the Southern region such as Surat Thani, Chumphon, and Krabi provinces, etc.

Super-districts in Bangkok had the lowest transport accident mortality followed by the three southernmost provinces in Thailand (Pattani, Yala, and Narathiwat) and the North-eastern region except in Nakhon Ratchasima, Khon Kaen, Chaiyaphum, Ubon Ratchathani, and Si Sa Ket provinces.

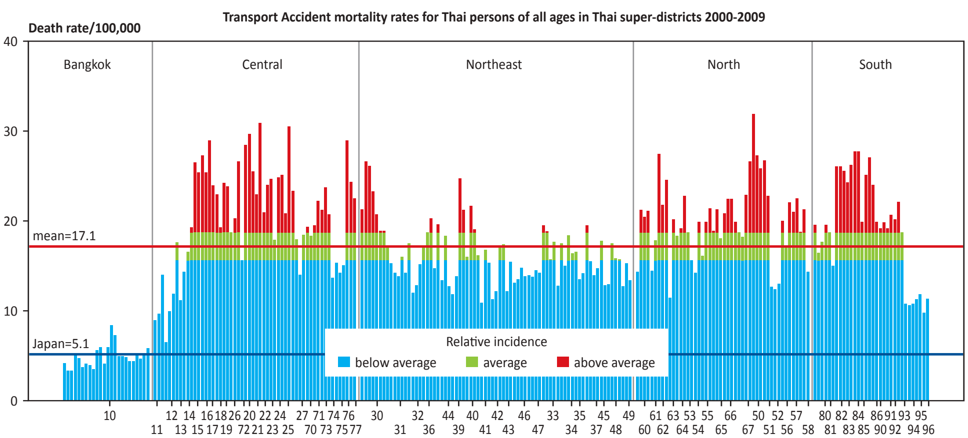


Figure 1. Bar Chart of Adjusted Death Rates

Figure 2 shows that the transport accident mortality rates for males older than 10 years were significantly higher than for females and for the mean of overall transport accident mortality rates. The transport accident mortality rates were at a peak for males aged 20-29 years. The transport accident mortality rates decreased by 30% over the 10-year period.

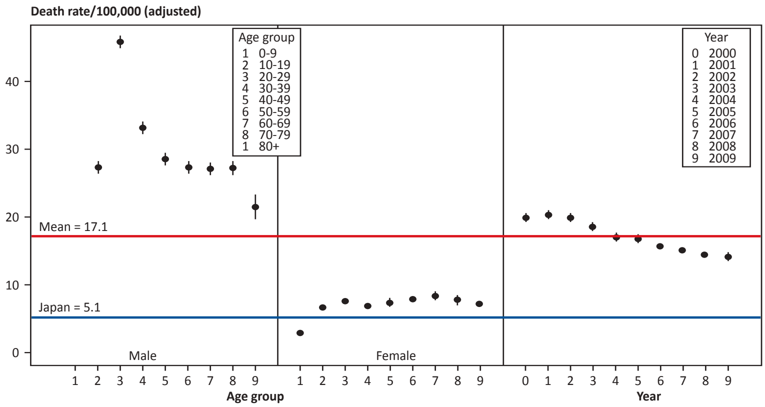


Figure 2. Transport Mortalities in Thai People, 2000-2009

4. DISCUSSION AND CONCLUSION

Death certificate data from years 2000 to 2009 showed that the number of Thai people in all age groups who were killed from transport accidents was more than 110,000 persons, while the Royal Thai Police data for the same period shows that the number of transport accident deaths in all ages was 124,855 persons (Thairoads Foundation and Thailand Accident Research Center, 2011). This difference can be explained in that the numbers of deaths from transport accidents in different database have different reporting systems with differences in the coverage and quality of data (Ditsuwan et al, 2011; Suriyawongpaisal and Aekplakorm, 2003; World Health Organization, 2009a) and the purpose of data collection (Suriyawongpaisal and Aekplakorm, 2003). However these two data sources show a consistent decreasing trend of transport accident fatality rates since 2005.

Transport accident mortality rates in Bangkok were the lowest in Thailand, followed by the three southernmost provinces. The highest transport accident mortality rate was in the Central region around Bangkok, Chiang Mai, and the other Southern provinces (apart from Nakorn Si Thammarat). There is a considerably lower accident mortality rate in Bangkok. This result is consistent with the report by The Thairoads Foundation and Thailand Accident Research Center in 2011 which reported that Bangkok is at lower risk area of death from transport accidents than other provinces. The lowest proportion of the motorcycle accident deaths may be due to higher percent of wearing helmets and fastening seat belts by both drivers and passengers than in the other provinces in Thailand. This is supported by a previous report by the National Highway Traffic Safety Administration (NHTSA) in 2008, which estimated that motorcycle helmets reduce the risk of a crash fatality by 37% for motorcycle riders and 41% for motorcycle passengers. Helmet use reduced the mortality and morbidity from head injuries (Chiu et al, 2000; Pitaktong et al, 2004).

The transport accident mortality rates for males were higher than for females by 4.3 times, consistent with the results from Thairoads Foundation and Thailand Accident Research Center (2011) which reported that the transport accident mortality rates in males were higher than females by 4 times. A similar result was reported in South Africa in 2001-2006 where the death rates from road traffic accident in males were higher than females by 2.5 times (Statistics South Africa, 2009). Death rates from road traffic accident in European countries also found that males were higher than females (OECD, 2010). The overall mortality rate decreased by 30% over the last decade. A similar result was reported by the Thairoads Foundation and Thailand Accident Research Center (2011) and Koedklai et al (2008) with a decreasing fatality rate since 2004 when the government launched the 5-E strategy: engineering, education, enforcement, emergency medical service, and evaluation (National Health Foundation, 2006; Tanaboriboon and Satiennam, 2005). This suggests that preventative measures are effective. Even though the transport accident fatality rate declined, it is still higher than the target road accident death rate of 14.15 per 100,000 population in 2012 and less than 10 per 100,000 population in 2020 (Thairoads Foundation and Thailand Accident Research Center, 2011).

The rising number of transport accident mortalities may relate to the increasing number of registered vehicles. However this information was not considered in this study. Even though the quality of death data in Thailand is poor and underreported with almost 40% reported as ill-defined causes and other vague causes (Porapakkham et al, 2010; Pattaraarchachai et al, 2010; Polprasert et al 2010; Tangcharoensathien et al, 2006), looking at the trend of mortality is still useful and provides significant information for control and prevention of unintentional deaths. A comparative study of death verification between verbal autopsy and death certificate for 2005 was conducted by Polprasert et al (2010) and found that deaths from transport accidents increased from 2.7% to 5.2% of 6,328 deaths outside hospitals (Polprasert et al, 2010) and increased from 2.8% to 6.4% of 3,316 hospital deaths (Pattaraarchachat et al, 2010). When results from using verbal autopsy (VA) were used of 2,558 hospital deaths with medical records it was show that road traffic accident cases had the highest sensitivity (97.8%) compared to other causes of death.

The results of this study may be useful for guiding health policy on the prevention of transport accidents in both national and local level such as legislation and enforcement of major road safety laws: Helmet for motorcycle riders and passenger, safety belts for car occupants and education of road users etc.

In conclusion, transport accident mortality varied by geographic area and age group. The strength of the implementation and measures for road safety activities can reduce these phenomena. Therefore, efforts to tackle road safety should focus on the risk area, and there should be further study of other determinants of transport accident mortalities, with emphasis on males who are particularly vulnerable group at risk, especially those of working age (20-39 years).

ACKNOWLEDGEMENTS

The authors are grateful to Prof Don McNeil for his helpful advice and suggestions. We also would like to thank the Thai Ministry of Public Health for providing mortality data.

REFERENCES

Anjuman T, Hasanat-E-Rabbi S, Siddiqui CKA, Hoque MM. Road traffic accident: A leading cause of the global burden of public health injuries and fatalities. Proceedings of the International Conference on Mechanical Engineering 2007, 29-31 December, 2007, Dhaka.

Bundhamcharoen K, Odton P, Phulkerd S, Tangcharoensathien V. Burden of disease in Thailand: changes in health gap between 1999 and 2004. *BMC Public Health* 2011; 11:53.

Chiu WT, Kuo CY, Hung CC, Chen M. The effect of the Taiwan motorcycle helmet use law on head injuries. *American Journal of Public Health* 2000, 90: 793-6.

Department of Highways and Faculty of Engineering Prince of Songkla University. Final report of The study of traffic accident costs in Thailand. Faculty of Engineering Prince of Songkla University. 2007. [Cited 2011 Nov 14]. Available from: http://siteresources.worldbank.org/INTTHAILAND/Resources/333200-1177475763598/Sep07-traffic_accident-full-report.pdf

Ditsuwan V, Veerman LJ, Barendregt JJ, Bertram M, Vos T. The national burden of road traffic injuries in Thailand. *Population Health Metrics* 2010; 9:2.

Koedklai S and et al. Sittuation and Trend of Injuries and Deaths related to Transportation Accident 1998-2007: Bureau of Epidermiology Ministry of Health, 2008. (in Thai)

National Health Foundation. Choice in determining the national agenda on control and prevention of traffic accidents. Bangkok: Octoberprinting Press, 2006. (in Thai)

National Highway Traffic Safety Administration. Traffic Salety Facts 2008 Data. 2008. [Cited 2011 Nov 30]. Available from: http://www-nrd.nhtsa.dot.gov/pubs/811159.pdf.

Nantulya, V.M., Reich, M.R. Equity dimensions of road traffic injuries in low-and middle-income countries. *Injury Control and Safety Promotion* 2003; 10: 13-20.

Odton, P. Statistical and Graphical Methods for Analyzing Spatial Counts with Applications to Mortality in Thailand. Doctor of Philosophy in Research Methodology. Prince of Songkhla University. 2010.

OECD. Health at a Glance. Paris: OECD Publishing, 2010.

Pattaraarchachai J, Rao C, Polprasert W, Porapakkham Y, Pao-in W, Singwerathum N, Lopez AD. Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Population Health Metrics* 2010; 8:12.

Pitaktong U, Manopaiboon C, Kilmarx PH, et al. Motorcycle helmet use and related risk behaviors among adolescents and young adults in northern Thailand. *Southeast Asian J Trop Med Public Health* 2004; 35: 232-41.

Polprasert W, Rao C, Adair T, Pattaraarchachai J, Porapakkham Y, Lopez AD. Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. *Population Health Metrics* 2010; 8:13.

R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for statistical Computing. 2011. Vienna: R Foundation for Statistical Computing. [Cited 2011 Sep 6]. Available from: URL: http://www.r-project.org.

Statistics South Africa. Road traffic accident deaths in South Africa, 2001-2006: Evidence from death notification. Pretoria: Statistics South Africa, 2009.

Suriyawongpaisal P, Aekplakorn W. Textbooks of epidemiology and traffic accidents. Bangkok: Holistic Publising, 2003.

Tanaboriboon Y, Satiennam T. Traffic accidents in Thailand. *IATSS research* 2005; 29: 88-100.

Tangcharoensathien V, Faramnuayphol P, Teokul W, Bundhamcharoen K, Wibulpholprasert S. A critical assessment of mortality statistics in Thailand: potential for improvements. *Bulletin of the World Health Organization* 2006; 84: 233-8.

Thairoads Foundation and Thailand Accident Research Center (TARC). Report road accidents in Thailand 2010. Bangkok: National Health Foundation, 2011. (in Thai)

Tongkumchum P, McNeil D. Confidence intervals using contrasts for regression model. *Songklanakarin J Sci Technol* 2009; 31: 151-6.

Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer-Verlag, 2002.

World Health Organization. Global status report on road safety: time for action. Geneva: WHO, 2009a.

World Health Organization. Regional Report on Status of Road Safety: the South-East Asia Region. India: WHO, 2009b.

World Health Organization. Age standardized death rates by country, 2008. Geneva: WHO, 2011. [Cited 2011 Sep 14]. Available from: URL: http://www.who.int/gho/mortality_burden_disease/countries/deaths/en.

CPS17: Demography & Social Welfare Statistics (2)

Assessing Household Welfare by Monitoring Specific Group Inflation Rate in Jakarta

Ayu Paramudita*
Jakarta, Indonesia – paramuditaayu@gmail.com

Abstract

Inflation should vary across households. The difference of consumption behavior inflicts different effect of rising goods price on each household group, so general inflation could not represent condition of all groups. Specific group inflation plays an important role in a variety of ways such as setting wage, measuring macroeconomics indicator precisely, and especially knowing how the marginalized group suffers due to the rising price. In this paper, an attempt has been made to measure specific group inflation rates for vulnerable poor and non-vulnerable poor in Jakarta during 2013-2015. The main purpose of this study is assessing welfare of each group by monitoring specific group inflation rates; vulnerable poor inflation and non-vulnerable poor inflation. Descriptive analysis, U Mann-Whitney nonparametrics, and modified laspeyres index are used to get the study results. Significantly, there was difference on consumption behavior between vulnerable and non-vulnerable poor. The difference was in accordance with the Engel's Law. Vulnerable poor consumption was dominated by food consumption, whereas non-vulnerable poor consumption was dominated by housing as the biggest expenditure. The Consumer Price Index (CPI) of vulnerable poor was higher than non-vulnarable poor along the periods. The inflation rates also show that vulnerable poor experienced higher inflation than non-vulnerable poor. The higher CPI, the more burdensome for vulnerable poor. This indicated worse welfare of vulnerable poor. Furthermore, the higher inflation rates meant that prices of goods consumed dominantly by vulnerable poor were always highly increased so that it weakened their purchasing ability. Yet, the prices of goods, which were considerably consumed by non-vulnerable poor, tended to rise slowly so the inflation rates were not as higher as vulnerable poor inflation. The goods that gave great share of vulnerable poor inflation rates were generally necessity goods such as rice, fish, red onion, chili, ciggarete, electricity, gas, and public transportation fare. As a suggestion, the government should pay more attention to stabilize prices of necessity goods in order to make vulnerable household welfare better.

Keywords: Inflation, Welfare, Vulnerable poor, Non-vulnerable poor.
JEL : E310, I310, I320

1. INTRODUCTION

Inflation as one of important economics indicator gives information about price fluctuation of goods consumed by household. Inflation by Boediono (1986) is tendency of prices to rise generally and continously, but rising price of one commodity isn't named as inflation unless it generates increasing price on the other comodities. The effect of rising price can be felt directly by household. Purchasing ability goes down and burden increases. Having this issue, Sukirno (2004) affirmed by stating that inflation affects household welfare because it diminishes the real revenue. Hence, household welfare could be monitored by inflation rate as a household burden indicator.

Inflation should vary accross households. According to Asra (1989), difference of consumption behavior inflicts different effect of rising price on various society groups. Therefore, specific group inflation is important to observe rising prices impact on each group that can't be represented by general inflation. Moreover, specific group inflation plays an important role in a variety of ways such as setting wage, measuring macroeconomics indicator precisely, and especially knowing how the marginalized group suffers due to the rising price.

Even though household groups might have experienced different effect of rising price as a consequence on different consumption behavior, some researches reveal whoever experiences the higher inflation depends on the structure of inflation itself at that period. Study in 6 regencies of Indonesia from Badan Pusat Statistik (BPS or Central Bureau Statistics of Indonesia) and Asian Development Bank (2006) showed that poverty inflation, which was composed of comodities consumed dominantly by poor household, was higher than national inflation rate. On the other hand, Chema and Malik (1986) and Bhorat and Oosthuizen (2005) studied about poor households in Pakistan and South Africa, who didn't experience worse impact of rising prices consistently when food price was cheap.

So, in order to asses household welfare in Jakarta, an attempt has been made to simulate measurement of specific group inflation rates for vulnerable poor and non-vulnerable poor during 2013-2015. Household whose expenditure per capita is under 1,2 poverty line is categorized as vulnerable poor group, and the rest household is categorized as non-vulnerable poor group. This grouping refers on BPS (2011) that stated vulnerable poor household as a target of goverment assistance receiver. Moreover, BPS (2011a) also identified that this group tended to encounter trancient poverty, a situation when one didn't belong to poor falling into poor because of rising prices. Asian Development Bank (2008) revealed that in many country in Asia, including Indonesia, rising food price would urge household down under the poverty line.

In this paper, CPI and specific group inflation rates are measured monthly and annually by using modified laspayers index. Consumption expenditure of 169 comodities from National Survey of Economics and Social (Susenas) 2012, is used to compile consumption diagram of each group. Comodity price is also used to calculate relative price during the period. All data are taken from BPS.

As a base on measuring specific group inflation rate, the difference of consumption behavior between two groups is tested firstly by using U Mann-Whitney nonparametric test. Nonparametric test is used because classic assumption to do parametrics test is not fulfilled in this case. According to Siegel (1956), U Mann-Whitney test is a best powerfull test as a parametric test substitution. Thus, U Mann-Whitney test is chosen to test the difference of consumption behavior between two independent groups in this paper.

2. THE DIFFERENCE OF CONSUMPTION BEHAVIOR BETWEEN VULNERABLE AND NON-VULNERABLE POOR

First indication of prosperity gap could be seen from groups's consumption expenditures. Vulnerable poor's expenditure was about Rp2.161.632 per month in 2012. It was very small amount compared to non-vulnerable poor whose consumption expenditure reached Rp6.165.471 per month. As initial view, vulnerable poor welfare was extremely lower than non-vulnerable poor welfare.

Compiled by six commodity groups, consumption diagrams of two groups are presented on Figure 1. Consumption of two groups were dominated by food, housing, and transportation expenditure which had different proportion among two groups. Over half of total expenditure was spent on food consumption by vulnerable poor. Contrarily, non-vulnerable poor consumption was dominated by non-food, which was 37,84% on housing as the biggest expenditure. Excluding food, vulnerable poor's housing, clothing, health, education, and transportation expenditure were lower than non-vulnerable poor's.

Vulnerable poor's consumption on food was so high, reaching 50,70%, whereas non-vulnerable poor's was only 33,09%. The greatest food consumption was rice. Vulnerable poor spent 9,97% on rice consumption, higher than non-vulnerable poor, who spent 3,68%. On the other hand, vulnerable poor group consumed meat only 1,65%, less than non-vulnerable poor group whose meat consumption was about 2,41%. Surprisingly, ciggarete consumption exceed other food such as meat, vegetables, and fruits. Vulnerable poor household spent 5,42% of total expenditure on ciggarete, while non-vulnerable poor spent 3,17%.

On housing expenditure, consumption was dominated by dwelling cost and fuel for housing. House rent was the greatest spending of dwelling cost. Non-vulnerable poor spent 24,83%, greater than vulnerable poor, who spent 15,31% for renting house. Contrary, vulnerable poor spent more than another group on fuel for housing, which was dominated by gas and electricity. It was 9,46% for vulnerable poor and 4,17% for non-vulnerable poor.

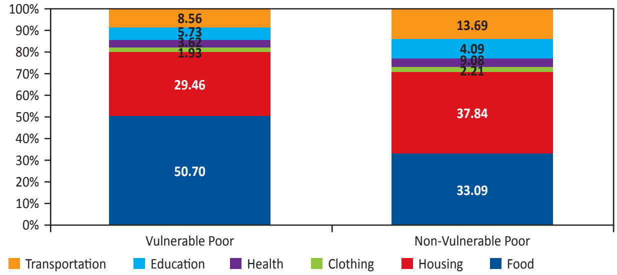


Figure 1. Percentage of Consumption Expenditure by Commodity and Household Groups, 2012

On transportation expenditure, great consumption was on public transportation and gasoline. Vulnerable poor spent 3,63% on public transportation, whereas non-vulnerable poor spent only 2,56%. In contrast, non-vulnerable poor was more consumtive on gasoline, reaching 3,12%, than vulnerable poor, who spent only 1,80%.

From the analysis of consumption diagram above, it is shown that household groups had different consumption behavior. However, significance of the difference of consumption behavior should be tested statistically to make sure the validity. U Mann-Whitney test is used by testing the difference of consumption behavior through food expenditure proportion between two groups. The result shows that consumption behavior among groups was significantly different at 1 percent significance level. Additional U Mann-Whitney test is also taken toward necessity goods. The result shows that consumption of rice, meat, ciggarete, house rent, gas, and public transportation fare were significantly different among vulnerable and non-vulnerable poor at 1 percent significance level.

Overall, consumption behavior between two groups were totally different. Generally, vulnerable poor consumption was dominated by primary consumption, especially food, while non-vulnerable poor tended to consume non-primary goods. This result was in accordance with the Engel's Law, which states as income gets higher, the proportion of food expenditure gets smaller. Unfortunately, the great spending on food might cause rising food price risk on vulnerable poor getting higher. It was because the fastest rising price occurred on food commodity.

3. SPECIFIC GROUP INFLATION RATE: VULNERABLE POOR INFLATION AND NON-VULNERABLE POOR INFLATION

Price level can be shown by CPI and its growth is called inflation. In this paper, simulation of measuring CPI and inflation rates are shown by figures below.

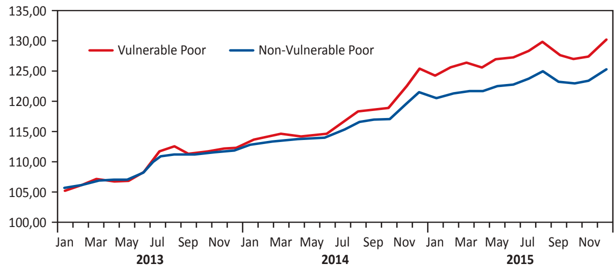


Figure 2. Consumer Price Index (2012=100) by Household Groups

Figure 2 shows that generally, vulnerable poor's CPI was higher than non-vulnerable poor's along 2013-2015. Even so, in the beginning of 2013, both groups experienced the same CPI caused by harvesting time which made commodity prices, especially food, cheap and stable. In the next period, higher vulnerable poor's CPI was caused by fluctuative price of commodity consumed dominantly by vulnerable poor. This fluctuative price was affected by some factors along the period such as crop failure and policy of rising administered price like gasoline, electricity, and gas.

Even though government set gasoline price down after declining of international oil price in 2015, this policy wasn't strong enough to push vulnerable poor's CPI into lower level. BPS reported that in 2015, volatile food inflation (4,84%) was higher than administered inflation (0,39%). So, despite of declining gasoline price, food price, which was highly increasing, made vulnerable poor's price level higher than non-vulnerable poor's. From that observation, vulnerable poor's CPI was greatly influenced by food price. On the contrary, commodity price consumed dominantly by non-vulnerable poor, like house rent and gasoline, rose slowly and caused non-vulnerable poor's CPI was not as high as vulnerable poor's.

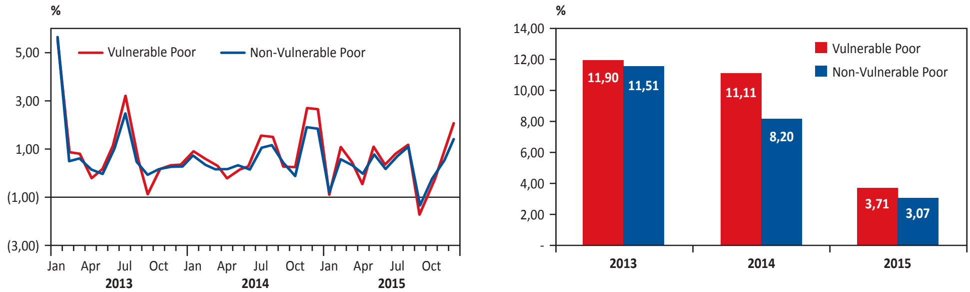


Figure 3. Vulnerable Poor and Non-vulnerable Poor Inflation Rates by Month and Annual

Seen as CPI growth on Figure 3, both monthly group inflation rates were fluctuative along the periods. Vulnerable poor inflation was commonly higher than non-vulnerable poor inflation. Some periods, like July 2013 and November 2014, vulnerable poor inflation were highly increased than another group inflation because of rising price on food, gasoline, and public transportation fare. On the other hand, vulnerable poor also experienced lower inflation, even deflation, on September 2013 and January 2015. On this periods, necessity goods like rice, fish, chili, red onion were on low prices so that could push down vulnerable poor inflation rates.

Annually, both group inflation rates were decreasing in 3 years. In 2015, price reduction of gasoline had main contribution on declining inflation rates. Administered commodity, especially gasoline and gas, which was recorded as deflation, was pressing non-vulnerable poor inflation into about 3,07% from 8,20% on the previous period. Vulnerable poor inflation was also declining from 11,11% into 3,71%, but it remained slightly higher than non-vulnerable poor inflation rate.

Along the periods, vulnerable poor's inflation rate was always higher than non-vulnerable poor's. It meant that rising prices caused a severe burden for vulnerable poor and brought down their purchasing ability badly. So, vulnerable poor welfare was getting extremely worse than non-vulnerable poor welfare throughout the periods.

Further illustration shows household welfare can be explained through a simulation of changing real revenue per capita (expenditure per capita is used as a proxy). Table 1 below is the result of measuring vulnerable poor's real revenue per capita by average vulnerable poor's and non-vulnerable poor's CPI as deflators. Vulnerable poor's real revenue per capita was smaller when facing higher CPI, which was vulnerable poor's CPI itself. The higher CPI, the smaller real revenue per capita. It meant more burdensome for household. Thus, vulnerable poor faced worse welfare than non-vulnerable poor.

Table 1. Simulation Result of Vulnerable Poor's Real Revenue per Capita, 2012-2015

Deflator	Real Revenue Per Capita of Vulnerable Poor Group			
	2012	2013	2014	2015
Vulnerable Poor's CPI	2.161.632,29	1.977.052,71	1.844.930,65	1.698.147,24
Non-Vulnerable Poor's CPI	2.161.632,29	1.981.845,81	1.868.693,14	1.759.075,47

4. SHARE OF SPECIFIC GROUP INFLATION

Commodities whose prices should be depressed to control inflation rate can be identified from the share of inflation. Knowing what kind of commodity which gives great share for inflation, goverment may get easier in making price policy in order to lift household welfare.

Both groups inflation were greatly contributed by food, housing, transportation, and education share as shown by Figure 4. Both groups faced dominant share of food along the periods. Even so, food gave greater contribution on vulnerable poor inflation. In 2015, its share was about 4,89%. On the other hand, share of housing on non-vulnerable poor inflation was 1,65%, greater than share of housing on vulnerable poor inflation. Beside housing, transportation and education also gave greater share on non-vulnarable poor inflation. The others comodity groups, health and clothing, contributed small value on both groups inflation.

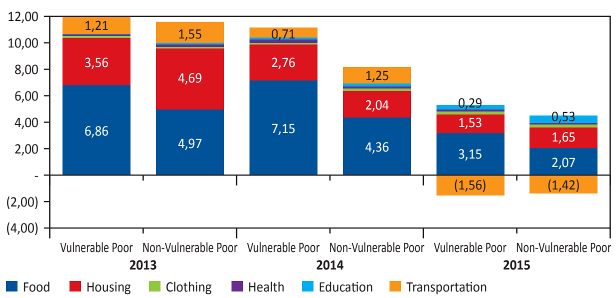


Figure 4. Inflation Share by Comodity and Household Groups

Commodities, whose share was significant toward vulnerable poor welfare, were mainly volatile food. Vulnerable poor inflation share was greatly contributed by rice, instan noodle, fish, egg, vegetable, tofu, fermented soybean, red onion, chili, cooking oil, ciggarete, electricity, gas, public transportation fare, and gasoline. If those prices increased, the impact would be worst on vulnerable poor welfare. So, government should pay attention on keeping these commodities price stable. On the other hand, goods, whose share were great on non-vulnerable poor inflation, were meat, milk, fruit, house rent, housing maintenance, education, and gasoline.

5. CONCLUSIONS

Consumption behavior between vulnerable poor and non-vulnerable poor differed significantly as shown by consumption diagram and U Mann-Whitney test. Based on consumption diagram, the greatest consumption of vulnerable poor was food, while non-vulnerable poor spent the greatest consumption on housing. The difference of consumption behavior distinguished the price level. Simulation of measuring group inflation rates shows that vulnerable poor's CPI was generally higher than non-vulnerable poor's, so was the inflation rates. This proved that vulnerable poor experienced severe burden caused by the drop of purchasing ability. So along 2013-2015, vulnerable poor group faced worse welfare than non-vulnerable poor group in Jakarta.

Some primary commodities, especially volatile food, gave great share toward vulnerable poor inflation along the periods. Government should mantain these prices to control inflation rate, especially for vulnerable poor because price fluctuation could affect marginal group welfare easily. Low and stable inflation is expected in order to attain vulnerable poor welfare better.

Another suggestion that can be given based on the result of this paper is in order to assess household welfare, a sharper indicator, which can use on reflecting the condition of specific group, is needed. This surely can help government on making proper policy. As an example, in case of regional minimum wage of labor policy, labor group inflation rate would be the most appropriate indicator to reflect labor welfare, so that government could set a proper wage for them. Thus, it is essential to have specific group inflation not only for assessing various household welfare but also measuring the others indicator precisely.

REFERENCES

Asian Development Bank. (2008). Food price and inflation in developing asia: is poverty reduction coming to end?. Philipina.

Asra, Abuzar. (1989). Inequality trends in Indonesia, 1969-1981: A re-examination. Bulletin of Indonesian Economic Studies, page 100-110.

Badan Pusat Statistik. (2011). Kompilasi bahan rateknas pimpinan bps provinsi dalam rangka pendataan program perlindungan sosial 2011. Jakarta, Indonesia. Central Bureau Statistics of Indonesia.

Badan Pusat Statistik. (2011a). Workshop wartawan 2011. Jakarta, Indonesia. Central Bureau Statistics of Indonesia.

Badan Pusat Statistik and Asian Development Bank. (2006). Laporan hasil studi pengembangan indikator proksi pemantauan kemiskinan. Jakarta, Indonesia. Central Bureau Statistics of Indonesia.

Bhorat, Haroon. and Oosthuizen, Morne. (2005). The relative inflation experience of poor urban South African households. Labour Market Frontiers September 2015, The South African Reserve Bank, page 1-6.

Boediono. (1986). Ekonomi makro. Yogyakarta, Indonesia, BPFE.

Chema, Aftab Ahmad. and Malik, Muhammad Hussain. (1986). Income specific-inflation rate in Pakistan. The Pakistan Development Review Vol XXV no. 1, page 73-84.

Siegel, Sidney. (1956). Nonparametric statistics for the behavioral science. New York. McGraw-Hill

Sukirno, Sadono. (2004). Makroekonomi teori pengantar. Jakarta, Indonesia. PT Raja Grafindo Persada.

Dominance of the Richest in Brazilian Income Inequality Measured With J-Divergence (1981-2015)

Marcos Dantas Hecksher
IPEA and ENCE/IBGE, Rio de Janeiro, Brazil – mdhecksher@gmail.com

Pedro Luis do Nascimento Silva*
ISI and ENCE/IBGE, Rio de Janeiro, Brazil – pedronsilva@gmail.com

Carlos Henrique Leite Corseuil
IPEA, Rio de Janeiro, Brazil – carlos.corseuil@ipea.gov.br

Abstract

The share of the income inequality explained by the 10% richest members of the Brazilian population is higher than 50%. This contribution is higher in Brazil than what is found for other countries such as the United States (45%), Germany (44%) and Great Britain (41%). The result for Brazil is stable for the period from 1981 to 2015. Inequality was measured using the J-divergence index, which is still not much used in the socioeconomic literature. The J-divergence index can be defined as the sum of Theil's T and L indices, but unlike these and the Gini index, can be easily decomposed as the sum of the individual contributions to the total inequality. We used the publicly available microdata on equivalised household total monthly income from the Brazilian National Household Sample Survey (PNAD), to estimate the J-divergence of income inequality indicators for each year from 1981 to 2015, and the corresponding shares of the inequality explained by each twentieth of the income distribution.

Keywords: Income concentration; Decomposition; Household sample survey.

1. INTRODUCTION

The synthetic indices most used to measure the size of inequality, such as Gini's and Theil's T and L, do not allow us to easily and directly distinguish between how much total inequality comes from each person or each income strata. However, the simple addition of Theil's T and L indices defines another index, less often used in this field of study to date, called J-divergence. This index preserves the main desired properties in an inequality index and adds other interesting ones, such as symmetry, applicability to some statistical tests and, above all, the possibility of being easily decomposed into individual non-negative contributions that aggregate into the total inequality (Rohde, 2016).

Brazil is known to have one of the highest income inequalities in the world (Milanovic, 2015). It is not by chance that reducing inequalities is one of the fundamental objectives of the country included in its Federal Constitution of 1988. In order to monitor and evaluate the fulfillment of this objective, different indicators can be used, each of which allows answering specific relevant questions.

This article presents the evolution, from 1981 to 2015, of the J-divergence index of income inequality calculated with the microdata of the National Household Sample Survey (PNAD) run by the Brazilian Institute of Geography and Statistics (IBGE), as well as the evolution of the contributions of specific groups of individuals. The analysis of these series allows to answer in a simple way some important questions related to the mentioned constitutional objective, but hardly raised with the generally used inequality indicators, such as:

a) What are the shares of the richest, the poorest and the middle group in total inequality?
b) How do these shares compare with those of similar groups in other countries?
c) How did these shares vary over time?

Before answering these questions, section 2 of this paper presents the formal definition of J-divergence and compares its properties to those of other inequality indices, summarizing the main aspects detailed by Rohde (2016). The results found in Brazil are shown and analyzed in section 3, followed by a conclusions section.

2. DEFINITION AND PROPERTIES OF J-DIVERGENCE AS AN INEQUALITY INDEX

This section is based essentially on Rohde (2016) to present the definition and summarize the properties of J-divergence as an inequality index. Formulated by different authors at distinct times, this is a statistic applied to several thematic areas, with various names in the international literature: symmetric Kullback-Leibler divergence, symmetric relative entropy, symmetric Theil measure or J-divergence, in honor of the pioneer article by Jeffreys (1946).

Just as the indices of inequality formulated by Theil (1967) belong to the broader set of generalized entropy indices, J-divergence belongs to a close family, the class of generalized symmetric relative entropy indices. All these indices, based on information theory, seek to synthesize in a number how much a probability distribution differs from another distribution considered a priori. In a fully egalitarian income distribution, any subgroup containing 18% of the population, for example, would share exactly 18% of the total income. In the case of Theil's T index, one starts from the proportions of the population in each subgroup and calculates how much informational content is needed to transform the expected proportions of income (equal to that of the population in an equal distribution) into the proportions of total income effectively earned by each subgroup. The Theil's L index, on the other hand, calculates the information needed to go the other way, from the proportions of income to the proportions of the population. J-divergence, in turn, is an unique index capable of representing indistinctly the information used in both ways, which characterizes its symmetry.[1]

The T and L indices of Theil for a population with N individuals can be expressed, respectively, as:

$$T = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i}{\mu} \ln\left(\frac{x_i}{\mu}\right) \quad \text{and} \quad L = -\frac{1}{N}\sum_{i=1}^{N} \ln\left(\frac{x_i}{\mu}\right),$$

where x_i is the income of each individual i, and $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the population average of the individual incomes. The J-divergence equals the sum $T + L$ and can be expressed by:

$$J = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i - \mu}{\mu} \ln\left(\frac{x_i}{\mu}\right).$$

Note that while Theil-T receives negative contributions from people with a lower than average income, Theil-L receives negative contributions from people with a higher than average income. In J-divergence, individual contributions are positive both below and above the average, or zero in the case of incomes equal to the average. Thus, each individual or group has its own divergence to the average that is always greater than or equal to zero. The sum of these partial divergences equals the total inequality, which makes it simple to calculate the share of each person or group in the total inequality for the whole population.

As Rohde (2016) highlights, besides this additional property, J-divergence has also a range of typical properties of inequality indices, such as:

a) *Scalar invariance*: inequality level does not change if all values are multiplied by a positive scalar.
b) *Anonymity*: inequality level does not change if two or more people switch positions while maintaining the same values as the original distribution.
c) *Population replication*: kept proportions of population with each fraction of total income, population size does not matter.
d) *Pigou-Dalton condition*: inequality increases when a regressive transfer is made (from poorer to richer) maintaining the mean.
e) *Diminishing transfers*: the previous effect diminishes with increasing relative incomes of the parties involved.
f) *Additive decomposability*: the index can be decomposed by the inequalities within each subgroup plus the inequality between the subgroups.

1 Symmetry is one of the necessary properties of a divergence, mathematical concept that fulfills almost all the properties of a measure of distance or metric, except the triangular inequality.

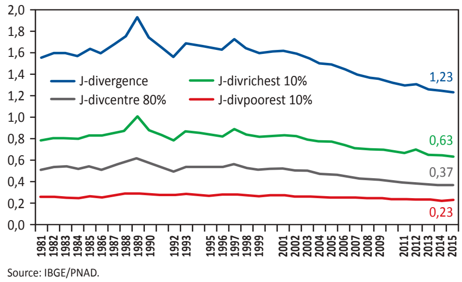
A disadvantage of J-divergence compared to Theil-T and the Gini index, on the other hand, is that the former ceases to be applicable when the distribution includes null incomes.[2] This limitation, also found in Theil-L, derives from the fact that it is enough that one person's income goes to zero for both the index and the contribution of that person to tend to infinity.[3]

3. APPLICATION OF J-DIVERGENCE TO BRAZILIAN HOUSEHOLD SAMPLE SURVEY

This section presents results of the Brazilian equivalised[4] household income inequality measured by J-divergence and the shares of various income strata in total inequality. They were calculated from the publicly available microdata of the National Household Sample Survey (PNAD) from 1981 to 2015, excluding households with null or unavailable income. These series are compared to those constructed by Rohde (2016), based on the Cross National Equivalent File (CNEF) household surveys from other three countries: United States, Germany, and Great Britain.

Graph 1 shows the evolution of the total J-divergence and its decomposition between the contributions of the richest 10%, the poorest 10% and the centre 80% in the distribution of the equivalised household income according to PNAD. The top line represents the sum of the three other lines. The results for the years that are not labeled in the horizontal axis are linear interpolations, because in those occasions PNAD was not collected.

With the top line, one can observe that, from 1981 to 2001, the point estimates of inequality fluctuate strongly in some years, but the period ends with the indicator at a level close to 20 years earlier. Between 2001 and 2015, the inequality measured in PNAD declines almost monotonically.[5] The line of contribution to inequality from the richest 10% runs well above the two other income strata and registers inflections in the same years as the total J-divergence. All lines decline in the period 2001-2015, but the strongest relative fall is in the centre 80%.



Graph 1. J-Divergence of Equivalised Household Income in Brazil and Income Strata Contributions

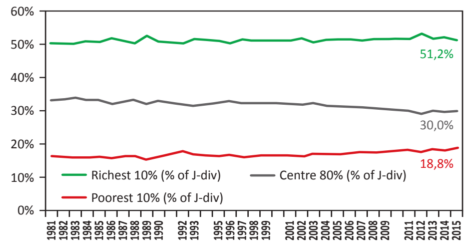
---

2 Note that $\ln\left(\frac{x_i}{\mu}\right) = \ln(x) - \ln(\mu)$ and that $\ln(.)$ is a monotonic transformation.

3 In the case of Theil-T, although it is not possible to define $\ln(0)$ for individuals with zero income, the value of 0 ln(0) is taken as zero, since the expression $\frac{x_i}{\mu}\ln\left(\frac{x_i}{\mu}\right)$ tends to zero when xi approaches zero (Foster, 1983). The first example of application of the index presented by Theil (1967) already included zero incomes: when a single person concentrates all income of the population, $T = \ln(N)$.

4 Using square-root equivalence scale, as it was done for the other countries compared.

5 This decline has been documented by many studies. However, Medeiros, Souza and Castro (2015) found stability of Brazilian personal income inequality by adjusting the PNAD series with estimates obtained from income tax returns.



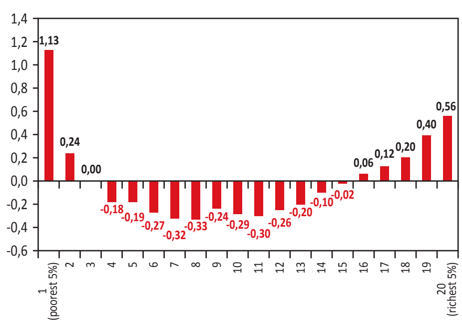
Source: IBGE/PNAD.

Graph 2. Shares of Income Strata in Equivalised Household Income Inequality in Brazil

By dividing the values of the three lower lines by the values of the top line in Graph 1, we obtain the shares of the richest 10%, the centre 80% and the poorest 10% in the total inequality in each year. These series are presented in percentages in Graph 2. The first fact to be highlighted in Graph 2 is that the richest 10% account for more than 50% of the total inequality in all 30 editions of PNAD analyzed here. The most recent share is 51.2%, higher than that observed in the three countries with comparable statistics available: Great Britain (41%), Germany (44%) and United States (45%).

Another fact that becomes clear in Graph 2 is that, in the period 2001-2015, where inequality decreased, only the centre 80% reduced modestly its share in inequality, from 32.1% to 30.0%. This was compensated by an equal increase in the share of the poorest 10%, from 16.7% to 18.8%, while the richest 10% mantained the same 51.2%.

Graph 3 highlights the changes in percentage points of the shares of each twentieth of the distribution between 2001 and 2014.[6] The 15th twentieth is the one which includes the average since the 1981, which limits large levels and variations of the local divergence to the mean. What happens to the other bars of the graph, meanwhile, should be related to factors that are important for the growth of incomes in each group. In particular, it is possible that the formalization of labor and the increases in the minimum wage that occurred during the same period were important to slightly change the distribution of inequality among the income strata, which could be investigated in a future study.



Source: IBGE/PNAD.

Graph 3. Variation in J-Divergence Share by Twentieth of the Distribution in Percentage Points (2001-2014)

6 The period 2001-2014 is analyzed separately because it differs significantly from what occurred in the year 2015 in PNAD. In 2015, affected by the current recession in Brazil, the survey showed a drop of 6.9% in the average of the inflation-adjusted equivalised household income, driven mainly by the richest and the poorest, not so much by the middle group. In 2001-2014, annual growth in the inflation-adjusted equivalised household income was 3.8% on average, 7.1% among the poorest 5% and 2.6% among the richest 5%.

4. CONCLUSIONS

The use of J-divergence as an index of inequality, though not yet popular in the socioeconomic literature, allowed us to explore issues that received little attention so far but relevant to the understanding of the phenomenon. The results for Brazil presented in this short article show that, in addition to the fact that national income is very concentrated in the richest strata (which was already known), these groups are also very dominant in determining the levels of inequality.

The concentration of more than half of the J-divergence in the richest 10% in all the years analyzed, from 1981 to 2015, adds a new aspect to the already known exceptionality of Brazilian income inequality, since this does not occur in any of the three countries for which Rohde (2016) calculated similar series. Moreover, the results show that, at least in Brazil, this synthetic index of inequality – as it is likely to occur to some others, certainly Theil-T – is mostly determined by what happens to the income share of the richest, being much less affected by the share of the poorest.

REFERENCES

Foster, J. E. (1983) 'An axiomatic characterization of the Theil measure of income inequality', *Journal of Economic Theory*, 31(1), pp. 105–121. doi: 10.1016/0022-0531(83)90023-6.

Jeffreys, H. F. R. S. (1946) 'An invariant form for the prior probability in estimation problems', *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007), pp. 453–461. doi: 10.1098/rspa.1983.0054.

Medeiros, M., Souza, P. H. G. F. and Castro, F. Á. de (2015) 'A estabilidade da desigualdade de renda no Brasil, 2006 a 2012: estimativa com dados do imposto de renda e pesquisas domiciliares', *Ciência & Saúde Coletiva*, 20(4), pp. 971–986. doi: 10.1590/1413-81232015204.00362014.

Milanovic, B. (2015) 'Global inequality of opportunity: how much of our income is determined by where we live?', *Review of Economics and Statistics*, 97(2), pp. 452–460. doi: 10.1162/REST.

Rohde, N. (2016) 'J-divergence measurements of economic inequality', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3), pp. 847–870.

Theil, H. (1967) *Economics and information theory*. Amsterdam: North-Holland.

A New Composite Indicator to Measure Wellbeing Index in Egypt

Mahmoud Mohamed Nagib ElSarawy (ISI member)
Central Agency for Public Mobilization and Statistics (CAPMAS), Egypt - Mahmoud_sa@capmas.gov.eg / M_sarawy@yahoo.com

Abstract

GDP has become the standard measure of economic progress in many countries, even though it was only intended as a macroeconomic accounting tool. In fact increasing GDP does not mean that increasing wellbeing, it depends on whether the social costs of such an increase outweigh the benefits. GDP does not include non-economic determinants of well-being such as social relationships, or the distribution of income and wealth, also GDP is a good measure of size, but at some point bigger is worse, not better, wellbeing is multidimensional encompassing all aspects of human life.

Measuring well-being level involves identifying the key components of a better life and then selecting a set of indicators and variables which provide data and information about the progress of society in this area with respect to these outcomes.

The importance of this paper is that policy makers do not focus single-mindedly on GDP. They rather seek to enhance the overall well-being of citizens, today and in the future, taking into account other factors so, it is important to compile wellbeing index in the country. The objective of this paper is compiling a composite indicator to measure well-being index in Egypt, using many variables which affect human life. The proposed indicator contain three elements economic, investments and sustainability that comprise ten dimensions income, economic stability, employment, health, education, infrastructure, income equality, civil society, governance, and environment according to data availability.

Keywords: Income equality, Governance, Human life, Civil society.

1. INTRODUCTION

The concepts of income and living standards are controversial concepts in the economic history and the challenging of raising the standard of living is the most important challenges facing planners and decision makers. Accordingly, the planners must determines the position of the components of the standard of living and also a clear plan that taken into account all of the income and standard of living variables before deciding to increase rates or improvement in living standards.

Wellbeing is about feeling good and functioning well and comprises an individual's experience of their life; and a comparison of life circumstances with social norms and values. Wellbeing exists in two dimensions Subjective wellbeing and Objective wellbeing:

Subjective wellbeing (or personal wellbeing) asks people directly how they think and feel about their own wellbeing, and includes aspects such as life satisfaction (evaluation), positive emotions (hedonic), and whether their life is meaningful (eudemonic). Objective wellbeing is based on assumptions about basic human needs and rights, including aspects such as adequate food, physical health, education, safety etc. Objective wellbeing can be measured through self-report (e.g., asking people whether they have a specific health condition), or through more objective measures (e.g., mortality rates and life expectancy).

The index of well-being appeared in the last decade and it has been measured using a variety of methodologies such as Canadian Index of Wellbeing, the Australian Unity Wellbeing Index and Better Life Index Which was adopted by the "Organization for Economic Cooperation and Development" (OECD), which groups 34 countries, mostly from developed countries. Despite the difference of the constituent elements of each of these indicators, but all of these indicators agree that measuring the well-being of any society must take into account both economic effervescence (or physical) and effervescence moral (or non-physical). For example, an estimation of the level of "human welfare" of the community, according to "index a better life", requires taking into account both Material Living Conditions and the Quality of Life (OECD, 2011), it includes two components:

- The first component is "physical well-being", it contains three elements are "incomes and wealth's of the community members", "jobs and wages" and "housing".
- The second component is "quality of life", it contains eight elements are: "health status" for the community members, "the balance between work and private life", "education and skills", "engage in civil activities", "social connections", "environmental quality ", " personal security "and" subjective well-being. "It is the elements that each of them is worth a careful look.

2. IMPORTANCE OF STUDY

We can get from the introduction that it is very important to compile an indicator to measure the level of wellbeing in Egypt, in order to put an objective basis to discuss the variables affecting the welfare of the community, and to monitor the value of the change in the levels of well-being, identifying the variables responsible for the largest share of change, whether positively or negatively, and helps over a number of years to measure the level of improvement in the living conditions and try to find strengths and weaknesses points.

3. OBJECTIVES

Building an indicator to measure the well-being of the community, using some variables such as Life expectancy at birth, GDP per capita (PPP), Health expenditure per capita (PPP), Employment (total ,% of total labor force) and Literacy rate (adult total ,% of people ages 15 and above).

4. METHODOLOGY

Before calculating the index of well-being it must build the five guides of the key dimensions of well-being separate. And to calculate the evidence of these dimensions, it should first determine by maximum and minimum values (Hypotheses for the guide):

Table 1. Five Guides of the Key Dimensions of Well-Being Separated

BIC	Well-being guide	Min	Max
1	Life expectancy at birth, total (years)	25	84
2	GDP per capita, PPP (constant 2011 international $)	100	40000
3	Health expenditure per capita, PPP (constant 2005 international $)	10	4000
4	Employment, total (% of total labor force)	0	100
5	Literacy rate, adult total (% of people ages 15 and above)	0	100

Second step calculate each component of the six components of the index of well-being as: (The actual value - minimum value) ÷ (maximum value - minimum value).

Then finally calculate the mean for the sex guides to reach the index of well-being.

5. CALCULATE THE WELL-BEING INDEX:

a. **Well-being guide no. (1):** Life expectancy at birth, total (years):

Table 2. Guide of Life Expectancy at Birth

Indicator Name / year	2010	2011	2012	2013	2014
Life expectancy at birth, total (years)	70.34	70.53	70.73	70.93	71.12
Life expectancy guide	76.8%	77.2%	77.5%	77.8%	78.2%

- life expectancy guide (2014) = $\frac{71.12-0}{84-0}$ = 78.2 %

b. **Well-being guide no. (2):** GDP per capita:

Table 3. Guide of GDP per Capita

Indicator Name / year	2010	2011	2012	2013	2014
GDP per capita, PPP (constant 2011 international $)	10102.10	10071.21	10067.08	10050.03	10048.57
GDP per capita guide	77.0%	77.0%	77.0%	76.9%	76.9%

- GDP per capita guide (2014) = $\frac{Log(10048.57) - Log(100)}{Log(40000) - Log(100)}$ = 76.9 %

c. **Well-being guide no. (3):** Health expenditure per capita

Table 4. Guide of Health Expenditure per Capita

Indicator Name / year	2010	2011	2012	2013	2014
Health expenditure per capita, PPP (Constant 2005 international $)	474.81	507.55	542.46	567.17	594.11
Health expenditure per capita guide	64.4%	65.5%	66.7%	67.4%	68.2%

- Health expenditure per capita guide (2014) = $\frac{Log(594.11) - Log(10)}{Log(4000) - Log(10)}$ = 68.2 %

d. **Well-being guide no. (5):** employment, total (% of total labor force):

Table 5. Guide of Literacy Rate of Labor Force

Indicator Name / year	2010	2011	2012	2013	2014
Literacy rate, adult total (% of people ages 15 and above)	72.05	73.32	73.87	75.06	75.43
Literacy rate guide	72.0%	73.3%	73.9%	75.1%	75.4%

- Employment guide (2014) = $\frac{86.8 - 0}{100 - 0}$ = 86.8 %

e. **Well-being guide no. (6):** Literacy rate, adult total (% of people ages 15 and above):

Table 6. Guide of Literacy Rate of People Ages 15 and Above

Indicator Name / year	2010	2011	2012	2013	2014
Literacy rate, adult total (% of people ages 15 and above)	72.05	73.32	73.87	75.06	75.43
Literacy rate guide	72.0%	73.3%	73.9%	75.1%	75.4%

- Literacy rate guide (2014) = $\frac{75.43 - 0}{100 - 0}$ = 75.4 %

f. **well-being index :**

$$\text{well-being index} = \frac{\text{(Life expectancy guide + GDP per capita guide + Health expenditure per capita guide + Population above poverty line at national poverty line guide + Employment guide + Literacy rate guide)}}{5}$$

$$= \frac{3.855}{5} = 77.1\ \%$$

6. MAIN FINDINGS

Table 7. Wellbeing index by guides of the key dimensions

Guide	2010	2011	2012	2013	2014
Life expectancy at birth guide	76.85%	77.18%	77.51%	77.84%	78.17%
GDP per capita guide	77.03%	76.98%	76.97%	76.95%	76.94%
Health expenditure per capita guide	64.43%	65.54%	66.65%	67.40%	68.17%
Employment guide	91.00%	88.00%	87.30%	86.80%	86.80%
Literacy guide	72.05%	73.32%	73.87%	75.06%	75.43%
Wellbeing Index	**76.27%**	**76.20%**	**76.46%**	**76.81%**	**77.10%**

We can find from table (7) that:

- The proposed welfare index contains 5 Guides, each of which is a guide to measure the level of health, educational, employment status and income level.
- The value of well-being index during the period from 2010 to 2014 remained constant between 76.2% and 77.1%. During the same period we found that Life expectancy at birth guide increased from 76.8% to 78.2%, GDP per capita guide decreased from 77% to 76.9%, Health expenditure per capita guide increased from 64.4% to 68.2%, Employment guide decreased from 91% to 86.8% and Literacy guide increased from 72% to 75.4%.

7. CONCLUSIONS

Select items from a strategic vision to raise the standard of living in Egypt should not be isolated from the perception of the overall development, expressing the attitudes and aspirations of the growth in the economy and society of Egypt after the revolution, January 25 and June 30, relying on the development of the financial resources are not sufficient to achieve a sustainable level of living of the society. Because of the characteristics of the weakness in the education, health, social, will have a negative distributional effects lead to polarization in the distribution of wealth and increasing gaps between social classes, and soon the deterioration of living standards in the community, to improve the level of well -being in the coming years to be:

1. The availability of a suitable position for the equitable distribution of shares of government spending on community members, which reflecting the state of justice in the distribution of income and reduce disparities between social classes.
2. A good, free and comprehensive health care system, covering all citizens, and there are no obstacles, social or financial constraints in access to the service, whether in rural or urban areas.
3. Correct the structural distortions in the economy are responsible for scalable rates of income growth, and the reduction of the value added, which leads to lower per capita GDP.
4. Obligation of the State to free education at all levels, with the continuous improvement in the educational materials and taking into account the contemporary moment includes everything changes and the lack of any distinction between males and females in educational opportunities.

REFERENCES

Anand, S. and A. Sen (2000) "The Income Component of the Human Development Index" *Journal of Human Development* Vol. 1, No. 1, Pp. 83-106.

Axel Dreher, (2005), "The Influence of Globalization on Taxes and Social Policy – an Empirical Analysis for OECD Countries", Thurgau Institute of Economic, Switzerland and University of Konstanz, Department of Economics.

Battle, Ken and Sheri Torjman (1993) *The Welfare Wall: The Interaction of the Welfare and Tax System* (Ottawa: Caledon Institute for Social Policy).

David S. Johnson, (2004), "Measuring Consumption and Consumption Poverty: Possibilities and Issues". Prepared for "Reconsidering the Federal Poverty Measure" American Enterprise Institute.

Daron Acemoglu, (2001), "Human Capital Policies and the Distribution of Income: A Framework for Analysis and Literature Review Treasury Working Paper 01/0 www.treasury.govt.r2/workingpapers/2001/twpd- 3.pdf

http://cdn.media70.com/national-accounts-of-well-being-report.pdf

Haider A. Khan, (2004), "Development as Freedom, University of Denver". www.e.u-tokyo.ac.jp/cirje/research/03research02dp.html

McGillivray, Mark. (2007). Human Well-being: Issues, Concepts and Measures. In Mark McGillivray, ed. Human Well-Being: Concept and Measurement.

Posted at www.csls.ca under conferences), also Research Paper R-99-3E, Applied Research Branch, Human Resources Development Canada, December.

Robert Holzmann and Steen Jorgensen, (2000), "Social Risk Management: A New Conceptual Framework for Social Protection and Beyond", Social Protection Discussion Paper Series No. 0006, Social Protection Unit, Human Development Network, the World Bank.

Thompson, S., & Marks, N. (2008). *Measuring well-being in policy: issues and applications*. Report commissioned by the Foresight Project on Mental Capital and Well-being, Government Office for Science.

Townson, Monica (1999), *Health and Wealth: How Social and Economic Factors Affect our Well Being* (Halifax: Lorimer and Ottawa: Canadian Centre for Policy Alternatives).

Van de Ven, Brugt Kazemier and Steven Keuning (1999), "Measuring Well-being with an Integrated System of Economic and Social Accounts," Division of Presentation and Integration, Department of National Accounts, Statistics Netherlands, May 17, unpublished report.

Wiebke Kuklys and Ingrid Robeyns, (2004), "Sen's Capability Approach to Welfare Economics". www.papers.mpiew-jena.mpg.de/esi/discussionpapers/2004-03.pdf

World Bank, (2006), "Equity and Development", World development report. Co-publication of the World Bank and Oxford University Press.

World Bank, Ghana Welfare Indicators Questionnaire (CWIQ) Survey 1997/98.

The Comparison of Three Methods Conjoint Analysis Based on Respondent Time to Determine the Choice of Stimulation Cards

Fitri Catur Lestari*
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – fitricaturlestari@stis.ac.id

Abstract

There were many types of method in conjoint analysis such as fullprofile design with ranking response (method A), fullprofile design with rating response (method B) and pairwise comparison with ranking response (method C). It was necessary to evaluate and compare them based on the result and the respondent time to determine their choice.

The purposive sampling method had been used to select 125 students as research samples. The consideration of samples selection was based on the segments that supposed as representative the population such as GPA, level/grade, high school origin, gender, and economic status (income) of the parents.

There were different result of three methods of conjoint analysis: respondent's preference of lecturer quality in method A and C were mastery of course material, ability to motivate students, character, communication and teaching method. While the method B, the sequence attributes that prefer were mastery of lecture, character, communication, ability to motivate students, and learning method.

By 5% significance level paired t test, the result weree: there were significant duration (time) differences between three methods (p value= 0.000), the method A was a method with the longest time for respondents to finished the stimulation card evaluation and in contrast the method C was a method in the shortest time for respondents to finished it.

Keywords: Fullprofile design; Ranking response; Rating response; Pairwise comparison.

1. INTRODUCTION

The conjoint analysis is a part of multivariate analysis. The conjoint analysis concern in preference study. Chaplin defined preference as the attitude that prefer something to other things (Chaplin, 2002). In the conjoint analysis there are varied methods. It rely on the attributes combination design (stimuli), data collection and analysis procedure. Many types of methods in conjoint analysis are such as fullprofile design with ranking response (method A), fullprofile design with rating response (method B) and pairwise comparison with ranking response (method C). It is necessary to evaluate and compare them based on the result and the respondent time to determine their choice (represented by the stimuli cards. The stimuli cards differ conjoint analysis with the other analysis).

The conjoint analysis can be applied in education field. The lecturer that is a main component in the university have an important role in determining the quality of the university (Sudiana, 2003). The lecturer quality significantly (sig. 0,01) influences the students learning motivation and has the highest correlation compared with other factors (Pujadi, 2007). The students that were directly involved with the lecturer in the learnig proccess surely have preference on the lecturer quality. Therefore, the students preference on the lecturer quality is neccessary to be investigated by conjoint analysis.

The objectives of this research are to know student preference of the lecturer quality on STIS by three methods of conjoint analysis and to know respondent time/duration to evaluate the stimuli cards by three methods of conjoint analysis.

2. METHODS

Three methods of conjoint analysis are used to know student preference of the lecturer quality on STIS. While to compare the respondent time/duration to evaluate the stimuli cards, this research used t test.

Consecutively five steps in the conjoint analysis are:

A. Formulate The Problem

The first step in conjoint analysis is formulate the problem (Aaker DA, 1980). Formulating the problem start from defining the product as a set of attributes that consist of several level. The information about attributes that represent the consumer preference can be gained from expert discussion, secondary data exploration or pre-survey (Rosada, 2002).

The lecturer quality is formed in attributes that consist of several levels. The attributes of lecturer quality in this research are determined by pre-survey. The result of the pre-survey are:

Table 1. The Attributes that Influence the Student Preference of Lecturer Quality

No.	Attribute	Percentage
1	Lecture delivery	24.00
2	Learning method	20.00
3	Lecture mastery	19.00
4	Personality	11.50
5	Ability to motivate the student	9.00
6	Other (8 attributes)	16.50

Based on table 1, there are 5 (five) attributes that have large percentage between 9,00% sampai 24,00%. Each attribute has two levels: lecture delivery (two-way communication and one-way communication), learning method (creative and not creative), lecture mastery (mastery and not mastery), personality (humorist and serious), and ability to motivate the student (motivate the students and do not motivate the students).

B. Design The Stimuli

The combination design used in this research is full profile or large factor evaluation (Sawtooth, 2001). Because there are 5 (five) attributes, there are 25 or 32 stimulation cards (that contain many characteristics of the lecturer specially exact field lecturer). Method A dan B are defined as conjoint analysis with fullprofile design.

Beside full profile design, pairwise design (Sawtooth, 2001) also used in this research and defined as method C. Because there are 5 attributes, the simulation cards that evaluated by the students or respondents are 5(5-1)/2 or 10 cards (that also contain many characteristics of the lecturer specially exact field lecturer). In this step, the questionnaire was made in which questions that are relevant with demography variable to know the characteristic of the students.

C. Determine the Method of Data Collection

The survey used primary data by interview and simulation cards evaluation that involved 125 students as respondents. They drawn by purposive sampling method with sample selection was based on the equalization in segments that allegedly there are different preference of the students on the quality of lecturers ie GPA, grade, high school origin, gender, and parent economic capacity (parent income). The evaluation response is rank response. The purpossive sampling method is one of sampling methods that is included in non-probability sampling by collecting the information from appropriate sources that are supposed have ability to give information that is needed or they are the only sources who can give the information that is needed by us (Wibisono, 2003). In the next writing, mentioning “students” means “respondents” because of the nonprobability sampling do not conclude for the population. The research was held in STIS on Wednesday-Thusrday, July 20th-21th, 2011.

D. Choose the Procedure in Conjoint Analysis

Analysis procedure used in this research is regression method with dummy variables. Generally the base model of conjoint analysis (Kuhfeld, 2000) with that procedure:

$$Y_{ij} = \beta_0 + \sum_{i=1}^{m}\sum_{j=1}^{k}\beta_j X_j + \varepsilon_k$$

Y_{ij} = Rank all respondents
β_o = Intercept
k = Number of levels in i-th attribute

m = Number of attributes
X_{ij} = Dummy variable in i-th attribute j-th level
β_{ij} = Part worth i-th attribute j-th level
ε_{ij} = Error

By regression model, part worth of the levels in each attribute (NKT) can be determined and are used to determine importance value of the level compared by other level in an attribute. After this proccess, Important Relative Value (NRP) can be counted by formula:

$$NRP_i = \frac{UT_i - UR_i}{\sum_{i=1}^{k}(UT_i - UR_i)}$$

NRP_i = NRP of i-th attribute
UT_i = the highest NKT of i-th attribute
UR_i = the lowest NKT of i-th attribute
k = Number of attributes

Thurstone procedure used to describe about pairwise of the attributes by pairwise comparison (Rosada, 2002). The steps are:

1. Count the frequency matrices by adding the score of all observations by this rules:

$$F_{ij} = \begin{cases} 1, & \text{if attribute i>attribute j} \\ 0, & \text{if attribute i<attribute j} \\ 0.5, & \text{if attribute i=attribute j} \end{cases}$$

   F_{ij}= frequency of i-th row and j-th column (column more important than row)

2. Count the proportion matrices (P_{ij}) by dividing each component of frequency matrices with the number of respondent.

$$P_{ij} = \frac{F_{ij}}{n}$$

3. Transform the proportion matrices component to standard normal (Z_{ij}).
4. Order the column of Z matric from column with smallest mean to greatest one.
5. Count the nearest columns difference.
6. Count scale value with first cale value is zero and the next scale value is the cummulative of the former scale value.
7. Conclude the important factors

The statistic method that is used to compare the respondent time/duration to evaluate the stimuli cards is t test (Walpole, 1992).

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$\bar{d}$ = average of samples data difference in 2 different treatment
μ_d = average of population data difference in 2 different treatment
s_d = standard deviation of samples data difference in 2 different treatment
n = number of samples/observation

The data preparation and analysis were proccessed by software: SPSS 13.0, Microsoft Excel 2003, dan SAS 9.1.

3. RESULT

By method A and C, the student preference on the lecturer quality based on important relative value (NRP) and scale value (NS) consecutively from prefered attribute until unprefered attribute are:

1. Lecture mastery (NRP=31.05%, NS=0.00)
2. Ability to motivate the students (NRP=21.01%, NS=0.32)
3. Personality (NRP=20.29%, NS=0.60)
4. Lecture delivery (NRP=17.44%, NS=0.72)
5. Learning method (NRP=10.21%, NS=0.86)

By method B, the attribute that have the highest influence to student preference is lecture mastery with NRP = 32.79%. The next attribute is personality, lecture delivery, ability to motivate the students and learning method with NRP in sequence 17.82%, 17.20%, 16.75% dan 15.44%. The lecturer who is liked most by the students is the smart lecturer (mastery in his/her field) with highest NKT 2.15. Consecutively the type of lecturer: humorist (NKT=1.17), communicative or two-way communication (NKT=1.13), able to motivate the students (NKT=1.10), and creative in learning method (NKT=1.01).

The student preference of the lecture quality is varied on the segments: GPA, grade, high school origin, gender, and parent income. The graphs below are describing the segmentation:

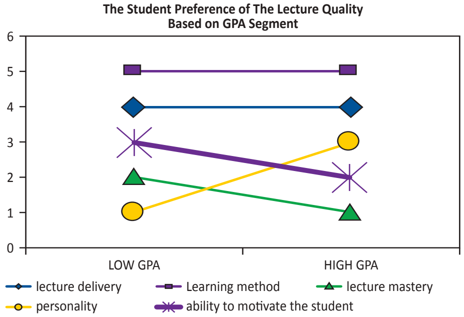


Figure 1. The Student Preference of The Lecturer Quality Based on GPA Segment

By Figure 1, the high GPA students prefer smart lecturer (mastery in his/her field). While students who have the lower GPA like the humorist lecturer. The high GPA student, after mastery attribute, like motivating lecturer and then humorist lecturer. While the low GPA student, after personality attribute, like smart lecturer and then motivating lecturer. Both of them place the lecture delivery attribute as 4th important attribute and learning method as the last important attribute.

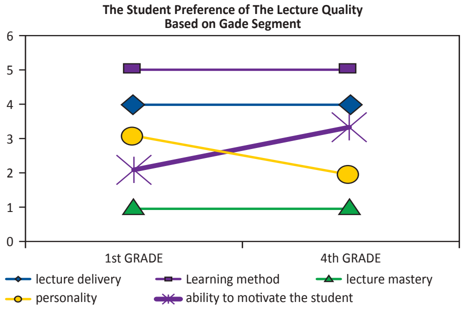


Figure 2. The Student Preference of The Lecturer Quality Based on Grade Segment

The student preference on lecturer quality is different in the 2nd and 3rd attribute (Figure 2). The first level students prefer lecturer who motivates the students than humorist lecturer. In contrast, the fourth level students prefer the humorist lecturer rather than the motivating lecturer. It is assumed because the first grade students need to be motivated at the first learning process in the university and the fourth grade students need a relaxation in the class by humorist lecturer on their last learning process in the university.

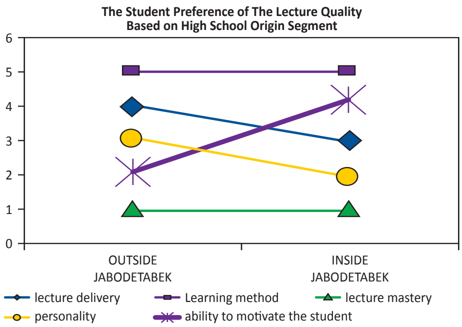


Figure 3. The Student Preference of The Lecturer Quality Based on High School Origin Segment

From figure 3, it could be describe that the students from outside Jabodetabek (Jakarta Bogor Tangerang Bekasi) prefer motivating lecturer, humorist lecturer and communicative lecturer but the students from Jabodetabek prefer humorist lecturer, communicative lecturer and motivating lecturer.

By presupposition, the students from outside Jabodetabek need the motivation because they live far from their family. The students from Jabodetabek live in urban city so they need the humorist lecturer in the class or learning proccess to minimalize the pressure of their environment.

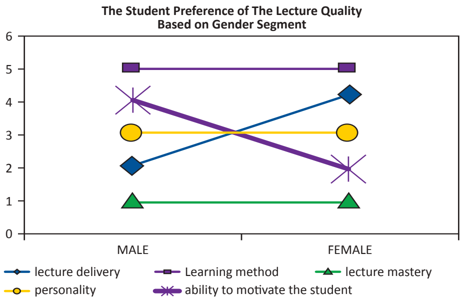


Figure 4. The Student Preference of The Lecturer Quality Based on Gender Segment

The female students prefer motivating lecturer to communicative lecturer while the male ones prefer lecturer who deliver the lecture by two way communication or communicative lecturer to motivating lecturer (Figure 4). It might caused that female students are dominated by their feeling and need motivation from the lecturer than the male one that are dominated by their mind or logic.

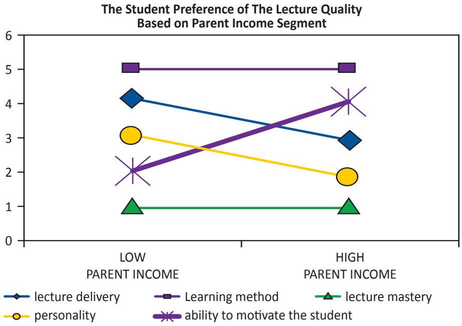


Figure 5. The Student Preference of The Lecturer Quality Based on Parent Income Segment

There are different preference by parent income segment like the previous segments. The students whose parent have the low income prefer consecutively motivating, humorist and communicative lecturer. Contrarily the students whose parent have the high income prefer consecutively humorist, communicative and motivating lecturer. It assumed that in the low parent economic status, students need motivation from the lecturer rather than in the high one. In the high parent economic status, students want to fullfill their tertiary need such as joke, entertainment, comfort situation, etc.

The next result is the comparison of three methods conjoint analysis based on respondent time to determine the choice of stimuli cards.

Table 2. Description of Time in Three Methods (minutes)

Method	Mean	Number of Samples	Standard Deviation	Standard Error Mean
A	2.71	125	3.7219	0.3329
B	2.04	125	1.4629	0.1308
C	1.25	125	0.7715	0.0690

Table 2 shows that the method A is a method with the longest time (average= 2.71 minutes) for students to finished the stimuli card evaluation and in contrast, the method C is a method in the shortest time (average= 1.25 minutes) for students to finished it. In method A, the evaluation time has the largest variance (square of standard deviation) while in method C, it has the smallest one.

Table 3. Time Difference Test

Paired Method	t-value	Degree of freedom	p-value
A and B	14.556	124	0.000
A and C	24.893	124	0.000
B and C	19.346	124	0.000

Based on table 3, the conclusion is there is time (for students to finished the stimuli card evaluation) difference in three methods in conjont analysis because all p-values are lower than alpha 5%. By 5% significance level, there is significantly duration (time) differences between three methods, method A is a method with the longest time for students to finished the stimuli card evaluation, in contrast, the method C is a method in the shortest time for students to finished it.

4. CONCLUSIONS

There were different result of three methods of conjoint analysis: respondent's preference of lecturer quality in method A and C were mastery of course material, ability to motivate students, character, communication and teaching method. While the method B, the sequence attributes that prefer were mastery of lecture, character, communication, ability to motivate students, and learning method.

By 5% significance level paired t test, the result weree: there were significant duration (time) differences between three methods (p value= 0.000), the method A was a method with the longest time for respondents to finished the stimulation card evaluation and in contrast the method C was a method in the shortest time for respondents to finished it.

REFERENCES

Aaker DA, Day GS. (1980). Marketing Research. New York : John Willey&Son, Inc.

Chaplin, JP. (2002). Kamus Lengkap Psikologi. Edisi kelima. Terjemahan: dr. Kartini Kartono. Jakarta: PT. Raja Grafindo Persada.

http://www.sawtooth.com, Conjoint Analysis, (2001).

Kuhfeld, WF. (2000). "Conjoint Analysis Examples", SAS Institut, Inc. http://www. sawtooth software. com. [28 Januari 2004].

Pujadi, A. (2007). Faktor-Faktor yang Mempengaruhi Motivasi Belajar Mahasiswa: Studi Kasus pada Fakultas Ekonomi Universitas Bunda Mulia. Business & Management Journal Bunda Mulia, Vol: 3, No. 2.

Rosada, R. (2002). Perbandingan Metode Pairwise Comparison dan Full profile dalam Pengumpulan Data untuk Analisis Konjoin. Skripsi. Jurusan Statistika IPB. Bogor.

Sudiana, IN. (2003). Peranan Profesionalisme Tenaga Pengajar (Dosen) terhadap Proses Pembelajaran di Perguruan Tinggi. Jurnal Pendidikan dan Pengajaran IKIP Negeri Singaraja, No. 4 TH. XXXVI.

Walpole, RE. (1992). Pengantar Statistika. Edisi ke-3. Jakarta: PT. Gramedia Pustaka Utama.

Wibisono, D. (2003). Riset Bisnis. Jakarta : PT. Gramedia Pustaka Utama.

Thursday,
23 March 2017

Plenary Seasson 02 (PS02):
Statistical Methodology and Applications

Quantitative Risk in Commercial Banking
Agus Sudjianto

Fusion Learning: Fusing Inferences from Diverse Sources for More Powerful Findings
Regina Y. Liu

Variable Selection Techniques for Analyzing Huge-Dimensional Datasets
Naveen Naidu Narisetty

This page is intentionally left blank

Quantitative Risk in Commercial Banking

Agus Sudjianto, Ph.D.
Executive Vice President, Head of Corporate Model Risk

Quantitative risk in Commercial Banking

Statistical methodology and applications

Agus Sudjianto, Ph.D.
Executive Vice President, Head of Corporate Model Risk

Presentation outline

- Introduce applications of statistical models in banking and finance and their practical challenges
- Discuss more in-depth credit model for stress testing and its challenge
  - Introduce credit modeling concept
  - Review available methodology in statistics that can be applied
  - Illustrate the problem using an example
  - Discuss the challenge for further research

1

Risk management and models

Credit, market, and operation

- **Credit risk.** Potential loss due to failure of a borrower to meet its contractual obligation to repay a debt in accordance with the agreed terms
  - Loan, a contractual agreement that outlines the payment obligation from the borrower to the bank
  - Credit events include bankruptcy, failure to pay, loan restructuring, etc.
- **Market risk.** Potential loss due to changes in market prices or values
  - Traded products are priced using models, particularly for less liquid products
  - Credit risk can arise in the trading book as counterparty credit risk

2

Risk management and models (continued)

Credit, market, and operation

- **Operational risk.** Potential loss due to failure in process, system, or people
- **Mathematical models.** Applied across the above risk disciplines
  - For retail and commercial banks such as Wells Fargo, credit risk is generally the most critical risk

3

Examples of risk management models

Credit risk models

Credit scoring Underwriting and account management: Rank order of accounts or customers in terms of probability of default Collection: Rank order to determine percent recoverable balance from delinquent account	Approach: Regression (GLM), Regression Tree or other machine learning technique Challenge: Making reject inference due to non-matched external trades in surrogate data; Model performance constrained by 'Fair Lending' regulation on variables to use
Regulatory capital (Basel) Calculate regulatory capital or Risk Weighted Asset by estimating PD, LGD, EAD to make sure that the bank sufficiently holding capital during downturn	Approach: Regression (GLM) or Regression Tree Challenge: Modeling portfolio with low default probability
Loss forecasting and stress testing Predict future expected loss (PD*LGD*EAD) under various macro-economic scenarios	Approach: Survival model (continuous or discrete), vintage decomposition, GLM, semi-parametric regression, time series Challenge: Impacted by policy and frequent operational changes (unobserved data)

4

Examples of risk management models

Market risk models

Trading model – pricing, Value at Risk (VaR) Pricing derivatives to market observable data and Profit/Loss attributions to market driver Calculate extreme exposure to market risk for trading limits or regulatory capitals	Approach: Pricing – Stochastic Partial Differential Equations solved using lattice or Monte Carlo methods; VaR – Approximation models using Greeks (partial derivatives) Challenge: Calibration can be difficult, some parameters (e.g., correlations) don't often have market observables to imply; Tail behavior is extremely dependent on distributional assumptions
Counter party Quantify (derivative) exposure in the event of counterparty default Credit valuation adjustment to calculate market implied loss Potential future exposure for limit setting and capital calculation	Approach: Similar to pricing models but simplifications are made to speed up calculation Challenge: Dealing with correlation (wrong-way risk); Computation

5

Examples of risk management models

Mortgage servicing right and asset liability management

Mortgage Servicing Right (MSR)

Stochastic Risk Model is designed to generate market risk sensitivities for MSR to various market risk factors, such as: Interest rates, volatility, mortgage basis, home price, unemployment rate, and primary-secondary spread risk

The output of the model is used to measure the MSR's risks, limit setting, and hedging strategy

Asset and liability management

Measure, report, and manage risk sensitivities to market environments such as interest rates

Models are used to measure economic value and duration of earning assets and funding sources, forecast expected net interest income, produce yield, option-adjusted duration, convexity profiles, and scenario analyses

Approach:

- SDE for interest rate model: A stochastic model used for the term structure of volatility
- Simulation/Lattice models: Measure expected values of instruments and associated embedded options under different rate paths
- Regression models for Prepayment/Default/Loss Severity
- Regression models for Deposit Yield/Attrition
- Econometric model to capture impact of economic drivers

Challenge:

- Illiquid market, lack of market prices to calibrate model parameters
- Less frequent data (e.g., quarterly) applied to daily measurement
- Sensitive to input assumptions to the models

6

Examples of risk management models

Operational risk models

Financial crime

Anti-Money Laundering: To estimate the inherent money laundering risk of bank customers. This risk level then drives due diligence applied to the customer

Fraud: To detect fraudulent activities

Approach: Rule-based, Regression Tree, and other machine learning approach both supervised and unsupervised

Challenge:

- Noisy training data: Risk factors' contributions may be counterintuitive, requiring manual optimization to create a stable, understandable set of coefficients
- Risk factors embody both judgment and multiple parametric controls

Operational risk model

Operational risk (including legal risk) is a risk "resulting from inadequate or failed processes, people, and systems or from external events."

To estimate 1-year op-risk exposure at 99.9th percentile of the enterprise-level distribution of potential aggregate operational losses

Approach: Loss Distribution Approach, Extreme Value Distribution, and Copula

Challenge:

- Operational loss data are very sparse
- Severity distributions are typically heavy-tailed; extreme large loss may play an oversized role
- Parameter stability and uncertainty

7

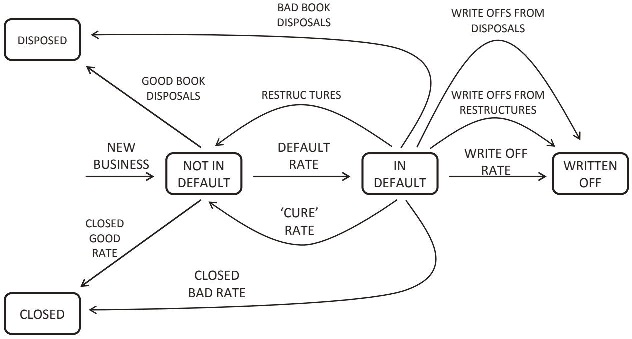
Credit risk stress testing model

- Banking organizations with consolidated assets larger than U.S. $10 billion are required to perform an annual Comprehensive Capital Analysis and Review (CCAR) by the U.S. regulators
  - To assess whether banks have sufficient capital to absorb losses and able to continue its operations during adverse economic conditions
  - The outcome of stress test is publicly announced
- Stress testing is a very large scale modelling exercise applied to all business aspects: Credit, market, operational, and revenue/income
  - Several ~~hundreds~~ very complicated models are built, validated, and applied
  - Model risk management is very crucial
- For a commercial bank, credit loss is the most crucial
  - Sophisticated credit models are applied for the entire portfolio at various granularity levels (account/obligor to pool level)
  - Models are subjected to various economic scenarios (internal and Federal Reserve Bank scenarios)
  - Multiple models (with different structure) are applied for a given portfolio: Primary, confirming, and challenger models to compare model outputs

8

Elements of credit loss model

- Probability of default (PD): The likelihood that the borrower will fail to make full and timely repayment
- Exposure at default (EAD): The expected value of the loan at the time of default
- Loss given default (LGD): The amount of the loss if there is a default, expressed as a percentage of the exposure at default
- Models often capture more detail stages



9

What we learn from event history analysis

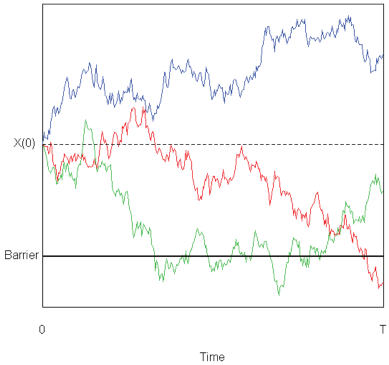
- Survival analysis
  - Parametric approach: Accelerated Life Model
  - Semi-parametric approach:
    - Cox Regression
    - Discrete time life model
- The shape of hazard rate from point process view
  - Brownian motion and the first hitting barrier and its discrete counterpart multi-states formulation
  - Inverse Gaussian Distribution
- Competing risk and recurring events
- Frailty
  - Heterogeneity
  - Unobserved factors
- Dynamic factors
  - Land-marking
  - Joint Longitudinal Survival
  - Multi-states

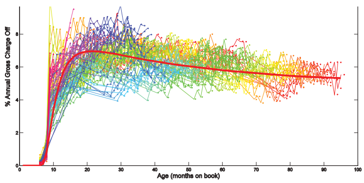
Credit modelling poses its own unique challenges!

10

Default process and modelling framework

- Consider the ability of borrowers to repay their loans following Geometric Brownian Motion
- Default event as threshold crossing of asset value
- The distribution of the *first passage time* (i.e. default time), t, follows Inverse Gaussian (IG) distribution
  - t~IG(m,s)
  - Covariates alter the default time distribution through its parameters
    - m = f(**x**) and s = g(**x**)
- Commercial loans: Rare default experience
  - Simplified parametric assumption is often used
- Consumer loans: Plenty default data are available
  - Model default directly using intensity model or semi-parametric model (e.g., survival analysis)
- In real life it is more complicated, which we will discuss further





11

Concept in credit modelling

Application to stress testing

- Credit event
  - A customer or obligor throughout the life of their loans may experience various events. Credit modelling is event history modelling to predict the future
- Duration (time)
  - Loan age: Dynamic effect of loan age since the time of origination or time to maturation
  - Calendar time: Effect of exogenous (i.e., macro-economic) factors
  - Prediction time: Single or multi-period prediction (time since last observation often called snapshot time)

12

Concept in credit modelling (continued)

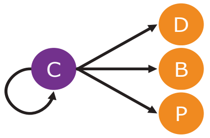
Application to stress testing

- Risk factors
  - Idiosyncratic or systemic factors
  - Static or dynamic factors
    - Characteristics at origination time or factors that are dynamically change throughout the life of the loan
      - Predictable factors: Future values are predicted
      - Unpredictable factors: Predicted future values are not available
  - Unobserved (latent) factors

13

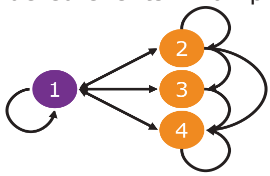
Credit events

- A customer or obligor throughout the life of their loan may experience various events:
  - Terminal events: Default, bankrupt, prepay (attrite), etc.
  - Non-terminal events: Credit rating migration, delinquency, work-out, etc.
- Competing events (risks)
  - Unordered events. Example: (D)efault, (B)ankrupt, (P)repay



Modeling approach:
- Censored treatment for other than the event of interest
- Multinomial (logit)

  - Ordered events. Example: Credit Rating Up/Down Grade

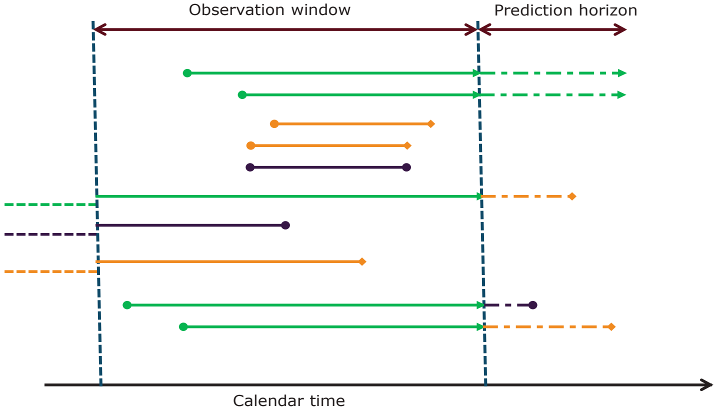


Modeling approach:
- Censored treatment for other than the event of interest
- (Generalized) proportional-odds (ordered) regression (logit)

14

Duration (time) in credit

- Triplet of time in credit model (m,t,k)
  - m: Loan age, time from origination (months on book)
  - t: Calendar time
  - k: Time horizon of prediction



15

Choice of modelling approach

- Time basis among the triplet time: Loan age, calendar time, or prediction time
- Continuous or Discrete Time
  - Continuous time: Survival Analysis such as Cox Regression
  - Discrete time: Generalized Linear Model (GLM) with a proper choice of link function

16

Choice of modelling approach (continued)

- Parametric or Semi-Parametric
  - Parametric/Structural model with proper choice of parametric distribution (Weibull, Gamma, Inverse Gaussian) regression
    - Useful when data is scarce
  - Semi-Parametric Regression (Continuous or Discrete Time)
    - Panel regression for discrete time
      - Parametric for known risk factors
      - Non-parametric for unobserved factors (represented by time)
- Granularity: Account/Borrower or Pool Level
  - Data size, computation time, accuracy
  - Granularity of risk factors and response variables
- How do we do this for predictive model?

17

General modeling framework

- Let T a random time of account closing (e.g., due to default or attrition/prepayment), the hazard function is modeled as a regression with $g(.)$ link function and covariates $Z(s)$

$$\lambda\{t|Z(s)\} = g[\lambda_0(t), Z(s)]$$

- Where $\lambda_0(t)$ is the baseline hazard to represent the effects of unobserved factors and s is the observation time which can be:
  - Static such as time of origination, $s = 0$
  - Dynamics
    - Last snapshot information without future prediction
    - Including future prediction, i.e. $s = t$ and prediction model $Z(t)$ is available (e.g., utilization or spend rate) or macro-economic factors

$$Z\{t|X(s)\} = h[Z_0(t), X(s)]$$

18

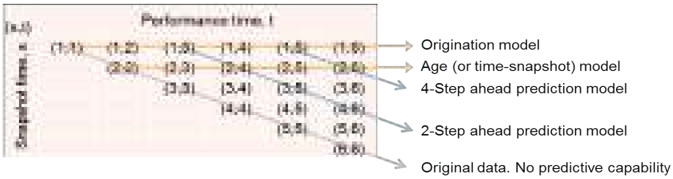
Dynamic covariates and data stacking

Dynamic factors that no future prediction are available but they are critical such as refreshed FICO, utilization, etc. and need to be handled through data stacking approach

Observation Data

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	675	Current	13-Jan(1)	1	7.2	0	0
[illegible]	690	Current	13-Feb(2)	2	7.2	0	0
[illegible]	630	30	13-Mar(3)	3	7.2	0	0
[illegible]	620	60	13-Apr(4)	4	7.2	0	0
[illegible]	620	90	13-May(5)	5	7	0	0
[illegible]	620	120	13-Jun(6)	6	6.7	1	0

Dynamic Factor without future predicted values



19

Dynamics stress testing model

Dynamics of Horizon Prediction:

$$\lambda_i(t|s) = \beta_0(k) + \boldsymbol{x}_i^T(s,t)\boldsymbol{\beta}(k)$$

Prediction of time *t* given the ‘snapshot’ information at time *s*

Dynamic covariates:

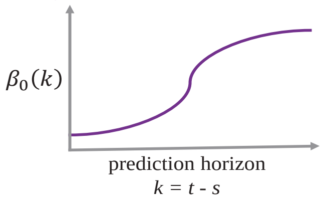
- Economic factor *s*<–*t*
- Behavioral covariate *t*<–*s*

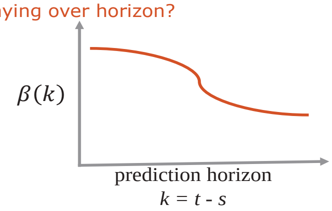
Is there effect from unobserved variables?

- e.g., baseline hazard in PD model

Is the sensitivity change over the horizon?

- e.g., is the effect of FICO at time snapshot decaying over horizon?





20

Joint age-by-horizon model: motivation

- Traditional credit risk models
  - Vintage model
    - Aggregates the account level panel data by vintages and then makes the forecast along the age dimension
    - Fails to utilize snapshot information (time at forecast) in prediction
  - Horizon model (account level)
    - Stacks the account level panel data into snapshot-forecast pairs
    - Often requires to sample the data into manageable size
  - Horizon model (pool level)
    - Aggregates the stacked snapshot-forecast pair data by the snapshot segments, and makes the forecast along the horizon dimension
    - Loses the age dynamics and vintage information in the modeling
- Newly proposed joint age-by-horizon model
  - Pool level time-varying coefficient model
  - Two-dimensional model – leverages both age and snapshot information

21

Joint age-by-horizon model: formulation

$$g\{E(Y_a(k)|X_a, Z_t)\} = f(a,k) + \alpha^T(k)X_a + \gamma^T(k)Z_t + \eta^T(k)O_v + \delta^T(k)X_aZ_t$$

- s = snapshot time; t = performance time; $k = t - s$ horizon
- a = vintage's age at s
- $Y_a(k)$ is the aggregated pool level response at performance age $m = a + k$
- X_a is a vector of vintage snapshot variables at snapshot age a
- Z_t is a vector of external macroeconomic variables at performance time t
- O_v is a vector of origination variables at vintage v
- $g(.)$ is a link function decided by the distribution of the response variable
- $f(a,k)$ is the two-dimensional baseline that vary with both snapshot ages and horizons
- $\alpha(k)$, $\gamma(k)$, $\eta(k)$ and $\delta(k)$ are the time-varying regression coefficient vectors

- **Model estimation**
  - **$\alpha(k)$, $\gamma(k)$, $\eta(k)$ and $\delta(k)$ can be modeled via regression splines or local polynomials.**
  - **$f(a,k)$ can be modelled with tensor product spline basis,**

$$f(a,k) \approx \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} f_{i,j}\phi_i(a)\psi_j(k).$$

  - **Lasso penalizes the likelihood with L_1 penalty,**

$$\hat{\theta} = argmin_{\psi}\{l(\theta) + \lambda\|\theta\|_1\},$$

Where θ is the vector of the coefficients of the base functions for $f(a,k)$, $\alpha(k)$, $\gamma(k)$, $\eta(k)$ and $\delta(k)$, and $\lambda \geq 0$ is a tuning parameter.

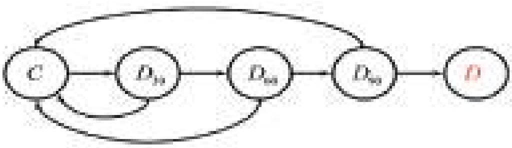
22

Insights from simulation and data analysis

- Joint age-by-horizon model captures the two-dimensional nature of the relationships well
- Account level horizon model yields generally similar results to joint age-by-horizon model in terms of accuracy and robustness
- Failure to consider age baseline and origination variables in the pool level horizon model causes bias in coefficient estimation
- The pool level vintage model usually generates noisier estimated coefficient curves over time, especially for old ages where sample sizes are smaller
- With coarser aggregation level, the coefficient estimation of the pool level vintage model and pool level horizon model is less robust and reliable than the joint age-by-horizon model

23

Credit cards or loans



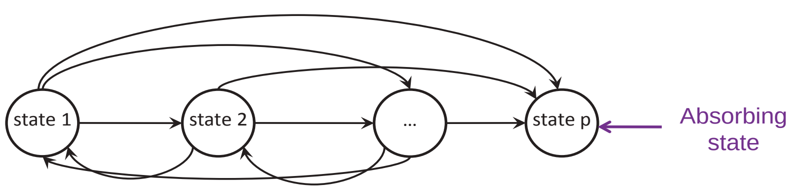
Status
Current, 0 day past due
D30, 1 cycle past due
D60, 2 cycles past due
D90, 3+ cycles past due
Attrition, closed with zero balance (option to re-open)
Default, 6+ cycles past due or in bankruptcy (termination)

Borrower quality rating in commercial credit

BQR	Description
1	Prime Borrowers
2	Near Prime Borrowers
3	Strong Borrowers
4	Acceptable Borrowers
5	Generally Acceptable Borrowers
6	Special Mention Borrowers
7	Substandard Borrowers
8	Doubtful Borrowers
9	Loss Borrowers

24

Multi-state models: framework



Subject or system can be in one of a finite number of states: $J(t) \in \{1, 2, \ldots, p\}$
Special case: Assumed here that one of the states (p) is absorbing

Initial state distribution π → system starts in state j with probability $\pi(j)$

Move among states according to transition probability matrix $P(s,t) = \{ p_{ij}(s,t) \}$

- Time nonhomogeneous and depends on path history in general
- Special cases: Markov, semi-Markov, hidden Markov,...

- Advantages of multi-state models over time-to-event models
- Challenges:
  - Need specification of transition structure – not robust
  - Non-Markov processes difficult to handle

25

Current work: testing for Markovian model and developing simple non-Markovian alternatives

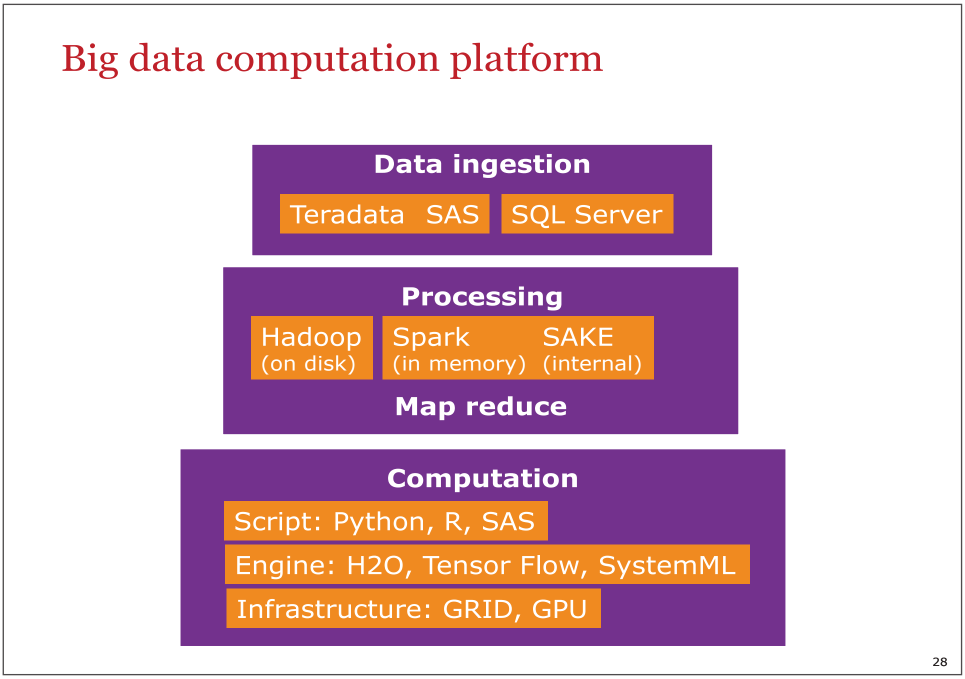
- Much of the finance literature assume Markov property
- Issues with non-Markov inference:
  - Transition matrix can explode in size
  - Prediction can be complex
- Our current work:
  - Testing for Markovian assumption
  - Developing model diagnostics
  - Simple non-Markovian alternatives

26

Desirable risk model characteristics in practice

- Simple/parsimonious
  - Should include as few factors and parameters as possible while still capturing key drivers
- Predictive
  - Should be forward looking, not just representative of the past
- Intuitive
  - Should include factors considered important by experts and react to changes in inputs as expected
- Transparent / understood
  - Structure and model behaviour including limitations is understood by users
- Stable
  - Should continue to perform if used in different circumstances or outside of development sample
- Computationally tractable
  - Should be able to deal with very large data with reasonable computation time

27


Big data computation platform
Data ingestion
Teradata SAS
SQL Server
Processing
Hadoop (on disk)
Spark (in memory)
SAKE (internal)
Map reduce
Computation
Script: Python, R, SAS
Engine: H2O, Tensor Flow, SystemML
Infrastructure: GRID, GPU
28

Fusion Learning:
Fusing Inferences from Diverse Sources for More Powerful Findings

Regina Y. Liu
Rutgers University, NJ, USA

Fusion Learning:
Fusing Inferences from Diverse Sources for More Powerful Findings

Regina Liu
Rutgers University
rliu@stat.rutgers.edu

ISI-RSC 2017, Bali, March 20-24, 2017

Fusion Learning:
Combine Inferences from Diverse Sources – Confidence Distribution (CD) Approach

Rutgers *CD BFF* group:
***Minge Xie,** William Strawderman, John Kolassa*
Joseph Chiarappa, Sifan Liu, Jieli Shen, Suzanne Thornton, Yilei Zhen
Dungang Liu, University of Cincinnati

Acknowledgement of research support from: *NSF, NSA, FAA*

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

4

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

6

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

e.g. **D&B –** global Consulting, has databases from companies (e.g.,*Target*, *Walmart*, *Sears*, …), provides forecasts/ratings of each company, for a sector, country, continent, …etc.

7

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

e.g. **D&B –** global Consulting, has databases from companies (e.g.,*Target*, *Walmart*, *Sears*, …), provides forecasts/ratings of each company, for a sector, country, continent, …etc.

*particularly suited for **Fusion** & ***i*Fusion** (more later)

8

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

e.g. **D&B** – global Consulting, has databases from companies (e.g.,*Target*, *Walmart*, *Sears*, …), provides forecasts/ratings of each company, for a sector, country, continent, …etc.

e.g. **Drug safety**: clinical trials from different lab (*FDA*)

e.g. **Regulatory issue**: aircraft performance/operations from different airlines (*FAA*)

… in almost all aspects of the real world!

9

Fusion Learning
Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

e.g. **D&B** – global Consulting, has databases from companies (e.g.,*Target*, *Walmart*, …), provides forecasts/ratings of each company, for a sector, country, continent, …etc.

e.g. **Drug safety**: clinical trials from different lab (FDA)

e.g. **Regulatory issue**: aircraft performance/operations from different airlines (FAA)

… in almost all aspects of the real world!

What do we gain from combining inferences

Validity (the 4th V) +

10

Fusion Learning

Combining Inference from Multiple Studies/Sources

Need: **Why**/Where/When/**How?**

- Information Explosion, Big Data (Volume, Velocity, Variety!)
  different databases, different data sources, different labs, …

e.g. **Big Data**: Storage/Computing – split & conquer, …

e.g. **D&B –** global Consulting, has databases from companies (e.g.,*Target, Walmart*, …), provides forecasts/ratings of each company, for a sector, country, continent, …etc.

e.g. **Drug safety**: clinical trials from different lab (FDA)

e.g. **Regulatory issue**: aircraft performance/operations from different airlines (FAA)

… in almost all aspects of the real world!

What do we gain from combining inferences

Validity (the 4th V) + Enhance/strengthen overall inference

11

Meta Analysis

"Statistical methods for combining and contrasting results from different studies, in the hope of identifying ***common patterns*** or ***sources of disagreement*** among those results, or ***other interesting relationships*** that may come to light in the context of multiple studies."

14

Meta Analysis (⊂Fusion Learning)

"Statistical methods for combining and contrasting results from different studies, in the hope of identifying ***common patterns*** or ***sources of disagreement*** among those results, or ***other interesting relationships*** that may come to light in the context of multiple studies."

15

Meta Analysis (⊂Fusion Learning)

"Statistical methods for combining and contrasting results from different studies, in the hope of identifying ***common patterns*** or ***sources of disagreement*** among those results, or ***other interesting relationships*** that may come to light in the context of multiple studies."

Reputation in the past –

- pseudo-science
- old fashioned
- "bad naming" – from Ingram Olkin (Stanford)

16

Meta-analysis cont.

Figure: Karl Pearson (1857∼1936)

Figure: Ronald Fisher (1890∼1962)

Meta-analysis cont.

Figure: Gene V. Glass (1940∼)

"my major interest currently is in what we have come to call ...the meta-analysis of research. The term is a bit grand, but it is precise and apt ... Meta-analysis refers to the **analysis of analyses***".*

Meta Analysis (⊂ Fusion Learning)

"Statistical methods for combining and contrasting results from different studies, in the hope of identifying ***common patterns*** or ***sources of disagreement*** among those results, or ***other interesting relationships*** that may come to light in the context of multiple studies."

Reputation in the past –

- pseudo-science
- old fashioned
- "bad naming" – from Ingram Olkin (Stanford)

Deserve a fresh new look -- take ownership of the subject – in light of

Data Explosion/Big Data/Automated data collections, …
(cf. CS, Math, Engineering, B-School, Data Science, …)

17

Introduction: Confidence Distribution (CD)

Statistical inference:

- Point estimate
- Interval estimate
- Distribution estimate (e.g., confidence distribution)

Example: $X_1, \ldots, X_n$ i.i.d. follows $N(\mu, 1)$

- Point estimate: $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Interval estimate: $(\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96/\sqrt{n})$
- Distribution estimate: $N(\bar{x}_n, \frac{1}{n})$

The idea of the CD approach is to use **a sample-dependent distribution (or density) function** to estimate the parameter of interest.

There are many ways of generating CDs
e.g., *P*-value function is a special case of a CD

1

Introduction to CDs

CD is very informative – Point estimators, confidence intervals, p-values & more

CD can provide meaningful answers for all questions in statistical inference –

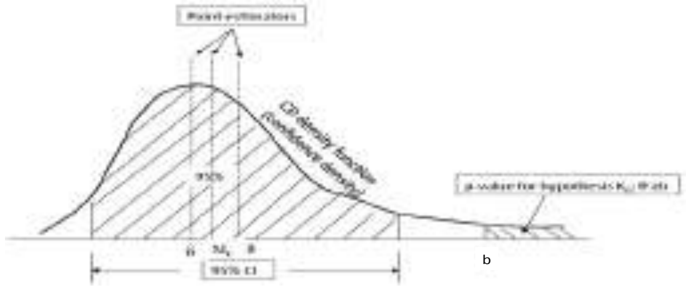


Figure 1. *The plot is a graphical illustration on making inference using a CD, including examples of point estimators (mode $\hat{\theta}$, median M_n and mean $\bar{\theta}$), a level 95% confidence interval and a one-sided p-value.*

Example 1: Normal CD function $N(\bar{x}, 1/n) \longrightarrow$ (i) Point estimator: $\bar{x}$; (ii) level $(1-\alpha)100\%$ CI: $\bar{x} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{n}$; (iii) p-value: $1 - \Phi(\sqrt{n}(b - \bar{x}))$, etc.

(cf., Xie & Singh 2013; Singh et al. 2007)

Introduction to CDs

Definition: Confidence Distribution

- **Definition**:
  A *confidence distribution* (CD) is a sample-dependent distribution function on parameter space that can represent confidence intervals (regions) of all levels for a parameter of interest.
  - Cox (2013, *Int. Stat. Rev.*): The CD approach is "to provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)."
  - Efron (2013, *Int. Stat. Rev.*): The CD development is "a grounding process" to help solve "perhaps the most important unresolved problem in statistical inference" on "the use of Bayes theorem in the absence of prior information."

Fusion learning through CDs

Fusion learning through confidence distributions

Key idea (steps)

- Summarize relevant data information using a confidence distribution in each study
- Synthesize information from diverse sources/studies via combination of the confidence distributions from these studies

Combining Independent CDs

$$p_c(\psi) = \Phi\left(\frac{1}{\sqrt{\sum_{i=1}^{K} w_i^2}}\left(w_1\Phi^{-1}(p_1(\psi)) + \cdots + w_K\Phi^{-1}(p_K(\psi))\right)\right)$$

Where p_i (.) is a *CD* function for the *i*-th study
and w_i is a combining weight

Note: Combining dependent *CDs* is also possible,
e.g. Design of temperature regulation system for an IBM data center – CDs are obtained in sequential spatial blocks

2

New (second) project - motivating example of spatial data

IBM data center thermal management project

- A data center is an integrated facility housing multiple-unit servers, providing application services or management for data processing.
- **Goal**: Design a data center with an efficient heat removal mechanism.
- Computational Fluid Dynamics (CFD) simulation ($n = 26820$; $p = 9$)

Figure : Heat map for IBM T. J. Watson Data Center

Q. ***Why not just buy a super computer and put all data in (i.e., data aggregation) to do statistical inference?***

A. May not necessarily be a good idea!

Fusion Learning is not the same as Data Aggregation!

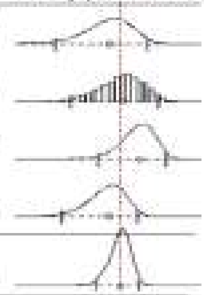
1

Numerical results

Illustrative Example I: Fusion learning across BFF paradigms (numerical results)

Table: Inference on correlation coefficient; four independent bivariate normal trials

Study & Method	95% Confidence Interval	CD density/bootstrap density/posterior distr.
Study-I: Fisher's Z method	[0.478,0.798]	
Study-II: Bootstrap BC_a	[0.543,0.836]	
Study-III: profile likelihood	[0.628,0.861]	
Study-IV: Bayesian/arc-sine prior	[0.499,0.762]	
Combination (via CDs)	[0.581,0.746]	



(*Thanks to Chengrui Li for the numerical results.*)

Numerical results

Illustrative Example I: Fusion learning across BFF paradigms (coverage)

Table : Inference on correlation coefficient; four independent bivariate normal trials; 1000 repetitions

	Coverage rate (1000 repetitions)	Mean length (& SD) of 95% CIs
Study-I: Fisher's Z method	0.943	0.294(0.0559)
Study-II: Bootstrap BC_a	0.948	0.293(0.0667)
Study-III: Profile likelihood	0.946	0.298(0.0533)
Study-IV: Bayesian/arc-sine prior	0.957	0.252(0.0474)
Combination (via CDs)	0.945	0.142(0.0224)

Introduction – Motivating examples

Illustrative Example II: iFusion with big data (motivating examples)

iFusion **- Individualized Fusion learning (by individual-to-clique)** tackles a special type of problems encountered in complex/heterogeneous/big data:

e.g. FINANCIAL FORECASTING (CONSULTING FOR DUN&BRADSTREET)
– Provide credit scores for each of 100,000^{+} companies

Q: Data collected for a single company may be limited, can we use companies of similar types to help improve inference?

e.g. PERSONALIZED/PRECISION MEDICINE.
Provide healthcare/treatment tailored to individual patients.

Q: Information on individual patient is often limited, can patients with similar traits help achieve better inference?

Goal: Make inference for an individual subject by borrowing information from similar subjects ('clique') to improve efficiency.

introduction – Methodology

Illustrative Example II: iFusion with big data

An iFusion approach following key steps:

S1 Obtain inference for each individual subject based on individual data (subset of data)

S2 Form a clique (or group) around the individual by the similarity of their share parameters or models

S3 Finally obtain inference or model from the clique

(Drawing inference from the clique allows borrowing strength from individuals with the same/similar parameters; thus enhance inference efficiency.)

introduction – Methodology

Illustrative Example II: iFusion with big data

An iFusion approach following key steps:

S1 Obtain inference for each individual subject based on individual data (subset of data)

S2 Form a clique (or group) around the individual by the similarity of their share parameters or models

S3 Finally obtain inference or model from the clique

(Drawing inference from the clique allows borrowing strength from individuals with the same/similar parameters; thus enhance inference efficiency.)

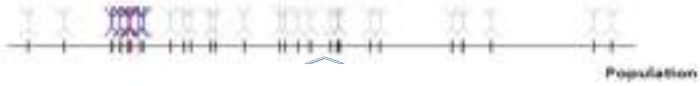
- **Bias-Variance tradeoff**
  - Inclusion: Similar individuals with small biases ("clique"); increased sample sizes reduce variance (improve efficiency)
  - Exclusion: Individuals with big biases (outside the "clique"); Reduced variance by increasing combined sample size can not overcome the big biases.

introduction – Methodology

Illustrative Example II: iFusion with big data

Red = Interest; Blue = Clique (similar ones)

- Personalized (precision) medicine/individualized inference



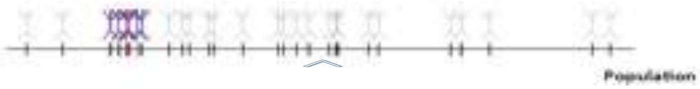
  ○ Bias-variance tradeoff

introduction – Methodology

Illustrative Example II: iFusion with big data

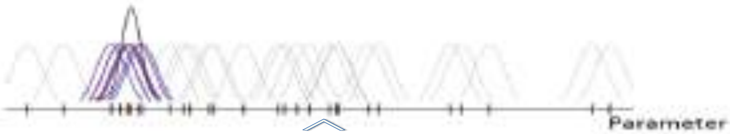
Red = Interest; Blue = Clique (similar ones)

- Personalized (precision) medicine/individualized inference



  - ○ Bias-variance tradeoff

- iFusion: Summarize ind. info. in CDs; form 'clique' (ties/near ties); combine info in clique – theoretical sound; division of labor



  - ○ Computationally feasible for big data (vs. conv. hierarchical/mixture/non-parametric Bayesian methods)

Proof of concept (simulation)

Illustrative Example II: iFusion with big data

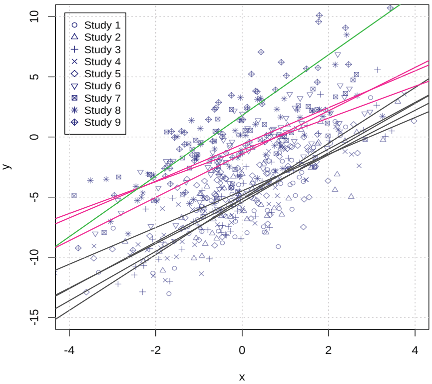
Simulation study/Proof of concept: Regression model -

$$y_{ik} = \beta_{0,k} + \beta_{1,k} x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, 2^2), \quad i = 1, \ldots, 60, \quad k = 1, \ldots, 9$$

True parameter values

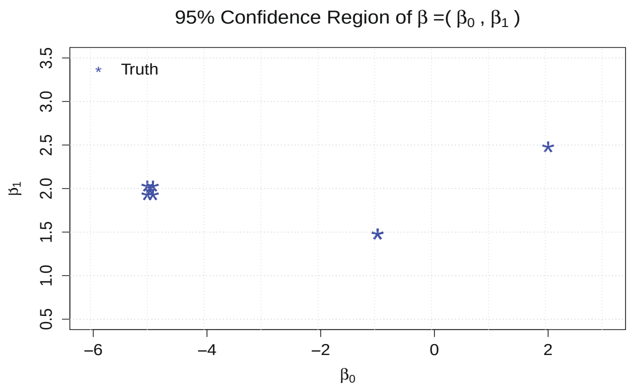
Study	β_0	β_1	n_k
1	-5.02	2	60
2	-4.98	2.02	60
3	-4.98	1.98	60
4	-5.02	2.02	60
5	-5.02	1.98	60
6	-1	1.5	60
7	-1	1.5	60
8	-1	1.5	60
9	2.0	2.5	60

Simulated data with individually fitted lines



Proof of concept (simulation)

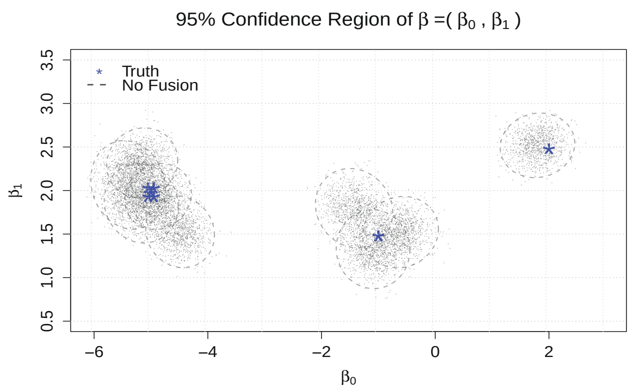
Illustrative Example II: iFusion with big data



Typical outcomes based on a simulated dataset.

Proof of concept (simulation)

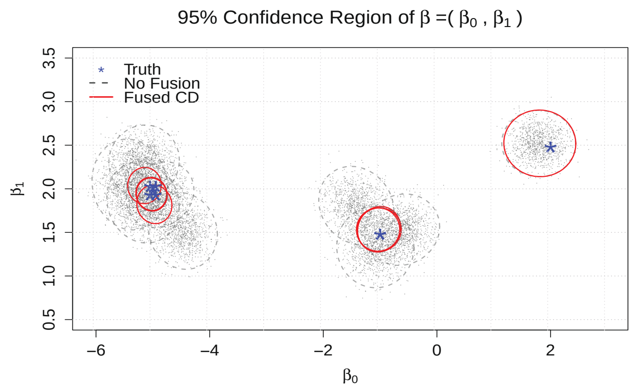
Illustrative Example II: iFusion with big data



Typical outcomes based on a simulated dataset.

Proof of concept (simulation)

Illustrative Example II: iFusion with big data



Typical outcomes based on a simulated dataset.

Proof of concept (simulation)

Illustrative Example II: iFusion with big data

Parameter	MSE		Coverage		Median Width	
	Indiv	Fusion	Indiv	Fusion	Indiv	Fusion
$\beta_{1,1}$	0.040	0.008	0.956	0.957	0.807	0.316
$\beta_{1,2}$	0.037	0.009	0.951	0.950	0.754	0.317
$\beta_{1,3}$	0.032	0.010	0.950	0.935	0.673	0.354
$\beta_{1,4}$	0.029	0.009	0.953	0.950	0.660	0.316
$\beta_{1,5}$	0.033	0.011	0.955	0.946	0.708	0.359
$\beta_{1,6}$	0.031	0.011	0.947	0.944	0.695	0.381
$\beta_{1,7}$	0.032	0.011	0.954	0.944	0.707	0.381
$\beta_{1,8}$	0.025	0.011	0.963	0.944	0.621	0.381
$\beta_{1,9}$	0.024	0.024	0.956	0.956	0.600	0.600

Long-run performance/1000 repetitions: $\text{MSE}_{1,k} = \text{E}_{1,k}(\hat{\beta}_{1,k} - \beta_{1,k})^2$ and 95% nominal confidence intervals. Similar results for intercepts β_0's (omitted here)

Methodology and Theory

Illustrative Example II: iFusion with big data

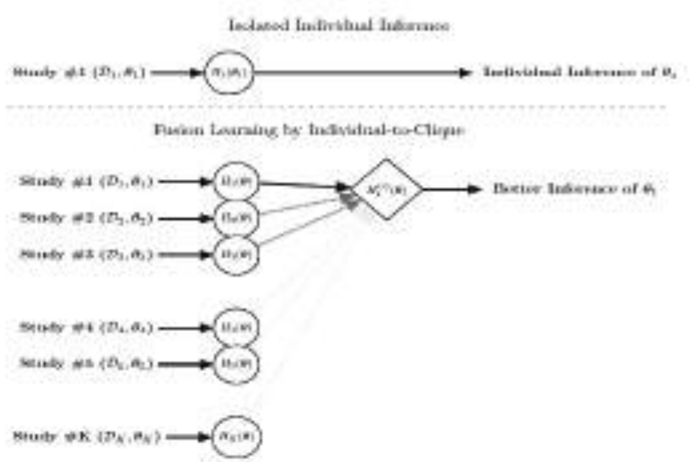
Desirable features

- Inference efficiency
  - "Borrowing strength" from relevant studies;
  - Oracle performance under mild conditions
- Parameter assumption free
  - No assumption on underlying individual parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$; Any of the $\boldsymbol{\theta}_i$'s may or may not equal to some others.
- Computational feasibility
  - "Divide, conquer, and combine"; scalable to big data.
- Generality and flexibility
  - **Confidence distribution** - a new, versatile, flexible and effective inference concept/tool.

Methodology and Theory

Illustrative Example II: iFusion with big data

Making inference on θ_1 - a diagram:



Methodology and Theory

Illustrative Example II: iFusion with big data

Theorem 1

Suppose $w_{1,k}^{(s)}$ satisfies

$$w_{1,k}^{(s)} = \begin{cases} c_{1,k} + o_p(n^{-1/2}) & \text{if } \boldsymbol{\theta}_k \in \mathcal{C}_1; \\ o_p(n^{-1/2}) & \text{otherwise.} \end{cases} \tag{1}$$

for some constant $c_{1,k} \in (0, 1]$. Then,

(i) $H_1^{(c)}(\boldsymbol{\theta})$ by iFusion is a valid CD for $\boldsymbol{\theta}_1$.

(ii) If further $c_{1,k} = 1$, then $H_1^{(c)}(\boldsymbol{\theta})$ by iFusion provides oracle inference, asymptotically.

Translation:

(i) Inference by iFusion is valid.

(ii) Inference by iFusion is asymptotically equivalent to the "best" (oracle).

i-Fusion has been applied to 3 projects:

- **D&B project** – proprietary data, good results but can not be presented here
- **Mine Detection project** –
- **Astrophysics project** -- Reconstructing Star Formation Histories to Reveal the Origin and Evolution of Galaxies

3

Fusion Learning:
Combine Inferences from Diverse Sources – CD Approach

Added values –

CD provides solutions for problems whose solutions are previously unavailable, unknown or not easily available

Fusion Learning:
– CD Approach

Added values – CD provides solutions for problems whose solutions are previously unavailable, unknown or not easily available

Combining Inferences from , e.g.,

- ***Inferences derived from different approaches/tests or paradigms (Bayesian, frequentist, fiducial, …)***
- ***Binary Experiments (with rare events, e.g., zero-cells)***
- ***Heterogeneous Studies***
- ***iFusion: individualized Fusion (Learning by Individual-to-Clique – similar to personalized medicine, …)***

Confidence Distribution (CD)

- **CD is a 100% frequentist approach**
- **CD can potentially provide a unifying framework under which Bayesian, fiducial and frequentist (BFF) inferences can be directly connected, compared, and combined.**

BFF: Bayesian, fiducial and frequentist

$\Longrightarrow$ ***BFF: Best Friends Forever (possible?)***

4th BFF Workshop – ***May 1-3, Boston, Harvard***

Variable Selection Techniques for Analyzing Huge-Dimensional Datasets

Naveen N. Narisetty
University of Illinois at Urbana-Champaign, USA

Variable Selection Techniques for Analyzing ~~Large~~ Huge-Dimensional Datasets

Naveen N. Narisetty

Department of Statistics,
University of Illinois at Urbana-Champaign

ISI RSC 2017, Bali, Indonesia

Outline

1. Introduction
   - High Dimensional Data
   - Variable Selection
2. Penalty Based Methods
   - Convex Penalties
   - Non-Convex Penalties
3. Bayesian Methods
   - Methodology
   - Computation

Datasets with Huge Dimensions

- In the era of Big Data, many datasets have a "huge" number of features (variables) per each observation
- These are called High-Dimensional or Large-Dimensional Data
- Many biological, economic, financial, and web applications generate such data
- Examples include high-frequency financial data (huge no. of time points), stock market data (huge no. of stocks), gene expression data (huge no. of genes), etc.

Stock Price Data

- How do the prices (or returns) of different stocks in the S&P 500 index depend on each other? Prices (or returns) of which stocks are associated with which others?
- To study this even for the 505 stocks in the S&P 500 index, there are around 128,000 possible pairwise associations
- Only 260 observations if we use 5-year weekly return data!

Gene Expression Data

- Understanding which genes have influence on a phenotype is a fundamentally important problem in biomedical research
- In most gene expression data, data on only a few subjects are available; but on many genes per subject
- **Example: PCR data.** gene expressions of $n = 60$ subjects and $p = 22,575$ genes per subject.

  Outcomes of interest: different phenotypes related to obesity

Example Discoveries

Variable Selection

- Finds a small subset of variables (or features) that are related to the outcome of interest
- A very commonly used statistical tool in practice
- Parsimonious models help for better prediction and estimation
- A classical (and **evergreen**) problem in statistics with both computational and theoretical challenges: 2^p possible models (combination of variables) to choose from

Classical Methods

- When p is small, classical approaches such as best subset selection, information criterion based methods, stepwise methods such as backward elimination are feasible
- For huge p, these approaches are either computationally infeasible or lack statistical properties!
- Huge number of publications on this topic

  (Google Scholar finds 30,000 entries since 2016 with "model selection" as keyword and 16,000 with "variable selection")

Outline

High Dimensional Linear Regression

- Consider the standard linear regression set-up

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

n - no. of observations
p - no. of features/variables

- When $p > n$, none of the standard techniques work as the problem is ill-posed
- Sparsity Assumption: only a few features are important

(Popular) Penalty Trick

- Maximum likelihood minimizes the negative log likelihood. That is,

$$\hat{\beta}_{MLE} = \arg\min_{\beta} \{-\log L(\beta)\} = \arg\min_{\beta} \|Y - X\beta\|^2.$$

- The penalty trick is to minimize it together with a penalty function $\rho_\lambda(\beta)$, to induce sparsity.

$$\hat{\beta}_{PEN} = \arg\min_{\beta} \{-\log L(\beta) + \rho_\lambda(\beta)\}. \quad (1)$$

L_0 Penalty

- A very natural penalty is the L_0 penalty

$$\rho_\lambda(\beta) = \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}.$$

- Several authors discussed the suitability of L_0 penalty for variable selection (Schwarz, 1978; Shen, Pan and Zhu, 2012).
- However, $\rho_\lambda(\beta)$ is non-convex and the optimization is NP-hard. Motivated by this, convex relaxations of this penalty are used in practice.

Convex Penalty Methods

- The well-known Lasso (Tibshirani, 1996) method uses L_1 penalty:

$$\rho_\lambda(\beta) = \lambda\|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Adaptive Lasso (Zou, 2006) uses a scaled L_1 penalty:

$$\rho_\lambda(\beta) = \lambda \left\| \frac{\beta}{\hat{\beta}} \right\|_1,$$

where $\hat{\beta}$ is a consistent estimator of β.

- Elastic Net (Zou and Hastie, 2005) stabilizes the correlations between predictors:

$$\rho_\lambda(\beta) = \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right),$$

for some $\alpha \in [0, 1]$.

Introduction | Penalty Based Methods | Bayesian Methods

Convex Penalty Methods

- The well-known Lasso (Tibshirani, 1996) method uses L_1 penalty:

$$\rho_\lambda(\beta) = \lambda\|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Adaptive Lasso (Zou, 2006) uses a scaled L_1 penalty:

$$\rho_\lambda(\beta) = \lambda \left\| \frac{\beta}{\hat{\beta}} \right\|_1,$$

where $\hat{\beta}$ is a consistent estimator of β.

- Elastic Net (Zou and Hastie, 2005) stabilizes the correlations between predictors:

$$\rho_\lambda(\beta) = \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right),$$

for some $\alpha \in [0, 1]$.

13/32
N. N. Narisetty | Variable Selection: A Review

Convex Penalty Methods

- The well-known Lasso (Tibshirani, 1996) method uses L_1 penalty:

$$\rho_\lambda(\beta) = \lambda\|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Adaptive Lasso (Zou, 2006) uses a scaled L_1 penalty:

$$\rho_\lambda(\beta) = \lambda \left\| \frac{\beta}{\hat{\beta}} \right\|_1,$$

where $\hat{\beta}$ is a consistent estimator of β.

- Elastic Net (Zou and Hastie, 2005) stabilizes the correlations between predictors:

$$\rho_\lambda(\beta) = \lambda \left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \right),$$

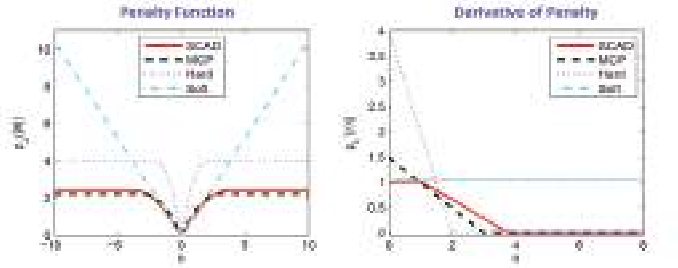
for some $\alpha \in [0, 1]$.

Pros and Cons of Convex Penalties

- Computationally appealing. Global minimizers can be obtained using efficient algorithms (Efron et al., 2004, Friedman et. al, 2009)
- Standard packages available for implementation (Lars, glmnet, elasticnet, ... in R)
- However, performance for variable selection is compromised. This is because these penalties are far from L_0
- For convex penalties to work well, correlations between variables needs to be very very small (Zhao and Yu, 2006)

Non-Convex Penalties

- SCAD (Fan and Li, 2001)
- MCP (Zhang, 2010)



Some commonly used penalty functions (left panel) and their derivatives (right panel)

(Figure from Fan and Lv, 2010)

Pros and Cons of Non-Convex Penalty

- Variable selection performance in general better than convex penalty methods
- Package ncvreg in R for implementation
- Computationally challenging. Local Linear or Local Quadratic Approximations are commonly used.
- Potential issues with convergence of standard algorithms. No guarantee of reaching global minimum.
- Promising advances made by Breheney and Huang (2011), Gasso et al. (2008), Mazumder et al. (2011), She (2009).

Tuning Parameter Selection

- A key component of penalty based methods is the choice of the tuning parameter λ
- Many algorithms provide solutions for a sequence of λ values
- Several approahces available in the literature to choose an optimal λ: Cross Validation, Information Criterion (AIC/BIC/EBIC/etc.),
- In general BIC/EBIC provide sparser model whereas CV provides better prediction

Outline

Bayesian Methods for Variable Selection

- Bayesian methods have also been popular in practice for various reasons; Google recommended :)
- Bayesian model selection can be asymptotically equivalent to information criterion
- The classical arguments of inducing belief and obtaining uncertainty estimation stay valid but have limitations under huge dimensions
- Provides alternative computational and/or shrinkage strategies

A Unified Bayesian Framework

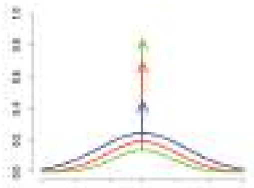
- Same linear regression model: $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$
- A Bayesian approach typically introduces binary variables Z_j to indicate $\beta_j \neq 0$.
- Priors on β_j (for $j = 1, \cdots, p$) are placed as

$$\beta_j \mid Z_j = 0 \sim \pi_0(\beta_j), \quad \beta_j \mid Z_j = 1 \sim \pi_1(\beta_j) \qquad (2)$$

- Once prior probabilities for $P[Z_j = 1]$ are assigned, the posterior of Z, $P[Z \mid Data]$, is used for variable selection.

Prior Specification

- The prior distributions $\pi_1(\cdot)$, $\pi_0(\cdot)$, and the probabilities $P[Z=1]$ determine a Bayesian variable selection method
- Typically, the ability to choose these distributions provide a greater flexibility than penalty based methods
- However, it comes at a cost of having to choose the hyper-parameters of these distributions.



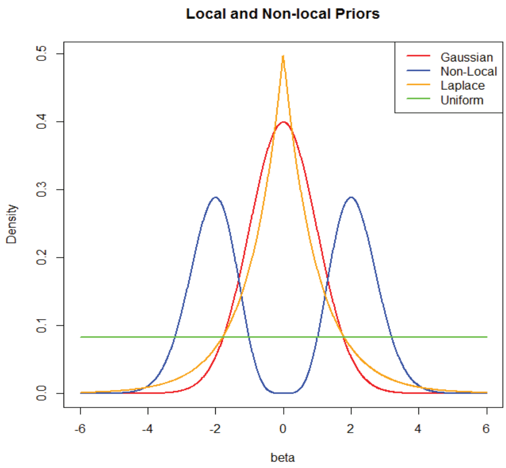
Spike and Slab Priors

- The prior π_0 (called spike) is often taken to be a point mass

  (Geweke, 1996; Scott and Berger, 2010; Liang, Song, and Wu, 2013; Yang, Jordan and Wainwright, 2015)

- Some common choices for the prior π_1 (called slab) include:
  - g priors (Zellner, 1983; Som, Hans, MacEchern, 2014)
  - Uniform (Mitchell and Beauchamp, 1988)
  - Gaussian (George and McCulloch, 1993)
  - Laplace (Park and Casella, 2008)
  - Non-local priors (Johnson and Rossell, 2012), & more ...

Different Slab Priors



Introduction | Penalty Based Methods | Bayesian Methods

Bayesian Approach and Penalization

- There is an interesting connection between the Bayesian approach and penalization
- A penalization method can be viewed as the maximum a posteriori model corresponding to a specific prior
- For example, Laplace slab prior will give raise to Lasso as the posterior mode
- Possible to think of the priors as a way to induce a shrinkage sparsity (as opposed to inducing belief)

Bayesian Computation

- Computation in the Bayesian paradigm often involves stochastic algorithms to sample from the posterior distribution
- Commonly used approaches are different types Markov Chain Monte Carlo (MCMC) algorithms including Metropolis-Hastings, Gibbs sampling, Slice Sampling, Reversible Jump MCMC, etc.
- An advantage is that non-convexity does not affect sampling algorithms as much as optimization algorithms
- However, they can be computationally intensive for large p problems. Very active area to develop efficient algorithms!

Existing Approaches - Point Mass Spike

- Optimization based Methods
  - EM Algorithm (Ročková and George, 2014)
  - Approximate L_0 Optimization (Bertsimas, King, and Mazumder; Cassidy and Solo, 2015)
- Metropolis-Hastings Random Walk Methods
  - Shotgun Stochastic Search (Hans, Dobra, West, 2007)
  - Bayesian Subset Regression (Liang, Song, and Yu, 2013)
- Laplace Approx. based Algorithms (Shun and McCullagh, 1995; Johnson and Rossell, 2012; Barber et al., 2015)

Pro: Point mass priors give good theoretical properties!
Con: Either computationally costly or approximation algorithms!

Existing Approaches - Point Mass Spike

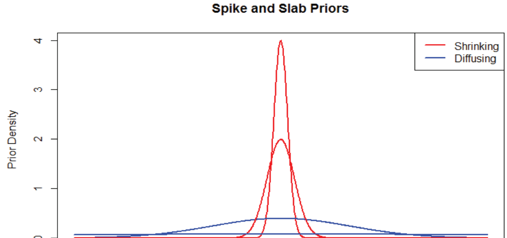
- Optimization based Methods
  - EM Algorithm (Ročková and George, 2014)
  - Approximate L_0 Optimization (Bertsimas, King, and Mazumder; Cassidy and Solo, 2015)
- Metropolis-Hastings Random Walk Methods
  - Shotgun Stochastic Search (Hans, Dobra, West, 2007)
  - Bayesian Subset Regression (Liang, Song, and Yu, 2013)
- Laplace Approx. based Algorithms (Shun and McCullagh, 1995; Johnson and Rossell, 2012; Barber et al., 2015)

Pro: Point mass priors give good theoretical properties!
Con: Either computationally costly or approximation algorithms!

Gaussian Spike and Slab Priors

George and McCulloch (1993)

- Spike: $\pi_0 \equiv N(0, \tau_0^2)$, the variance τ_0^2 is small
- Slab: $\pi_1 \equiv N(0, \tau_1^2)$, the variance τ_1^2 is fixed at a larger value
- This allowed the use of a Gibbs sampler for posterior sampling
- With appropriately chosen parameters, theoretical properties hold even for huge p! (Narisetty and He, 2014)



Posterior Computation

- Standard Gibbs Sampler: conditional distributions are
  - Z_j (given β): Independent Bernoulli
  - β (given Z): Multivariate Normal: $N\left(VX'Y, \sigma^2 V\right)$
- Easy to implement but not easily scalable for large p; involves p dimensional matrix computations
- Common difficulty for many Gibbs sampling algorithms (Ishwaran and Rao, 2005; 2010; Bhattacharya et.al, 2015)

Can we have a scalable Gibbs sampler?

Posterior Computation

- Standard Gibbs Sampler: conditional distributions are
  - Z_j (given β): Independent Bernoulli
  - β (given Z): Multivariate Normal: $N\left(VX'Y, \sigma^2 V\right)$
- Easy to implement but not easily scalable for large p; involves p dimensional matrix computations
- Common difficulty for many Gibbs sampling algorithms (Ishwaran and Rao, 2005; 2010; Bhattacharya et.al, 2015)

Can we have a scalable Gibbs sampler?

Posterior Computation

- Standard Gibbs Sampler: conditional distributions are
  - Z_j (given β): Independent Bernoulli
  - β (given Z): Multivariate Normal: $N\left(VX'Y, \sigma^2 V\right)$
- Easy to implement but not easily scalable for large p; involves p dimensional matrix computations
- Common difficulty for many Gibbs sampling algorithms (Ishwaran and Rao, 2005; 2010; Bhattacharya et.al, 2015)

Can we have a scalable Gibbs sampler?

Skinny Gibbs Sampler

- Sparsify the covariance matrix of β
- That is, use the covariance matrix S in place of V:

$$S = \begin{pmatrix} (X_A' X_A + \tau_{1n}^{-2}\,\mathrm{I})^{-1} & 0 \\ 0 & (n + \tau_{0n}^{-2})^{-1}\,\mathrm{I} \end{pmatrix}$$

- Sample Z (given β) from conditional Bernoulli draws
- This makes the computations linear in number of variables! Huge improvement!

Features of Skinny Gibbs

- Extends readily to non-linear models including logistic regression, SVM, quantile regression, etc.
- Continues to work in non-convex settings unlike optimization based methods
- Computationally scalable without sacrificing strong statistical properties (Narisetty, Shen and He, 2017+)
- Mixing properties of Skinny Gibbs and other high dimensional Bayesian algorithms require future work!

Implementation of Bayesian Methods

- Several packages available in R to implement Bayesian variable selection methods
- For non-local priors: mombf; Gaussian spike and slab: spikeslab; for skinny Gibbs: skinnygibbs (under development and available on request)
- The software Stan can be used to implement gradient-based MCMC and variational Bayes algorithms
- WinBUGS is a specialized software useful for Bayesian computation with MCMC

Screening Methods

- When p is really really really REALLY HUGE, none of the methods we discussed may be feasible.
- A first step of screening is needed in those cases
- Several promising approaches available: Sure Independent Screening (Fan, 2007, 2010), Model free Screening (Zhue et al., 2011), Quantile based Screening (He et al., 2013), MDD based Screening (Zhao et al. 2017)
- R package SIS can be used!

Open Questions

- There have been exceptional advances made in the last two decades in terms of methodology, computation and theory
- There are many open challenges especially for better computation using non-convex penalties and Bayesian approaches
- There is a huge need to effectively communicate the developments to practitioners across disciplines

Thank You!

This page is intentionally left blank

IPS13:
Payment System Data and Leading Indicators

Recent Developments in Payment Systems

Christian Dembiermont*
Bank for International Settlements, Basel, Switzerland – Christian.Dembiermont@bis.org

Abstract

Payment, clearing and settlement systems are a core element of the financial infrastructure at the national and international level.

The Committee on Payments and Market Infrastructures (CPMI) periodically publishes reference works on payment, clearing and settlement systems in the CPMI member countries. These reports are widely known as Red Books. The statistics include indicators of retail payment systems and payment instruments, as well as of wholesale systems used among banks. Moreover, they include data on trading platforms, clearing houses and settlement systems for securities, as well as on the systems used to perform cross-border transactions. All data are national data collected by the central banks participating in the exercise.

Keywords: Payment systems, Clearing, Settlement systems, Central bank's statistics.

1 For examples of this public debate refer to Bindseil et al. (2015)

Money Talks – Nowcasting Real Economic Activity with Payment Systems Data

Luís Teles Dias
Banco de Portugal, Lisbon, Portugal – ldias@bportugal.pt

Abstract

Payment systems play a central role in the functioning of modern market economies, by enabling the transfer of money and financial instruments between economic agents in a safe and efficient way. In this context, researchers have devoted a great deal of attention to the assessment of the spillover effects that the development of payment systems could induce in an economy and on the demand for currency. More recently, the usage given to the data generated by these systems have been considerably broadened, to encompass issues as, *inter alia*, the assessment of financial integration and nowcasting private consumption. In this paper, we explore possible uses of payment systems data in the specification of coincident and/or leading indicators for key macroeconomic aggregates – such as the gross domestic product and private consumption of households. Given the strong connection between the phenomena underlying these data and the above-mentioned macroeconomic variables, we highlight the comparative advantages of this information in relation to competing indicators, which is based on its inherent quality, competitive costs and high frequency. In addition, we present the payment systems data available in Portugal, with particular emphasis on Automated Teller Machine and Point of Sale driven data, and describe how the current institutional environment and the recent enhancements in the reported information have gained further importance and relevance in meeting *Banco de Portugal's* main goals.

Keywords: GDP forecasting; Private consumption forecasting; Nowcasting; Leading indicators

Statistics to Support The Smooth Operation of Payment Systems in The European Union

Rodrigo Oliveira Soares* and Hanna Häkkinen
European Central Bank, Frankfurt, Germany - Rodrigo.Oliveira@ecb.int, Hanna.Hakkinen@ecb.int

Abstract

Statistics on payments are collected from service providers and payment systems operators. These data are important for monitoring the payment systems activity and for research and further development of payment systems. The statistics also provide an overview of the whole industry, which can be used by reporting entities in their strategy and planning processes.

The single euro payments area (SEPA) and the entry into the market of new service providers under the EU Directive have changed retail payments in Europe ensuring a higher degree of integration, e.g. a lower and more homogeneous pricing, and more value and consumer protection for households. These changes also had implications in payments statistics, which in 2013 were brought within the scope of a regulation (ECB/2013/43) to enhance the coverage and foster a better comparability across countries.

This paper will review payment statistics in the European Union, namely as regards payment services, retail and large-value systems.

Keywords: Euro, Payment services, Retail payments, Large-value payments.

JEL classification: E4

Payment System Statistics to Support Policy Formulation in Indonesia

Farida Peranginangin[1]
Bank Indonesia, Indonesia – faridap@bi.go.id

Abstract

Bank Indonesia sets monetary, financial system stability, and payment system policy in accordance with its mandate of achieving and maintaining the stability of the Rupiah value. In Indonesia's payment system, Bank Indonesia operates high value payment system–Bank Indonesia Real Time Gross Settlement (BI-RTGS)–and regulates retail payment systems that are operated by industry: card-based payment instruments (ATM-Debit and credit cards) and electronic money. In conducting payment system policy, Bank Indonesia synchronously considers the monetary and financial system stability policymaking process. In this framework, the role of payment system is to support the monetary policy transmision and the stability of financial system. In order to enhance the robustness of policy formulation, Bank Indonesia continuously performs in-depth analysis of data (research-based policy), including data that is produced through payment system. In this session, we will present several use case of payment system statistics for payment and settlement systems monitoring and strengthen Bank Indonesia's policy formulation on payment system as well as in monetary and financial system stability

Keywords: Statistical simulation, Retail payment, Economic growth, Financial development

1. INTRODUCTION

As the authority of payment system in Indonesia, Bank Indonesia strives for a safe, efficient, smooth and reliable payment system with considerations of expanding access and protecting consumer. In Indonesia's payment system, Bank Indonesia functions both as the operator of high value dan systemically important payment system known as BI-Real Time Gross Settlement (BI-RTGS) system and as regulator of various industry-operated retail payment systems, including ATM-Debit card, credit card, and electronic money. Thus, Bank Indonesia is in the most pertinent position to harness the economic insights from payment and settlement system data in Indonesia. These statistics are used for enhancing the robustness of Bank Indonesia's policy formulation, both in payment system policy or monetary and financial system policy.

In order to achieve its mandated objective–the stability of the Rupiah–Bank Indonesia harmonizes monetary, financial system stability and payment system policies. In this framework, payment system has a role in ensuring the transmission of monetary policy and support financial system stability. Bank Indonesia continuously performs in-depth analysis of data in order to enhance the robustnessof its policy formulation, particularly but not limited to payment system. The data are collected monthly from Bank Indonesia-operated payment system (BI-RTGS) and industry players-operated payment systems (card-based payment and electronic money). The statistics of BI-RTGS serves as an input bank's liquidity monitoring and in monetary and financial system stability policymaking. Bank Indonesia also uses figures from retail payment system in fostering an efficient transaction cost and promoting consumer protection. Bank Indonesia has also formulated the Payment System Liquidity Index (PSLI) as a measure of general liquidity conditions in the payment system by utilizing statistics of BI-RTGS settlement data. Due to PSLI's characteristic as a composite index that is constructed from BI-RTGS settlement data, it is suitable to indicate the condition of banking liquidity in the payment system.

2. PAYMENT SYSTEM DEVELOPMENT IN INDONESIA

BI-RTGS settlement statistics consists of settlement data of BI-RTGS's participants that includes all conventional and sharia banks in Indonesia. Currently there are 141 BI-RTGS participants, 136 of these

1 Director of Payment System Policy and Regulation, Bank Indonesia

are the aforementioned banks with 69% of Rp477.9 trillion daily transaction nominal (value) and 95% of 37.456 daily transaction volume. As a consequence, the banking industry dominates the payments system liquidity in BI-RTGS and this figure could serve as a proxy of banking liquidity conditions in Indonesia. BI-RTGS statistics is available at T+1 both in nominal and volume for all individuals BI-RTGS participants with each participant transaction's detail. The availability of BI-RTGS data in minute detail allows Bank Indonesia to conduct banking liquidity monitoring in a timely manner. In addition to analyzing the liquidity conditions of banking industry, the availability of individual bank's statistics in BI-RTGS allows Bank Indonesia to analyze the statistics both in terms banking group/ classification as well as individual bank. Furthermore, BI-RTGS statistics can be broken down to the type of the transactions, such as government transactions, customer transactions, interbank money market transactions, etc. The abundance of high frequency data in BI-RTGS with a lot of variations and low data lag, prompts Bank Indonesia to employ various payment system data analyses as part of its policy formulation in payment system as well as macroeconomic and financial system stability.

In terms of retail payment systems, BI manages statistics of card-based payment instruments which includes ATM Debit card, credit cards and electronic money. These statistics includes statistics of payment instrument providers, the number of outstanding payment instruments, and transaction statistics both in volume and nominal per providers/issuers of the payment instrument. These statistics are presented on a monthly basis and are obtained from bank or non bank. The data is reported online through Commercial Bank Head Office Report (LKPBU) and Non Commercial Bank Report (CB and NCB Report) application with a lag of one month. The analysis from retail payment statistics is used for policy formulation in payment system area in order to increase the efficiency of the economy by promoting consumer protection. In addition, the analysis is also used to fortify Bank Indonesia's policy mix formulation on monetary and financial system stability.

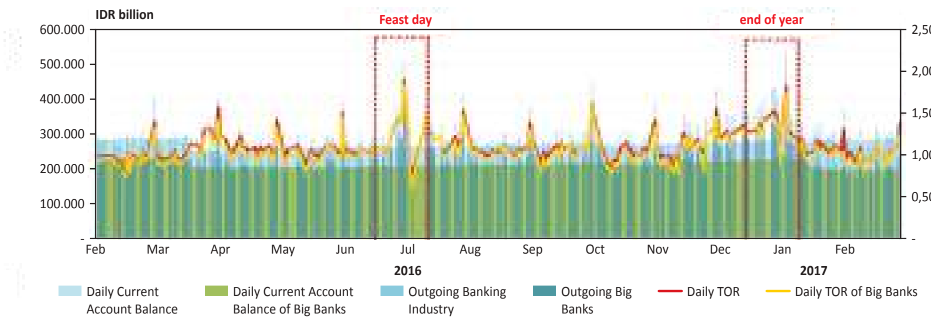
3. PAYMENT SYSTEM STATISTICS AS A TOOL FOR FINANCIAL MARKET RISK MONITORING

The availability of BI-RTGS statistics in a daily basis allows the use of these statistics for either weekly or monthly monitoring. The indicators used for monitoring are Turn Over Ratio (TOR), throughput zone III, and interconnectedness which formulated as follows:

Table 1. Indicator of Payment System for Monitoring

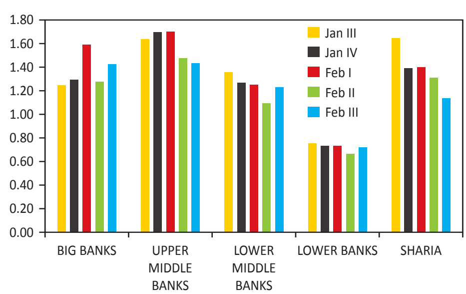
Indicator	Formula	Interpretation
Turn Over Ratio	$TOR_i = \frac{\Sigma \text{Outgoing transaction}}{\text{Current account balance}}$ $TOR_{industry} = \frac{\Sigma^n_{i=1} \text{Outgoing transaction}}{\Sigma^n_{i=1} \text{Current account balance}}$	TOR ≤1 : settlement obligation can be settled only with current accounts balance in Bank Indonesia. TOR > 1 : settlement obligation can be settled with incoming funds from other participants.
Throughput zone III	$\frac{\Sigma^n_{i=14.30} \text{Outgoing transaction}}{\Sigma^n_{t=0} \text{Outgoing transaction}}$	The share of transactions settled during 3rd BI-RTGS window settlement. Bank Indonesia's Guideline: 30% in zone I (system open-10.00 AM), 30% in zone II (10.00 AM – 2.00 PM), and max 40% (zone III: 2:00 PM-closed).
Interconnectedness	$\frac{\Sigma^n_i \Sigma^n_j X_{ij}}{n^2 - n}$ $X_{ij} = 1$, *if there is a transfer of funds from bank i to bank j,i ≠ j* *n = number of banks*	The payment system relationship between the participants in the BI-RTGS. Interconnectedness = 100% is reached when all RTGS participants doing transactions with all of other participants.

Bank Indonesia monitors the payment system liquidity indicators in both weekly and monthly basis with two stream of analysis: banking industry ownership group and banking industry BUKU group. Analysis on the indicators is conducted by reviewing the historical pattern, such as the increase of banking industry's TOR around feast day (Ramadhan), holiday and end of year which is a common phenomenon due to increasing payment system activities by customers and government (Figure 1). Bank Indonesia observes whether the significant increasing of TOR beyond this seasonality period due to its impact on the smooth operation of payment systems. If this shift occurs continuously within a long time period, the result of this analysis is delivered as input in macroeconomic and financial system policymaking mix.



Source: EDW, BI-RTGS

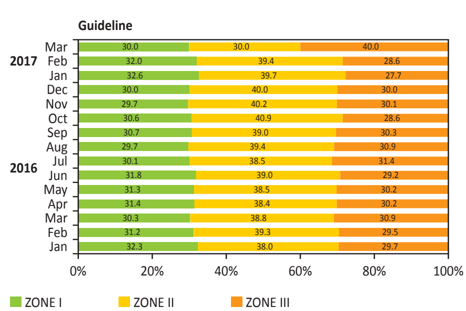
Figure 1. Turn Over Ratio (Industry)



Source: EDW, BI-RTGS

Figure 2. Turn Over Ratio (Group of Bank)

TOR monitoring is also applied on group of banks to see whether there is a change on banks' behavioral pattern on their liquidity management in BI-RTGS. Historically, groups of BUKU 4 has the lowest TOR (TOR<1), which means that settlement obligations of this group of banks can be settled only by using their current account balance in Bank Indonesia. Meanwhile, other groups of banks have TOR>1, means that their settlement obligations can be settled only if there exists incoming fund transfer since their current account balance in Bank Indonesia is not sufficient to cover their settlement obligations (Figure 2). The significant increase on the TOR of group of BUKU 4, especially when it exceeds 1,



Source: EDW, BI-RTGS

Figure 3. Throughput

should be closely evaluated since these banks cover 70% of liquidity in the payment system which could impact the payment system liquidity of the rest of banking industry. Monitoring on TOR indicator can be applied daily if necessary.

Throughput guideline is set by Bank Indonesia to avoid accumulation of BI-RTGS settlement risk at the end of day. Buckle and Camppell (2003) observed that the delay in payment by RTGS participant leads an increase in aggregate liquidity requirement, thus reduce efficiency of liquidity turnover. In this case, the throughput guidelines can be used to achieve the desired liquidity turnover efficiency. Therefore, throughput guideline could have risk-reduction benefits since they help to reduce the level of tiering in the financial system. The guideline substantially reduces the overall demand for intraday liquidity in the RTGS system. Bank Indonesia's guideline expects that the maximum value of transactions settled on zone III (throughput zone III) be at 40% from all the settlement transaction during the day. The distribution of BI-RTGS settlements that follow this guidance indicates that payment system liquidity condition is well managed (Figure 3). Furthermore, the increase of throughput zone III indicates the escalation of risk in BI-RTGS settlement. It also reflects the increase of liquidity tightening in banking industry. Historically, throughput zone III was increased due the increasing macroeconomy and financial system instability in 2005 (high inflation) and 2008-2009 (Global Financial Crisis).

Bank Indonesia also monitors level of connectivity among BI-RTGS participants (interconnectedness). This variable indicates the distribution of risk in the payment system. High interconnectedness indicates that risk is equally distributed and fewer banks dominate BI-RTGS transactions. Historically, interconnectedness will increase significantly at the end of year period following the increase of RTGS transactions from both customers and government approaching end of year/book closing (Figure 4). The significant decrease of interconnectedness on certain period should be closely monitored since it indicates the increase of concentration risk in BI-RTGS system, or in other words only a few banks that have big role in payment system and these banks' settlement failure will give spillover effect to other banks. The attention is mainly focused on banks with a high interconnectedness and net outflow pattern in BI-RTGS, in this case "bank A". These banks have more settlement obligations than incoming fund transfers and have transaction interconnectivity with many banks. Settlement failure on this type of bank will give spillover effect to other banks which have transaction relation.

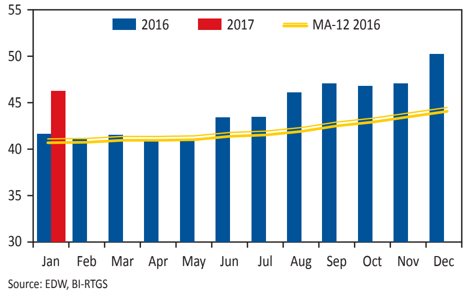


Figure 4. Industry's Interconnectedness

PSLI built by Bank Indonesia is a composite of three variables which are processed from settlement statistics. It consists as follows: (i) liquidity of RTGS (turnover ratio, queue, placement on Monetary Operations), (ii) Interbank Money Market Transactions (frequency, spread of interbank money O/N with policy rate), and (iii) interconnectedness. Historically, PSLI were able to capture changes in banking industry's behavior liquidity, at the time of magnitude change in looseness or tightness liquidity, driven by changes in macroeconomic and financial system stability. The increasing of PSLI can be read as an increase of tightness of payment system's liquidity. Therefore, Bank Indonesia can decide whether a further policy measures is needed. PSLI increases significantly in the period mid-2005 due to the increase of inflation pressure driven by rising fuel prices in that period (Figure 5). Furthermore, PSLI also showed an increase in the period between 2008 and 2009 when the Global Financial Crisis occurs due the increased volatility of exchange rate (Figure 6). Currently, PSLI is used both as a monitoring tool as well as an input to the formulation of macroeconomic stability policy and financial system policy.

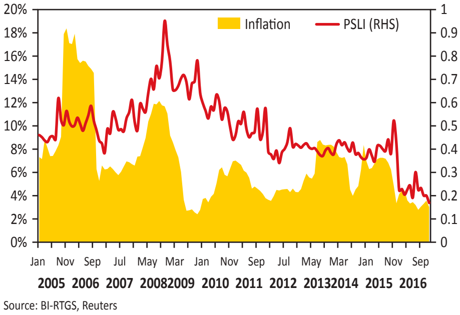


Figure 5. PSLI and Inflation

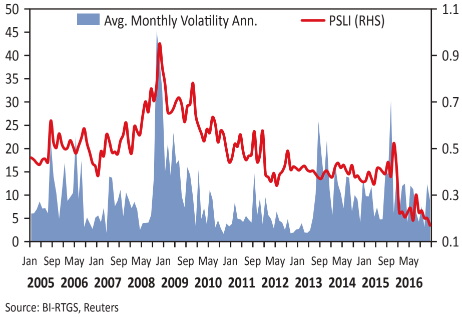


Figure 6. PSLI and Exchange Rate

4. PAYMENT SYSTEM STATISTICS AS A PROMPT INDICATOR FOR ECONOMIC ACTIVITY

Bank Indonesia uses statistics of retail payment transactions (ATM-Debit, credit card and electronic money) as a tracking indicator for household consumption, since this statistics could describe the dynamics of household consumption reasonably well (Figure 7). The retail payment statistics is available before the release of official Gross Domestic Product (GDP) that includes household consumption, therefore it gives an early figure on the growth of household consumption. The figure on current household consumption is very important for Bank Indonesia's policymaking ing macroeconomics and financial system considering it is the highest contributor in Indonesia's GDP, reaching 55.31% on 4th quarter 2016.

Non-cash transactions statistics are used to monitor the National Non-Cash Movement–launched in 2014–progress to decide the design of subsequent policies. Based on current statistics, National Non-Cash Movement progress fares fairly well. However, public preference of using cash is still moderately high. It is reflected from the slope ratio where it shows that cash transactions slope is higher than the slope of purely non-cash transactions (Figure 8). Therefore, Bank Indonesia has to intensify the effort of promoting the efficiency of non-cash transactions.

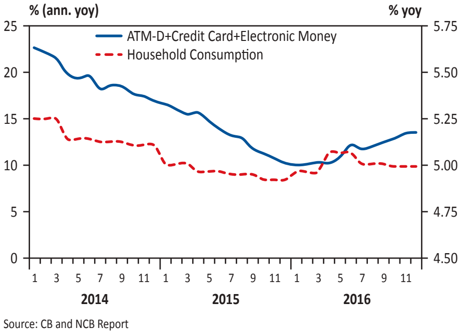


Figure 7. Retail Transaction

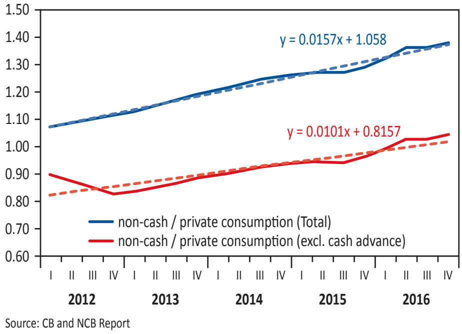


Figure 8. Growth of Non Cash Transactions

5. PAYMENT SYSTEM STATISTICS FOR POLICY REVIEW AND POLICY FORMULATION

Bank Indonesia also uses retail payment statistics to reinforce the policies of advancing consumer protection interest. In the end of 2016, Bank Indonesia adjust the maximum interest rate of credit card after considering the current macroeconomic conditions, financial data of issuer and the growth of credit card transactions. Along with the economic slowdown in the past few years, there is a declining trend of credit card usage (Figure 9) which followed by an increase in Non-Peforming Loan (NPL), especially in lower credit limit segment or more specifically people with low income levels (Figure 10).

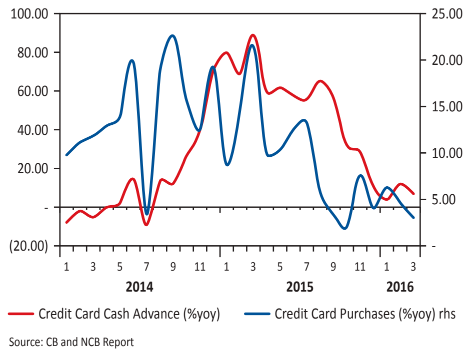


Figure 9. Credit Card Transactions

On the credit card provider/issuer side, the highest cost is attributed to Cost of Allowance for Impairment Losses which covers the upward trend in credit card NPL. In order to increase public confidence for using credit cards and to boost healthier credit cards industry, Bank Indonesia decreased the maximum interest rate of credit card from 2.95% per month to 2.25% per month. Along with this interest rate policy, Bank Indonesia requires issuers to deliver a closing statement to credit card customers who already closed their credit card. This measure was taken as a step to strengthen consumer protection of payment system due to the significant increase of credit card closure in mid-2016 which has the potential to put consumer's best interest at risk (Figure 11).

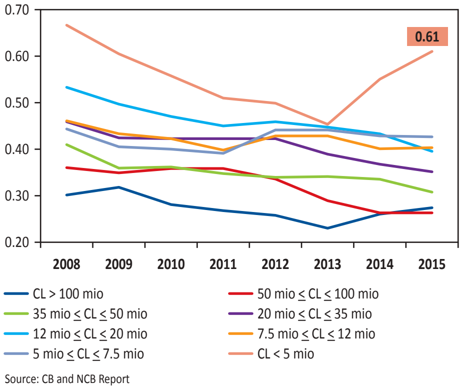


Figure 10. Credit Card Revolving Rate

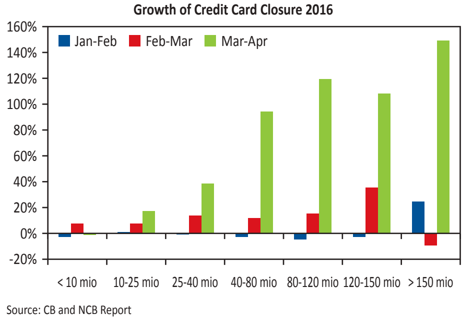


Figure 11. Credit Card Closure

6. CONCLUSIONS

BI-RTGS settlement data can reveal the broad picture of banking liquidity. PLSI devised by Bank Indonesia is able to reflect the liquidity conditions influenced by changes in both macroeconomic condition and financial system condition. Indicators of banks liquidity can be used as a monitoring tool in payment system as well as strengthening policies formulation in macroeconomic and financial system stability. In addition, retail payment statistics can capture the dynamics of the household consumption in Indonesia reasonably well and become an input for macroeconomic and financial system stability policy package. Retail payment statistics can also reflect the adoption of non-cash payment instruments in the society which will be considered as an input for payment system policy that aims to boost non cash transaction. Retail payment statistics will also be of service to Bank Indonesia's policy assessment and formulation concerning consumer protection framework and financial inclusion strategy.

REFERENCES

Buckle, S dan Campbell, E. (2003). Settlement Bank Behavior and Throughput Rules in an RTGS Payment System with Collateralized Intraday Credit. Bank of England Working Paper no. 209

Christian, S.(2011). Liquidity Requirements and Payment Delays Participant Type Dependent Preferences.European Central Bank.

Committee on Payment and Settlement Systems. (2014). Assessment of Compliance with the Core Principles for Systemically Important Payment System. The Fed Funds Services.

Drehmann, M. dan Tarashev, N. (2011). Systemic importance: some simple indicators. BIS Quarterly Review, Bank for International Settlements.

De Bandt, O. dan & Hartmann, P. (2000). Systemic Risk: A Survey. Working Paper Series 0035, European Central Bank.

European Central Bank (2010). The Payment System: Payments, Securities and Derivatives, and the Role of the Eurosystem. European Central Bank.

ECB-Bank of England Conference.(2008).Payments and Monetary and Financial Stability. Jurnal European Central Bank.

Federic, Hervo.(2008). Recent developments in intraday liquidity in payment and settlement system. Financial Stability Review of Banque de France No.11.

Hasan, Iftekhar, Heiko Schmiedel dan Liang Song (2012). Retail Payment and Bank Performance. Journal of Financial Services and Research, 41 (3): 163 – 195.

Hasan, Iftekhar, Tania De Renzis dan Heiko Schmiedel (2013). Retail Payments and the Real Economy. Working Paper No.1572/August 2013. European Central Bank.

International Monetary Fund (2011), Global Financial Stability Report. Durable Financial Stability: Getting There from Here.

International Monetary Fund /Bank for International Settlements/Financial Stability Board (2009), Guidance to Assess the Systemic Importance of Financial Institu-tions, Markets and Instruments: Initial Considerations.

Scholnick, Barry, Nadia Massoud, Anthony Saunder, Santiago Carbo-Valverde dan Francisco Rodriguez Fernandez (2008). The Economics of Credit Cards, Debit Cards and ATM: A Survey and Some New Evidence. Journal of Banking and Finance, 32, 1468 – 1483.

Troy Gill, Dilhan Perera and David Sunner (2012), Electronic Indicators of Economic Activity, Reserve Bank of Australia.

Zandi, Mark, Virenda Singh dan Justin Irving, 2013. The Impact of Electronic Payments on Economic Growth. Economic & Consumer Credit Analytics. Moody's

This page intentionally left blank

IPS14: Educating Students To Be Workforce-Ready Practitioners And Users of Statistics

Preparing Engineers For Practice

Professor Helen MacGillivray
President-Elect,
International Statistical Institute
Australian Senior Learning and Teaching Fellow

Abstract

Although engineering students are as diverse as the careers and workplaces they go to, there are some characteristics in common which must to be taken into account for sustained, authentic student engagement and current and future learning. In all disciplines 'serviced' by statistics, it is key to observe, listen and translate what is said by staff, employers and students into what is needed. This applies particularly in engineering, often complicated by constant searching by faculty and the profession for the 'holy grail' in engineering education. A brief overview of ongoing lessons from 40 years of working with engineering student, staff and past students is given in this presentation, including demonstration that decades-long advocacy for the training of professional statisticians also applies to statistically preparing engineering graduates for the workplace.

Keyword : Engineering student, Engineering practice, Statisticians.

IPS15: Time Series – Novel Methods and Applications

Applications of Distance Correlation to Time Series

Richard A. Davis*
Columbia University, New York, USA – rdavis@stat.columbia.edu
Muneya Matsua
Nanzan University, Nagoya, JAPAN – mmuneya@gmail.com
Thomas Mikosch
University of Copenhagen, Copenhagen, DENMARK – mikosch@math.ku.dk
Phyllis Wan
Columbia University, New York, USA – phyllis@stat.columbia.edu

Abstract

The use of empirical characteristic functions for inference problems, including estimation in some special parametric settings and testing for goodness of fit, has a long history dating back to the 70s (see for example, Feuerverger and Mureika (1977), and Csörgö (1981)). More recently, there has been renewed interest in using empirical characteristic functions in other inference settings. The distance covariance and correlation, developed by Szekely and Rizzo (2007) for measuring dependence and testing independence between two random vectors, are perhaps the best known illustrations of this. We apply these ideas to stationary univariate and multivariate time series to measure lagged auto- and cross-dependence in a time series. Assuming strong mixing, we establish the relevant asymptotic theory for the empirical auto- and cross-distance correlation functions. We also apply the auto-distance correlation function (ADCF) to the residuals of an autoregressive processes as a test of goodness of fit. Under the null that an autoregressive model is true, the limit distribution of the empirical ADCF can differ markedly from the corresponding one for an iid sequence. We illustrate the use of the empirical ADCF for testing dependence and cross-dependence of time series in a variety of different contexts.

Keywords: Auto- and cross-distance correlation function, Testing independence, Time series, Strong mixing, Ergodicity, Fourier analysis, *U*-statistics, AR process, Residuals.

Efficiently Estimating Discrete and Continuous Time GARCH Models with Irregularly Spaced Observations

William T.M. Dunsmuir[1]
University of New South Wales, Sydney, Australia – W.Dunsmuir@unsw.edu.au

Abstract

The simplest generalized autoregressive conditional heteroscedastic, GARCH(1,1), model for the conditional variance of incremental changes in a time series observed in discrete time has been a "workhorse" in modelling financial returns. There are essentially two continuous time limits of the GARCH(1,1) process as the time between observations shrinks to zero. The bivariate diffusion limit process, due to Nelson, involves two possibly correlated Brownian motion but does not allow jumps in the continuous time limit process. More recently the COGARCH process defined in terms of a Levy process has been proposed and studied. This can also be obtained as the continuous time limit of the discrete time GARCH(1,1) process. The COGARCH process has a single source of driving noise and allow jumps.

Because the number, location in time and size of jumps cannot be observed directly using equally or irregularly spaced observations on the continuous time process., a major challenge for the COGARCH model is estimating the unknown parameters. The likelihood for the COGARCH model is intractable and requires careful computational implementation. Sequential Monte Carlo (SMC) with a continuous resampling method to estimate the likelihood function and ensure it is continuous in the parameters is used. We show that the SMC based method outperforms the quasi-maximum likelihood methods previously proposed in the literature in terms of bias and standard errors of estimation. Application to high frequency financial returns data will be presented.

The SMC approach can also be used to estimate the parameters of traditional discrete time GARCH models and variants such as the Markov regime switching GARCH model when they are irregularly observed. Illustrations for financial time series as well as high frequency wind measurements will also be presented.

Keywords: Volatility; COGARCH and GARCH; Sequential Monte Carlo; Intractable likelihood.

1 Joint work with Damien Wee and Feng Chen, University of New South Wales

Volatility Modelling Using Box–Cox Stochastic Volatility Model: Evidence from Bank Indonesia (2010–2015)

Didit B. Nugroho*
Dept. of Mathematics, Satya Wacana Christian University, Salatiga, Indonesia – didit.budinugroho@staff.uksw.edu
Tundjung Mahatma
Dept. of Mathematics, Satya Wacana Christian University, Salatiga, Indonesia – t.mahatma@staff.ukse.edu
Yulius Pratomo
Dept. of Economics, Satya Wacana Christian University, Salatiga, Indonesia – yulius.pratomo@staff.ukse.edu

Abstract

Volatility of the foreign exchange rate in the Indonesia's central bank is modelled and estimated. The empirical analysis is based on daily data on selling four foreign currencies to Indonesia rupiah (IDR); these include the Swiss franc (CHF), the Euro (EUR), the British pound (GBP), the Japanese yen (JPY), and the US dollar (USD) over the period of January 2010 to December 2015. This study employs a class of nonlinear asymmetric stochastic volatility model by applying the Box–Cox transformation to the volatility equation and calls it "Box–Cox asymmetric stochastic volatility (BCASV) model". The empirical results show that the CHF and GBP returns series are in favour of the Box–Cox specification in terms of the 95% highest posterior density interval for the Box–Cox parameter. Results also indicate that the leverage effect is absent for all returns series. In particular, we find evidence of significant positive correlation between returns and volatility in adopting the CHF, JPY, and USD returns series. Furthermore, the log-marginal likelihood scores confirm that the BCASV model is superior to the asymmetric stochastic volatility (ASV) model for CHF, EUR, and GBP returns series.

Keywords: Asymmetric effect; Box–Cox transformation; Exchange rate volatility; Stochastic volatility.
JEL Classification: C32, G12, G13

1. INTRODUCTION

Volatility of the asset price returns is defined as a statistical measure of the risk of investors holding. It can be measured by using the standard deviation or variance of logarithmic returns. Higher volatility (wider range of fluctuations) usually denotes higher asset risks. A higher risk may increase the agency costs of debt financing (Botta, 2014). This means that debt is less attractive for firms with a higher asset volatility.

Recognizing the importance of asset returns volatility for portfolio allocation, risk management, and valuation of financial derivatives, the study of volatility modelling and forecasting of financial time series have become a significant area of research within financial econometrics. It is well known that asset returns volatility is not constant, but varies substantially over time. In the literature, two models commonly used to model the time varying conditional volatility of asset returns are the autoregressive conditional heteroscedasticity/generalized autoregressive conditional heteroscedasticity (ARCH/GARCH) models respectively proposed by Engle (1982) and Bollerslev (1986), and the stochastic volatility (SV) models introduced by Taylor (1982, 1994). It is also well known that asset returns volatility exhibits an asymmetric (leverage) effect. This feature, first recognized by Black (1976) for stock returns, is defined as a negative correlation between current return and future volatility, which means volatility is greater after bad news (negative return) than after good news (positive return) of the same magnitude in the asset market.

Theoretically, the SV model is much more flexible, realistic, and better performing than the ARCH/GARCH-type models since an innovation term is embedded in the volatility dynamics. The SV model specifies the logarithmic squared volatility as a first order autoregressive Gaussian process. Yu *et al*. (2006) presented an extension to the SV model by applying the Box–Cox transformation to the volatility equation, which they called the Box–Cox stochastic volatility (BCSV) model. Ignoring the leverage effect, they applied the BCSV model to daily returns of the dollar/pound exchange rate and found highly significant evidence against the SV model.

The aim of this study is to empirically investigate the existence of Box–Cox parameter and leverage effect in the returns of the Indonesia's currency using the asymmetric BCSV model. This study is important as it is the first to apply the BCSV models to daily returns of the foreign currency rates to Indonesia rupiah (IDR). The rest of the paper is structured as follows: section 2 presents the literature review, section 3 the model and its estimation, section 4 the characteristic of data and econometrics analysis, and section 5 presents the conclusions and future works.

2. LITERATURE REVIEW

On the empirical level, several applications on the development of Indonesian market mainly use the GARCH model. For example, Sumiyana (2007) studied the behavior of daily returns volatility in Indonesian Stock Exchange market by using size, trading volume, bid-ask spreads, and up-down market. Recently, Saadah and Panjaitan (2016) identified the volatility of IPO shares price on Indonesian Stock Exchange market during 2000–2013 through GARCH model. Recently, in the Indonesian Foreign Exchange market, Saputri *et al*. (2016) and Salim *et al*. (2016) investigated the behaviour of exchange rate volatilities for the EUR and JPY data sets using the ARCH and GARCH models, respectively. Saputri *et al*. (2016) found that ARCH model with skewed generalized Student-t error distribution give better results than ARCH models with normal, Student-t, non-central Student-t distribution. Meanwhile, Salim *et al*. (2016) found that GARCH model with Student-t error distribution give better results than GARCH model with normal distribution.

In the SV context, the previous study indicate that only few researches have been conducted based on Indonesian Market (e.g., see Asai & Unite (2010) and Chiba & Kobayashi (2013)) and none of them examine foreign exchange volatility nor compare the performance of ASV and BCASV models. Asai & Unite (2010) examined the asymmetric effect for the stock market indices of the Association of Southeast Asian Nations (ASEAN5 (*op cit.*)) countries on return volatility using ASV model and found that there is no asymmetric effect for the Indonesia data set.

3. BCASV MODEL AND ITS ESTIMATION

The BCSV model was introduced by Yu *et al.* (2006) and generalized into the BCASV model by Zhang & King (2008) that incorporates the asymmetric relationship captured by allowing a nonzero correlation between the current asset return and future volatility. The BCASV model can be expressed as

$$\left.\begin{aligned} R_t &= \sqrt{g_t}e_t \\ \alpha_{t+1} &= \mu + \phi(\alpha_t - \mu) + \sigma_u u_{t+1} \\ \alpha_1 &= \mathcal{N}\left(\mu, \tfrac{\sigma_u^2}{1-\phi^2}\right) \\ corr(\epsilon_t, u_{t+1}) &= \rho \end{aligned}\right\} \tag{1}$$

where $e_t, u_{t+1} \sim \mathcal{N}(0,1)$ with $\mathcal{N}$ represents normal distribution, and g_t is defined by

$$g_t = g(\alpha_t, \delta) = \begin{cases} (1+\delta\alpha_t)^{\frac{1}{\delta}} & \text{if } \delta \neq 0 \\ \exp(\alpha_t) & \text{if } \delta = 0 \end{cases}$$

The model assumes that the returns have no autocorrelation, and it also allows for possible correlation between the return and volatility errors. In addition, the values of $\delta = 0$, $\delta = 0.5$, and $\delta = 1$ correspond to the "classical" parametric SV models (see Yu *et al.* (2006))

Here, we perform the Bayesian approach for parameter estimations by the MCMC algorithm. In particular, the HMC method developed in Nugroho & Morimoto (2015) is employed in the MCMC algorithm to generate the values of parameter that cannot be sampled directly. Furthermore, all empirical results are obtained via the code implementation of MATLAB for the MCMC-simulation.

4. EMPIRICAL ANALYSIS

This section describes the data, report the descriptive statistics, econometrics analysis and discuss the main findings.

4.1. Exchange rate data

We start our analysis of the foreign exchange volatility by illustrating some stylized facts of Indonesia currency market with intradaily data for the CHF, EUR, GBP, JPY, and USD selling exchange rates to IDR over a period from January 2010 to December 2015, excluding weekends and holidays. Notice that the five currencies are the major currencies in the world economy in 2010 (Salvatore, 2011). The data sets were obtained from Bank Indonesia's Internet website (http://www.bi.go.id)

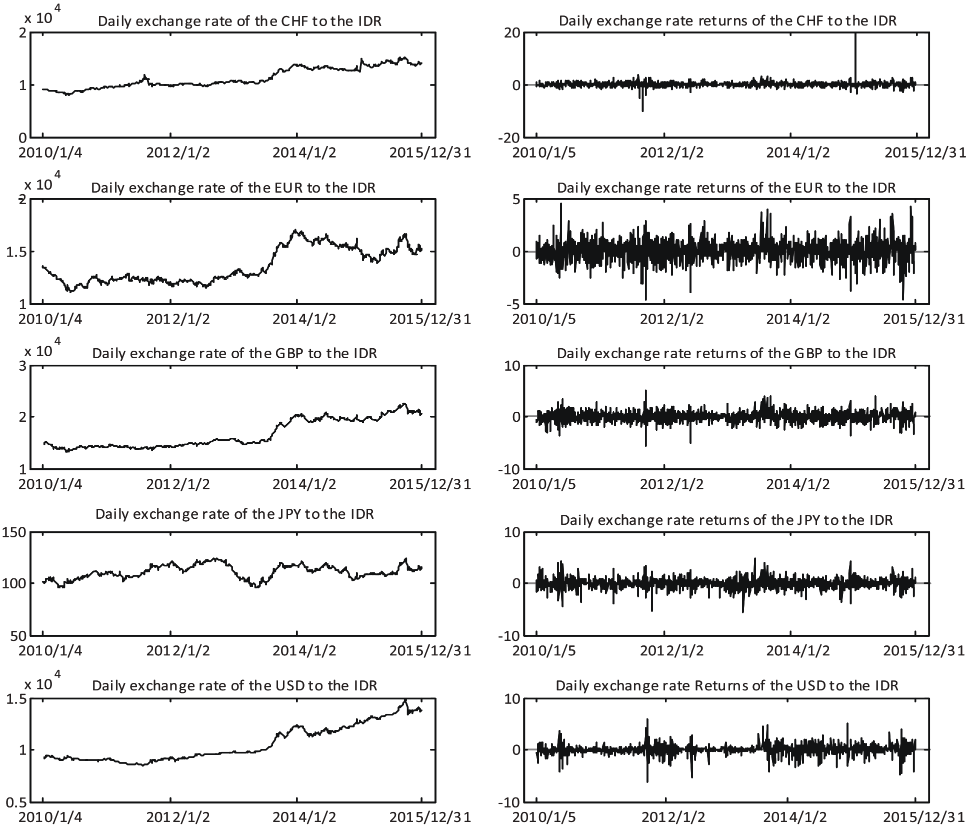


Figure 1. Time Series Plots for Daily Selling Exchange Rate and Correspondence Percentage Daily Returns

Let P_t be the price of day t. The return (in percent) of day t, denoted by R_t, is computed using mean-corrected and variance-scaled returns defined by

$$R_t = 100 \times \frac{Y_t}{s(Y_t)} \text{ with } Y_t = \ln\left(\frac{P_t}{P_{t-1}}\right) - \frac{1}{T}\sum_{t=1}^{T} \ln\left(\frac{P_t}{P_{t-1}}\right),$$

where $s(Y_t)$ is the sample standard deviation (SD) of Y_t. We omit all returns equal to zero because they will provide poor estimates of volatility. The series of daily exchange rates and returns are displayed in Figure 1. It shows that there is an upward trend for all exchange rates, but all returns are rather stable around a mean of zero.

Table 1 summarizes the descriptive statistics of the returns data. From the Jarque–Bera (JB) normality test and the Ljung–Box (LB) correlation test for 8 lags, all observed returns are neither normally distributed nor serially correlated. In this study, we ignore this non-normality in order to provide a framework for developing more general models. The mean of each return is in average around zero percent, while the standard deviation is equal to one percent.

Tabel 1. Summary Statistics for Mean-Corrected and Variance-Scaled Selling Exchange Rate Return Data

Returns	Number of Observations	Mean	Min.	Max.	JB stat. (normality)	LB stat. (lag-8 autocorr.)
CHF	1471	0	−10.22	19.75	756284.7 (No)	11.40 (No)
EUR	1471	−0.000	−4,72	4.54	184.5 (No)	6.45 (No)
GBP	1472	−0.000	−5.50	4.93	276.8 (No)	15.17 (No)
JPY	1469	−0.000	−5.77	4.69	671.4 (No)	10.70 (No)
USD	1415	0	−6.24	5.81	2231.2 (No)	1.68 (No)

4.2. Empirical Results

Table 2 summarizes the empirical results, including the posterior means, standard deviations, 95% highest posterior density (HPD) intervals for all the parameters, and the log-marginal likelihood values for both models. The main results are as follows.

Table 2. Parameter Estimates (Standard Deviation Printed in Italic, HPD Intervals in The Next Row) of The ASV and BCASV Models for The Selling Exchange Rate Return Data

Returns	Model	LML	α	ϕ	σ_u	ρ	δ
CHF	ASV	−1696.0	−0.603 (0.096)	0.887 (0.034)	0.349 (0.051)	0.180 (0.078)	
			(−0.755,−0.445)	(0.836,0.940)	(0.261,0.426)	(0.055,0.310)	
	BCASV	−1678.2	−0.888 (0.132)	0.932 (0.019)	0.293 (0.032)	0.241 (0.081)	−0.742 (0.074)
			(−1.150,−0.633)	(0.894,0.965)	(0.240, 0.350)	(0.070,0.392)	(−0.872,−0.617)
EUR	ASV	−2025.1	−0.136 (0.098)	0.953 (0.015)	0.148 (0.022)	0.018 (0.119)	
			(−0.297,0.024)	(0.930,0.977)	(0.110,0.182)	(−0.169,0.222)	
	BCASV	−2019.9	−0.208 (0.107)	0.951 (0.016)	0.156 (0.024)	0.034 (0.115)	−0.565 (0.326)
			(−0.421,0.007)	(0.917,0.978)	(0.112,0.203)	(−0.201,0.246)	(−1.252,0.030)
GBP	ASV	−2005.4	−0.166 (0.094)	0.935 (0.020)	0.197 (0.031)	−0.012 (0.100)	
			(−0.330,−0.022)	(0.904,0.966)	(0.147,0.251)	(−0.176,154)	
	BCASV	−1998.7	−0.261 (0.099)	0.928 (0.024)	0.203 (0.033)	0.003 (0.107)	−0.614 (0.259)
			(−0.455,−0.006)	(0.885,0.969)	(0.138,0.261)	(−0.198,0.197)	(−1.061,−0.075)
JPY	ASV	-1903.8	−0.317 (0.110)	0.920 (0.022)	0.304 (0.044)	0.252 (0.088)	
			(−0.494,−0.139)	(0.882,0.954)	(0.236,0.383)	(0.110,0.401)	
	BCASV	-1904.6	−0.395 (0.125)	0.921 (0.025)	0.310 (0.050)	0.238 (0.084)	−0.247 (0.155)
			(−0.636,−0.140)	(0.872,0.965)	(0.220,0.412)	(0.073,0.400)	(−0.573,0.038)
USD	ASV	-1465.7	-0.827 (0.226)	0.938 (0.014)	0.490 (0.049)	0.149 (0.068)	
			(−1.174,−0.440)	(0.916,0.962)	(0.408,0.569)	(0.041,0.263)	
	BCASV	-1477.3	−0.797 (0.243)	0.940 (0.013)	0.471 (0.047)	0.143 (0.068)	0.024 (0.068)
			(−1.316,−0.361)	(0.914,0.965)	(0.384,0.568)	(0.004,0.273)	(-0.116,0.146)

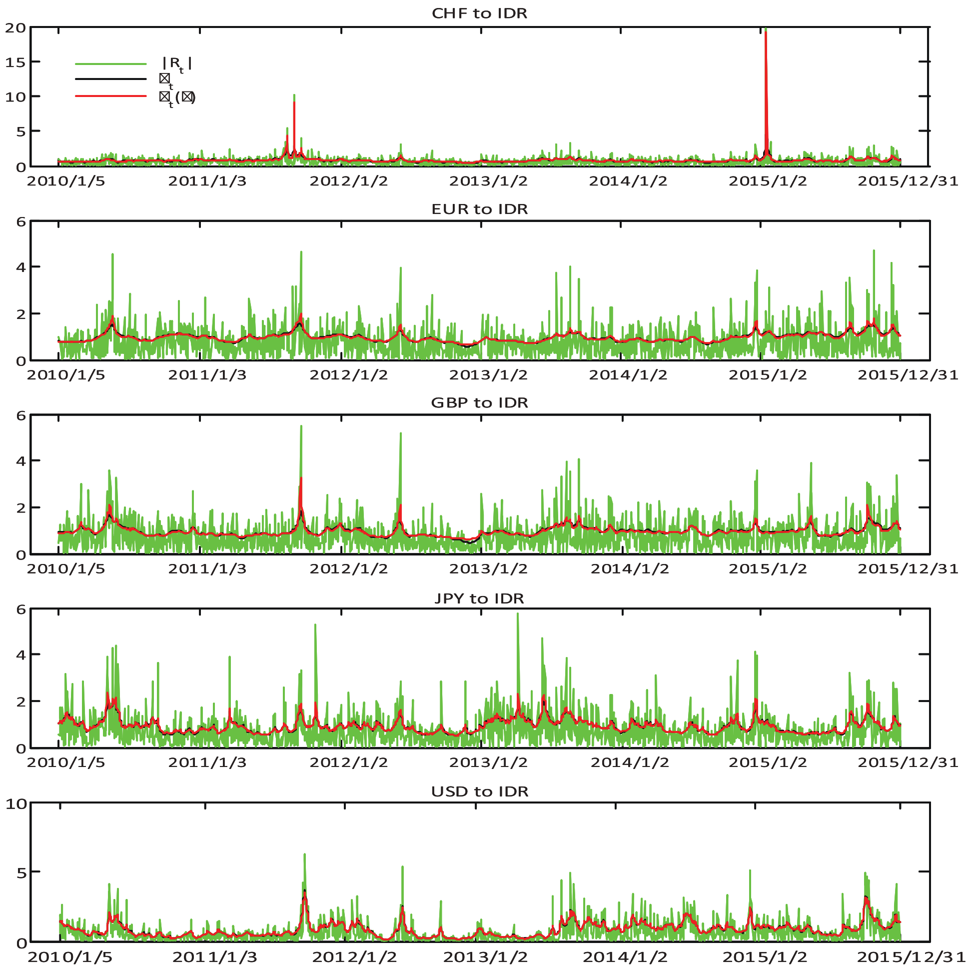


Figure 2. Plots of Daily Absolute Returns and Volatility Estimates for The Selling Exchange Rate Return Data

First, the posterior mean of ρ in both models adopting all returns, save the ASV model adopting GBP returns, is positive and the 95% HPD interval excludes 0 in the both models adopting CHF, JPY, and USD returns series. It means that the impact in the CHF, JPY and USD returns is symmetric, which implies that positive and negative news have the same effect on volatility. We even found that the 99% HPD interval of ρ exclude 0 for the CHF and JPY return series (results not shown). In the case of AVS model adopting GBP returns, the posterior mean of ρ is negative but not significantly different from zero. All these results indicate the non-existence of the leverage effect in the five currencies during the sample periode. It means that the asymmetric news effect is absent.

Second, the posterior mean of δ in the BCASV model adopting the CHF and GBP returns is negative and significantly different from 0, 0.5, and 1. Thus, the CHF and GBP returns provide highly significant evidence against all classical ASV models. Although not reported, we find that even 99% HPD interval of ρ in the BCASV model adopting CHF excludes 0.5 or 1.

Third, the log-marginal likelihood suggests that volatility estimates using BCASV model for the CHF, EUR, and GBP returns series are more accurate than the estimates computed by the ASV model. The results confirm the previous significant evidence of applying Box–Cox transformation. The only exception is the model adopting EUR returns, where its posterior distribution is mostly located in the negative range (results not shown).

The estimates of the unobserved volatilities, $\sqrt{g_t}$, under the ASV and BCASV models are given in Figure 2. We can see that the estimates from the two models are comparable and all returns, save the JPY returns, was very volatile in September 2011. In addition, the extremely high volatility period in the CHF returns occurred in January 2015.

5. CONCLUSIONS AND FUTURE WORKS

In this study we have modeled and estimated returns volatility in Indonesian currency market by applying a class of nonlinear asymmetric stochastic volatility model for daily observations on the five currency returns series over the period of January 2010 to December 2015. Empirical results showed evidence supporting the Box–Cox transformation of volatility process in adopting the CHF, GBP, and JPY returns and did not suggest the existence of the leverage effect in all exchange rate returns. Furthermore, the model may be extendable to non-Gaussian density of return errors.

REFERENCES

Asai, M., & Unite, A.A. (2010). General asymmetric stochastic volatility models using range data: Esyimation and empirical evidence from emerging equity markets. *Applied Finansial Economics 20 (13)*, pp. 1041–1049.

Black, F. (1976). Studies of stock price volatility changes. Paper presented at the proceedings of the 1976 meetings of the american statistical association, Bus. Econ. Stat.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, pp. 307–327.

Botta, M. (2014). *Firm financing in the Euro Area: How asset risk affects capital structure decisions within the monetary union*. Hamburg: Anchor Academic Publishing.

Chiba, M., & Kobayashi, M. (2013). Testing for a single-factor stochastic volatility in bivariate series. *Journal of Risk and Financial Management 6*, pp. 31–61.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, pp. 987–1007.

Nugroho, D. B., and T. Morimoto. 2015. Estimation of realized stochastic volatility models using Hamiltonian Monte Carlo-based methods. *Computational Statistics* 30 (2), pp. 491–516.

Saadah, S., & Panjaitan, Y. (2016). The Green Shoe Option's Effectiveness at Stabilizing the IPO'S Stock Price on the Indonesian Stock Exchange (2000-2013). *Gadjah Mada Int. Journal of Business 18 (1)*, pp. 71–82.

Salim, F.C., Nugroho, D.B., & Susanto, A. (2016). Model volatilitas GARCH(1,1) dengan *error Student-t* untuk kurs beli EUR dan JPY terhadap IDR [GARCH(1,1) volatility model with Student-t error for the EUR and JPY buying rates to IDR]. *Jurnal MIPA 39 (1)*, pp. 63–69.

Salvatore, D. 2013. *International Economics, 11th ed.* John Wiley & Sons.

Saputri, E.D., Nugroho, D.B., & Setiawan, A. (2016). Model volatilitas ARCH(1) dengan *returns error* berdistribusi *skewed Student-t* [ARCH(1) volatility model with Student-t returns error distribution]. *Jurnal MIPA 39 (1)*, pp. 63–69.

Sumiyana. 2007. Behavior of stock price variability over trading and nontrading periods, and daily return volatility. *Gadjah Mada Int. Journal of Business 9 (3)*, pp. 399–443.

Taylor, S.J. (1982). Financial returns modelled by the product of two stochastic processes---a study of the daily sugar prices 1961–75 (reprint as pages 60–82 in Shephard, 2005).

Taylor, S.J. (1994). Modelling stochastic volatility: A review and comparative study. *Mathematical Finance 4*, pp. 183–204.

Yu, J., Yang, Z., & Zhang, X. (2006). A class of nonlinear stochastic volatility models and its implications for pricing currency options. *Comput. Stat. Data Anal. 51*, pp. 2218–2231.

Zhang X., & King, M.L. (2008). Box–Cox stochastic volatility models with heavy-tails and correlated errors. *J. Empir. Financ. 15 (3)*, pp. 549–566.

IPS17: Input-Output Analysis

Functional Analysis of Industrial Clusters in Malaysia

Norhayati Shuja' *
Department of Statistics Malaysia, Putrajaya, Malaysia - yatishuja@stats.gov.my

Yap Bee Wah
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia - beewah@tmsk.uitm.edu.my

Abstract

Economic growth is influenced by the industrial total output rate which, on the other hand, depends on industrial linkages. Forward and Backward linkage indices deal with the performance of individual industries which thus highlights the importance to study linkages among industries. Cluster analysis of input and output coefficients can reveal important clusters which are groups of industries that have strong supply and buyer relationships. The national industrial clusters in Malaysia were determined by examining the inter-industry selling and purchasing relationship based on the Malaysia Input-Output Table 2010. The input and output coefficients were obtained from Input-Output table, 124 industries by 124 industries. Cluster analysis of the correlation matrix for pairwise linkage ratios was carried out using Ward's hierarchical clustering algorithm. The results of the hierarchical cluster analysis show that there are 11 distinct industrial clusters. Based on the results of functional analysis conducted, it was found that the largest contribution to the economy was cluster 3 (manufacturing) which contributed the highest to the total output.

Keywords: Input-output analysis, Industry clusters, Linkage ratio, Cluster analysis.

1. INTRODUCTION

Industrial clusters are industrial aggregations formed by industrial sectors with close economic links in a specific location. The industrial clusters depends on industrial linkages of goods and services among industries sectors. They comprise of industries that have strong supply and buyer relationship and thus provide better information of inter-industry dependence. The relationship of supply and buyer in functional links among industries will exhibit growth and development of industrial clusters in a national economy. The factors affecting the growth and development of industrial clusters refer to the factors leading to structural changes of links among industries, that is, technology changes, changes of output levels and changes of input goods prices (Zhu, 2010).

An important point to note is that, clusters tend to have strong inter-dependency effects if the industry sectors are related to each other, for example, industry that uses the same technology or industry that have a supplier and buyer relationship. It follows that cluster analysis is important technique used to aggregate industry sectors that have similar economic characteristics, including inter-industry linkages. There are a large number of industry sectors in input-output tables and it is necessary to aggregate the sectors in order to publish the results in a comprehensive way. Different approaches can be employed on how the sectors should be aggregated (Hoen, 2002). There are possibilities to search for clusters of sectors with strong linkages (Aroche-Reyes, 2001) or the sectors that have the same input structure (Dietzenbacher & Lahr, 2001). There have been several empirical studies in the field of industrial clusters (O'Huallachain, 1984; Akgungor *et al.*, 2002; Titze *et al.* 2011 & Brachet *et al.*, 2011). However, they differ not only in definitions used and areas studied, but also in the objectives and methodology. Rasmussen (1956) used the Leontief inverse matrix and Chenery & Watanabe (1958) used the input coefficient matrix to identify key sectors by analyzing Input-Output tables. Hypothetical

Extraction method is also an alternative method in studying key linkages (Dietzenbacher *et al.*, 1993 & Dietzenbacher *et al.*, 1997). Key linkage indices deal with individual industries while industry cluster is a group of industries that has strong complementary linkages. Industry clusters comprise of industries that have strong supply and buyer relationship and thus provide better information of inter-industry dependence.

This study will focus on economic links among industrial sectors in a national economy. There are various methods for analyzing and identifying industry clusters. The multivariate statistical analysis and input-output analysis are the main methods in identifying industrial clusters. Hence, the objective of this paper is to identify the national industry clusters in Malaysia by examining the inter-industry selling and purchasing relationship based on the Malaysia's 2010 Input-Output table using cluster analysis approach. Cluster analysis is useful in analyzing which sectors are strongly connected to each other.

This paper is organized as follows. Section II provides a review on industrial clusters identification. The methodology is explained in Section III. The results are presented in Section IV. Finally the conclusion is given in Section V.

2. LITERATURE REVIEW

An industrial cluster is a group of industries that has relatively strong complementary linkages within, but has relatively weak linkages outside the group. Different kinds of multivariate analysis have been applied to aggregate the industry sectors in the identification of industrial clusters. The most commonly used methods are factor analysis and cluster analysis. Czamanski (1971), Bergsman *et al*. (1972) identified groups of industries that have similar location patterns among Standard Metropolitan Statistical Area (SMSA) in the United States. The employment data of 144 manufacturing activities and 42 non-manufacturing activities were used in their studies. The grouping of industries with similar location pattern was identified using factor analysis. The aggregation of 186 industries resulted in 42 clusters. They extended their studies using a data base which consists of 480 industrial classifications over 311 metropolitan areas of SMSAs. The 480 x 480 symmetric matrix was analysed using hierarchical clustering analysis. The analysis revealed 40 groups of industries (Bergman *et al*., 1975).

An application of factor analysis to an input-output table as a means of identifying functionally related groups of industries was carried out by Roepke *et al*. (1974). Three separate factor analyses were applied to a 44 x 44 interindustry transaction matrix of the Canadian Province of Ontario, 1965 Input-Output table. They first applied conventional factor analysis on aggregated transaction matrix and thirteen industrial (dimensions) clusters were identified. Another two factor analyses were carried out using R-mode and Q-mode factoring. The factoring analysis show some similarities to the results of the aggregate matrix analysis. Blin & Cohen (1977) applied hierarchical algorithms to the U.S. 1967 input-output table. This method was developed on Sokal & Mechener (1958) and Ward (1963) approach. They compared the dendrogram obtained from centroid algorithm with Ward's method. The centroid method gives less discrimination for alternative specifications of technological similarity. Cabrer *et al*. (1991) carried out six hierarchical cluster analyses using the Euclidean distance (centroid, median and Ward) on the Leontief inverse matrix and on the technical coefficient matrix of the Spanish Input-Output table for 1985. The median linkage gives the best cluster aggregation method. Feser & Bergman (2000) followed the methodology used by Czamanski (1979) which involves construction of matrices of correlation coefficients relating industries on the basis of their selling and buying relationships. Four correlation coefficients for each pair were calculated. They used the symmetric matrix with each element as the maximum correlation for factor analysis. Hill & Brennan (2000) provided more insights on industrial cluster analysis using data for the Cleveland-Akron metropolitan area. They combined cluster analysis with discriminant analysis to identify the driver industries.

Similarly, Oosterhaven *et al*. (2001) used the intermediate purchases and sales structure to group different industries into one cluster. The industries with large loadings were treated as members of an industrial cluster (Hofe & Bhatta, 2007, and Kelton *et al*. 2008). In the following section, we present the methodology for determining the clusters in Malaysia Input-Input table, 2010.

3. METHODOLOGY

In this paper, the identification of linkages and industrial clustering involve four stages and is summarised in Figure 1.

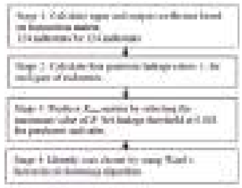


Figure 1. The Methodology Flow

Stage 1 involves obtaining the forward and backward linkages based on input and output coefficient matrix. The input and output coefficients of an individual industry are calculated based on transaction matrix, 124 industries by 124 industries, which was obtained from Malaysia Input-Output table, 2010.

An intermediate input matrix (*Z*) represents the inter-industry transaction of goods and services. Each element (*ij z*) in the *Z* matrix provides the value of goods and services sold by row industry *i* to column industry *j*.

For each industry, total intermediate goods purchases, $P_j = \sum_i z_{ij}$ and sales $S_i = \sum_j z_{ij}$. Purchases and sales coefficients between any two industries, i and j represent forward and backward linkages, and is derived as follows;

$$x_{ij} = \frac{Z_{ij}}{P_j}, \quad y_{ij} = \frac{Z_{ij}}{S_i}$$

Where, z_{ij} is the dollar value of purchases by industry j from industry i

P_j is the total purchases by industry *j*

x_{ij} is the ratio of purchases by industry *j* from industry *i* to the total purchases by industry *j*

S_i is the total sales by industry *i*

y_{ij} is the ratio of sales by industry *i* to industry *j* to the total sales by industry *i*

Each coefficient is an indicator of dependence between industry *i* and *j* in terms of relative purchases and sales linkage. Each vector in *X* indicates the purchasing pattern of industry *j* and each vector of in *Y* indicates the selling pattern of industry *i*.

In stage 2, we apply Feser & Bergman (2000) approach based on four correlation coefficients to capture each pair of industries with significant similar purchasing or selling patterns.

$r(x_i,x_j)$: industry *i* and industry *j* are correlated by having similar purchasing patterns

$r(y_i,y_j)$: industry *i* and industry *j* are correlated by having similar selling patterns

$r(x_i,y_j)$: industry *i* has a purchasing pattern which is similar as industry *j*'s selling pattern

$r(y_i,x_j)$: industry *i* has a selling pattern which is similar as industry *j*'s purchasing pattern

In Stage 3, the *R*max matrix which consists of the maximum correlations among the purchases and sales coefficients for each pair of industries was obtained.

According to Feser (2005), the input matrix is a symmetric matrix of the strongest linkages among the purchases and sales coefficients for each pair of industries. An input-output transaction matrix with a set of purchasing industries (buyers) from industry *i*, is denoted as *Bi*, and a set of supplier industries (sellers) to industry *i*, is denoted as *Si* . If an industry is acting as a supplier to both industry *i* and industry *j*, it is in the group I_{ij}^{SS}. Similarly, if an industry is purchasing from both industry *i* and industry *j*, it is in the group I_{ij}^{BB}. For industries that are selling to industry *i* and buying from industry *j*, it is define a group I_{ij}^{SB}. Alternatively, I_{ij}^{SB} is a group which contains industries buying from industry *i* and selling to industry *j*. Mathematically, it is defined as:

$$I_{ij}^{SS} = S_i \cap S_j, U_{ij}^{SS} = S_i \cap S_j$$

$$I_{ij}^{BB} = B_i \cap B_j, U_{ij}^{BB} = B_i \cap B_j$$

$$I_{ij}^{SB} = S_i \cap B_j, U_{ij}^{SB} = S_i \cap B_j$$

$$I_{ij}^{BS} = B_i \cap S_j, U_{ij}^{BS} = B_i \cap S_j$$

where, *SS*

U_{ij}^{SS} contains a group of suppliers to either industry *i* or industry *j*, or both.

U_{ij}^{BB} contains a group of buyers from either industry *i* or industry *j*, or both.

U_{ij}^{SB} contains a group of suppliers to industry *i* and buyers from industry *j*.

U_{ij}^{BS} contains a group of buyers from industry *i* and suppliers to industry *j*.

Some industries purchase intermediate input from industry *i (j)* and supply output to industry *j (i)*. Based on Czamanski's four-dimension linkage pattern, Feser (2005) has constructed the following four coefficients to represent the functional relationship between each pair of industries *Si* and *Bi* .

$$R_{ij}^{SS} = \frac{I_{ij}^{SS}}{U_{ij}^{SS}},\ R_{ij}^{BB} = \frac{I_{ij}^{BB}}{U_{ij}^{BB}},\ R_{ij}^{SB} = \frac{I_{ij}^{SB}}{U_{ij}^{SB}},\ R_{ij}^{BS} = \frac{I_{ij}^{BS}}{U_{ij}^{BS}}$$

were obtained, where

R_{ij}^{SS} is the ratio of the number of common suppliers to industries *i* and *j* over the total number of suppliers to industries *i* and *j*.

R_{ij}^{BB} is the ratio of the number of common buyers to industries *i* and *j* over the total number of buyers to industries *i* and *j*.

The highest ratio represents the strongest linkages between industries *i* and *j* by joint sourcing from the same supplier (buyer). R_{ij}^{SB} and R_{ij}^{BS} are measures of second-tier of linkages between industries *i* and *j*. If the ratio of $R_{ij}^{BB}(R_{u}^{BB})$ equals to 1 or 0, then industry *i* and *j* have the same supplying (buying) pattern.

In the calculation of the *R* measures, a linkage threshold, 001. 0 = α for purchases and sales are assigned to industry *i* and industry *j* (Feser, 2005). Key suppliers must account at least 0.1 percent of industry *i*'s to intermediate input purchases, while key buyers must also account for at least 0.1 percent of *i*'s intermediate sales. Selecting the maximum of the four *R* measures produces a new matrix (*R*max matrix).

$$R_{ij} = R_{ji} = \max\left[R(x_i, x_j), R(y_i, y_j), R(x_i, y_j), R(y_i, x_j)\right]$$

where, the *x* and *y* values are inter-industrial purchases and sales coefficients, respectively.

In Stage 4, the hierarchical clustering technique using the Ward's method (Ward, 1963) and applying squared Euclidean Distance as the distance or similarity measure to the *R*max matrix was carried out to identify the clusters. Ward's method uses an analysis of variance approach to evaluate the distances between clusters. The cluster membership was assessed by calculating the total sum of squared deviations from the mean of a cluster.

After identifying the industrial clusters, the study of the functions of industrial clusters is calculated based on the indexes of industrial clusters. The member industries of an industrial cluster are arranged in *m* columns and *m* rows, of which *m* is the number of all industries of industrial clusters and *n* is the number of industries of inter-industry transaction in Malaysia's input-output table, *n*=124. The final demand is from *n+1*,…,*l* and the primary input is from *n+1*,…,*h*. The functional types indexes of industrial clusters is defined based on the ratios as follows;

a) the ratio of inter-industry input to total output of an industrial cluster,

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} z_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{l} z_{ij}} \times 100\% \quad \text{or} \quad \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} z_{ij}}{\sum_{i=1}^{h}\sum_{j=1}^{m} z_{ij}} \times 100\%$$

where, z_{ij} is an inter-industry transaction

$\sum_{i=1}^{m}\sum_{j=1}^{l} z_{ij}, \sum_{i=1}^{h}\sum_{j=1}^{m} z_{ij}$ are the total outputs

b) the ratio of inter-industry output to total output of an industrial cluster,

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} z_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{l} z_{ij}} \times 100\%$$

c) the ratio of intra-cluster transaction to inter-industry input of an industrial cluster,

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} z_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}} \times 100\%$$

d) the ratio of intra-cluster transaction to inter-industry output of an industrial cluster.

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} z_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}} \times 100\%$$

Based on the four types of ratios, six types of functional industrial cluster can be identified.
1. Primary input-oriented industrial clusters if the ratio of (a) is less than 60%,
2. Inter-industry input-oriented industrial clusters if the ratio of (a) is greater than 60%,
3. Final demand-oriented industrial clusters if the ratio of (b) is less than 60%,
4. Inter-industry output-oriented industrial clusters if the ratio of (b) is greater than 60%,
5. Intra-cluster demand oriented industrial clusters if (c) is greater than (d),
6. Intra-cluster supply-oriented industrial clusters if (c) is less than (d).

4. SECTION 4

This section presents the findings as a result of performing the agglomeration coefficients hierarchical cluster analysis to identify industries that share similar economic characteristics, including inter-industry linkages.

Table 1. Number of Cluster, Agglomeration Coefficients and Slope of Agglomeration Coefficients

Stage	No. of Clusters in Solution	Agglomeration Coefficient	Slope (a)	Acceleration (b)
100	24	43.480	1.828	2.904
101	23	44.270	1.817	-.623
102	22	45.066	1.797	-1.129
103	21	45.918	1.888	5.080
104	20	46.781	1.882	-.279
105	19	47.648	1.850	-1.704
106	18	48.524	1.843	-.378
107	17	49.473	1.955	6.062
108	16	50.425	1.924	-1.588
109	15	51.396	1.926	.136
110	14	52.411	1.975	2.581
111	13	53.441	1.965	-.520
112	12	54.584	2.140	8.921
113	11	55.917	**2.442**	**14.117**
114	10	57.448	2.739	12.132
115	9	59.082	2.842	3.767
116	8	60.752	2.828	-.507
117	7	63.191	**4.015**	**41.979**
118	6	65.714	3.993	-.552
119	5	68.996	4.995	25.080
120	4	72.984	5.781	15.786
121	3	78.659	7.775	**34.487**
122	2	86.503	9.973	28.250
123	1	98.602	13.987	40.264

(a) The slope coefficient of the agglomeration schedule is the percentage change in the agglomeration coefficient given in the third column.

(b) Acceleration is the percentage change in the slope of the agglomeration coefficient, calculated as the percentage change in the number contained in the fourth column.

Table 1 presents only part of the agglomeration schedule of the cluster analysis performed for the Malaysia Input-Output Table for 2010. It comprises of only the final 24 of all 123 stages. The first column of the table displays the stage of the cluster solution, the second column lists the number of clusters in the solution and the third column gives the agglomeration coefficient. The fourth column shows the slope or the percentage change in agglomeration coefficient and the fifth lists the acceleration or the percentage change in slope coefficient. The change in the agglomeration coefficient is the most common indicator used as a guide to determine the clusters member of the agglomeration coefficient. It is the sum of the within-group variance of the two clusters combined at each successive stage.

The decision rule for selecting the candidate cluster solution is when there exists corresponding large changes in slope and acceleration of the agglomeration coefficient (Hill & Brennan, 2000). Therefore, there are three possible solutions identified: stage 113th with 11 clusters, stage 117th with 7 clusters and stage 121th with 3 clusters. Similarly the percentage change of the slope and acceleration is also high at the three stages. Among the three clusters, the 11-cluster is more homogeneous based on the characteristics of the industries.

The 3, 7, and 11-cluster solutions for the 124 industries is shown in Appendix A. The heterogeneity of the industry classification depends on the size of the clusters. Hence, when reading from right to left of Appendix A, we observe that the industrial clusters get more and more heterogeneous. On the other hand, homogeneity of the industrial clusters depends on the size of the cluster solution.

The relationship among the industries within a cluster is shown in Figure 1 as an example (Appendix B). The structures of industrial clusters have strong correlation and inter-dependent relationship among industries. The directions of arrowheads denote the input-output relationship among industries, while the type of line denotes the strength of the supplier-buyer relationship. The correlation coefficient of industries in which the value exceeds 0.3 is displayed in the figures.

The total output of an industrial cluster reflects the relative status of each industrial cluster in the economy. Appendix C.1 shows the total output, inter-industry input, inter-industry output, intra-cluster transactions of 11 industrial clusters. Cluster 3 recorded the largest cluster with 35 industries, while cluster 9 is the smallest cluster with 4 industries (residential, non-residential, civil engineering and special trade works). The largest contribution to the economy was cluster 3 (manufacturing) which contributed RM571 billion to the total output. The largest distribution of inter-industry input was cluster 6. Meanwhile, cluster 10 recorded the largest distribution for inter-industry output and intra-cluster transaction.

As shown in Appendix C.2, Cluster 8 registered a higher ratio (57.38%) of inter-industry input to total output of an industrial cluster. Therefore, cluster 8 is categorised as inter-industry input-oriented industrial cluster. Meanwhile, clusters 1-7 and 9-11 are categorised as primary input-oriented industrial cluster. Similarly, based on the ratio of inter-industry output to total output of an industrial cluster, cluster 1, 8 and 11 have higher ratios of inter-industry output. This indicates that they are inter-industry output-oriented industrial clusters. Cluster 2-7 and 9-10 have lower ratio of inter-industry output. Therefore, they can be considered as the final demand-oriented industrial cluster.

Measured in term of relative size, the ratio of intra-cluster transaction to inter-industry input and the ratio of intra-cluster transaction to inter-industry output of an industrial cluster, it was found that clusters 1, 3, 4, 7 and 10 are intra-cluster demand-oriented industrial clusters. The inter-industry economic links of an industrial cluster are relatively independent of industries outside the industrial cluster. Therefore, they are more specialized relative to industries outside the industrial cluster. Meanwhile, clusters 2, 5, 6, 8, 9 and 11 are intra-cluster supply-oriented industrial cluster. The member industries of the industrial cluster absorb more inter-industry output than industries outside the industrial cluster.

5. CONCLUSIONS

In conclusion, structural and functional analyses of industrial clusters provide information on the interconnecting values of chain of industries. Cluster analysis is useful to identify industries that have strong supply and buyer relationship. Functional analysis is, on the other hand, a useful technique to determine the orientation of the clusters. Hence, it is proposed that further research could include regional level data to identify industrial driver for each region.

REFERENCES

Akgungor, S., Kumral, N. and Lenger, A., 2003. National Industry Clusters and Regional Specializations in Turkey. European Planning Studies 11 (6), 647-659.

Bergsman, J., Greenston, P. and Healy, R., 1972. The Agglomeration Process in Urban Growth. Urban Studies 9, 263-288.

Bergsman, J., Greenston, P. and Healy, R., 1975. A Classification of Economic Activities Based on Location Patterns. Journal of Urban Economics 2, 1-28.

Bergman, E. and Feser, E., 1999. Industrial and Regional Clusters: Concepts and Comparative. The web-book of Regional Science. Regional Research Institute, West Virginia University, URL.://www.rri.wvu.edu/WebBook/Bergman-Feser/contents.htm.

Brachert, M., Titze, M. and Kubis, A., 2011. Identifying Industrial Clusters from a Multidimensional Perspective: Methodical Apects with An Application to Germany. Papers in Regional Science 90 (3), 419-439.

Chenery, H.B. and Watanabe, T., 1958. International Comparison of the Structure of Production. Econometrica XXVI (4), 487-521.

Czamanski, S., 1971. Some Empirical Evidence of The Strengths of Linkages Between Groups of Related Industries in Urban-Regional Complexes. Papers in Regional Science 7 (1), 137-150.

Czamanski, S. and De Q Ablas, L.A., 1979. Identification of Industrial Clusters and Complexes: A Comparison of Methods and Findings, Urban Studies 16, 61-80.

Dietzenbacher, E., Van der Linden, A.J. and Steenge, A.E., 1993. The Regional Extraction Method: EC Input-Output Comparisons. Economic Systems research 5 (2), 185-206.

Dietzenbacher, E., and Van der Linden, A.J., 1997. Sectoral and Spatial Linkages in the EC Production Structure. Journal of Regional Science 37 (2), 235-257.

Feser, E. and Bergman, E.M., 2000. National Industry Cluster Templates: A Framework for Applied Regional Cluster Analysis. Regional Studies 34 (1), 1-19.

Feser, E., 2005. Benchmark Value Chain Industry Clusters for Applied Regional Research. Regional Economics Applications Laboratory (REAL), University of Illinois at Urbana-Champaign.

Hill, E.W. and Brennan, J.F., 2000. A Methology for Identifying Drivers of Industrial Clusters: The Foundation of Regional Competitive Advantage. Economic Development Quarterly 14, (1), 65-96.

Hoen, A.R., 2002. Identifying Linkages with a Cluster-based Methodology. Economic Systems Research 14 (2), 131-146.

Hofe, R.V. and Bhatta, S.D. 2007. Method for Identifying Local and Domestic Industrial Clusters Using Interregional Commodity Trade Data, The Industrial Geographer 4 (2), 1-27.

Kelton, C.M.L., Pasquale, M.K. and Rebelein, R.P., 2008. Using the North American Industry classification System (NAICS) to Identify National Industry Cluster Templates for Regional Analysis. Regional Studies 42 (3), 305-321.

O'Huallachain, B., 1984. The Identification of Industrial Complexes. Annals of the Association of American Geographers 74 (3), 420-436.

Oosterhaven, J., Eding, G.J. and Stelder, D., 2001. Clusters, Linkages and Interregional Spillovers: Methodology and Policy Implications for the Two Dutch Mainports and The Rural North. Regional Studies 35 (9), 809-822.

Roepke, H., Adams, D. and Wiseman, R., 1974. A New Approach to the Identification of Industrial Complexes Using Input-Output Data. Journal of Regional Science 14 (1), 1-9.

Titze, M., Brachert, M. and Kubis, A. 2011. The Identification of Regional Industrial Clusters Using Qualitative Input-Output Analisis (QIOA). Regional Studies 45 (1), 89-102.

Rassmussen, P.N., 1956. Studies in Inter-Sectoral Relations, Amsterdam: North-Holland Publishing Company.

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58 (301), 236-244.

Zhu Yingming, 2010. Analysis of Industrial Clusters in China. Science Press. Appendix A : Map of Group of Industries in Malaysia, 2010 According to their Clusters

The Supply and Use Framework of National Accounts

Joerg Beutel
Konstanz University of Applied Sciences, Konstanz, Germany - beutel@htwg-konstanz.de

1. INTRODUCTION

The System of National Accounts 2008 - SNA 2008 (United Nations et al., 2009) provides a comprehensive conceptual and accounting framework for compiling the main macroeconomic data. Supply and use tables and institutional sector accounts are key structural elements of the SNA. The process of compiling these tables is a powerful way of ensuring consistency between the various data sources available to the compiler. The supply and use framework is also appropriate for calculating much of the economic data contained in the national accounts and detecting weaknesses. This is particularly important for the decomposition of values of flows of goods and services into prices and volumes for the calculation of an integrated set of price and volume measures.

The supply and use framework (Beutel, 2017) consists of two types of tables: supply and use tables and symmetric input-output tables. In principle, supply and use tables are rectangular tables with often many more products than industries. Most statistical offices aggregate the rectangular supply and use tables to symmetric supply and use tables. Transformation models are used to transform the symmetric supply and use tables into symmetric input-output tables with equal column and row totals.

The symmetric input-output tables constitute the data base for many macroeconomic models and different approaches for impact analysis. Input-output tables are therefore analytical constructs that inevitably involve some degree of modelling in their compilation. Symmetric input-output tables offer considerable advantages for many analytical purposes. As an analytical tool, input-output data are conveniently integrated into various macroeconomic models. Input-output analysis also serves a number of other analytical uses. However, the user of input-output data should keep in mind that rectangular supply and use tables are closer to actual observations and statistical sources than symmetric input-output tables.

In addition to the roles of the supply and use framework, this accounting framework serves also as a basis for various interconnections with satellite accounts, such as employment statistics, linkages with physical flows (land use, energy), linkages with other physical flows related to environmental issues (emissions, waste, sewage) and other forms of satellite systems (tourism, transport, health, education). Important examples for such systems are Social Accounting Matrix (SAM), System of Environmental-Economic Accounting (SEEA), Physical Input-Output Tables and Extended Input-Output Tables.

2. SUPPLY, USE AND INPUT-OUTPUT TABLES AS INTEGRAL PART OF THE NATIONAL ACCOUNTS

Supply and use tables are an integral part of the System of National Accounts (SNA). They play an important role as an integration framework of the national accounts. As a key feature of national accounts, supply and use tables provide the ideal concept for balancing supply and demand. The supply and use system is the best framework for compiling both GDP at current and at constant prices in an integrated approach. Supply and use tables constitute the centre piece of the internationally compatible accounting framework for a systematic and detailed description of the economy, its various components on the supply and demand side and its relations to other economies.

While supply and use tables are data-oriented in nature, the symmetric input-output tables are always constructed from having made certain analytical assumptions, usually from existing supply and use tables. The system recommends that the statistical supply and use tables should serve as the foundation from which the analytical input-output tables are constructed. This explains the importance given to the description of supply and use tables in the SNA 2008, while the more technical description of constructing symmetric input-output tables are shown in the forthcoming Handbook on Supply, Use and Input-Output Tables with Extensions and Applications (United Nations, 2017) of the United Nations Statistical Division.

SNA and data sources

Many basic sources are used to compile the national accounts. In general, a mixture of administrative records and statistical surveys is used as sources for the national accounts. The sample frame for the main surveys is determined by an Establishment Census or the General Business Register. The General Business Register is an important instrument of the statistical system. It comprises in principle all production units, listing names and addresses of all financial and non-financial corporations operating in the economy.

Basic structure of the SNA

The SNA includes four types of tables: goods and services account; institutional sector accounts; supply and use tables and analytical input-output tables. The SNA comprises supply and use tables and institutional accounts which ideally are balanced at the same time. For analytical purposes the supply and use tables are transformed to symmetric input-output tables. The institutional accounts include the goods and services account, the rest of the world account and the institutional sector accounts for non-financial corporation, financial corporations, general government, households, non-profit institutions serving households.

Supply and use framework

The supply and use framework is that part of the national accounts system which focuses on the production in an economy. It reflects the production of industries in which intermediate products and primary inputs (labour, capital, land) are required. Supply and use tables show where goods and services are produced and where they are used in intermediate consumption, final consumption, gross capital formation and exports.

The supply and use framework provides the most important macroeconomic aggregates such as GDP, value added, consumption, investment, imports and exports. The supply and use system is the adequate accounting framework for compiling consistent and reliable national accounts data. In consequence, it is recommended that the compilation of national accounts data - both in current prices as well as in constant prices - should be based on a supply and use framework.

Supply and use tables provide an ideal framework for checking the consistency of statistical data on flows of goods and services obtained from different sources. They also serve as a framework for economic statistics, both conceptually for ensuring the consistency of definitions and concepts used and as an accounting framework for ensuring numerical consistency of data obtained from different sources. The framework also allows basic economic data to be entered into the system in exactly the same structure in which the basic data can be surveyed and observed.

Gross domestic product

The annual estimates of gross domestic product (GDP) are obtained from the production, income and expenditure approaches and reconciled using supply and use tables. Some countries have a long tradition and much experience in utilising detailed production data based on local KAU (the establishment) as the statistical unit for compiling GDP estimates according to the production approach. The main classification schemes used are the activity classification based on The International Standard Industrial Classification of All Economic Activities (ISIC) and the product classification based on the Central Product Classification (CPC).

Supply, use and input-output tables

Supply and use tables manly serve statistical purposes. They are closer to the main sources of national accounts than input-output tables. They form a central part of the SNA and provide the framework for checking the consistency of statistics on flows of goods and services and primary inputs obtained from quite different statistical sources – industrial surveys, household expenditures inquiries, investment surveys, foreign trade statistics etc.

Supply and use tables provide a detailed picture of the supply of goods and services by domestic production and imports and the use of goods and services for intermediate consumption and final use (consumption, gross capital formation, exports). The use table also shows how the components of value added (compensation of employees, other net taxes on production, capital consumption, net operating surplus) are generated by industries in the domestic economy. Thus, supply and use tables provide detailed information on the production processes, the interdependencies in production, the use of goods and services and generation of income generated in production.

The supply and use tables also form the basis for deriving symmetric input-output tables by applying certain assumptions to the relationship between outputs and inputs. Symmetric input-output tables are the basis for input-output analysis. They mainly serve analytical purposes and are often used for impact analysis, macroeconomic models, price models, productivity analysis, etc. Input-output data also serve a number of other analytical purposes by linking other major statistics (employment, capital, energy, and environment) to the system of national accounts.

The classifications in the symmetric input-output tables match with those in the supply and use tables, as the former is a transformation of the latter. The symmetric input-output table is accompanied by a symmetric input-output table for domestic output and a matrix showing the use of imports. The classifications are totally compatible with those used within the framework of the United Nations. The supply and use tables and input-output tables of the SNA relate products and industries. The classification used for industries is the 'International Standard Industrial Classification of Economic Activities (ISIC) and the classification employed for products is the 'Central Product Classification (CPC).

A comprehensive system of supply, use and input-output tables includes the following tables:

SUPPLY AND USE TABLES AT PURCHASERS' PRICES

1. Supply table at basic prices, including a transformation into purchasers' prices
2. Use table at purchasers' prices

TRADE AND TRANSPORT MARGINS

3. Wholesale trade margins table
4. Retail trade margins table
5. Inland transport margins table
6. Water transport margins table
7. Air transport margins table

TAXES LESS SUBSIDIES ON PRODUCTS

8. Non-deductible VAT table
9. Taxes on products table
10. Subsidies on products table

SUPPLY AND USE TABLES AT BASIC PRICES

11. Supply table at basic prices
12. Use table at basic prices
13. Use table of domestic production at basic prices
14. Use table of imports at basic prices

SYMMETRIC INPUT-OUTPUT TABLES AT BASIC PRICES

15. Input-output tables at basic prices
16. Input-output tables of domestic output at basic prices
17. Input-output tables of imports at basic prices

Supply and use tables are rectangular tables with many more products than industries describing the domestic production process of industries and the transaction of products for intermediate and final use of the national economy in great detail. The valuation matrices for trade and transport margins and taxes less subsidies are also rectangular matrices corresponding to the outlay of the supply and use tables for products. The symmetric input-output tables are square matrices with the same number of products and industries and equal column and row totals for each product and corresponding industry.

3. COMPILATION OF SUPPLY AND USE TABLES

The compilation of supply, use and input-output tables is explained in detail in the Eurostat Manual of Supply, Use and Input-Output Tables (Eurostat, 2008). Supply and use tables are fully integrated into the SNA. The institutional approach to compile GDP with the production approach, the income approach and the expenditure approach are fully reflected in the supply and use tables. At the same time the supply and use tables are fully compatible with the sector accounts. They constitute the ideal framework for balancing the national accounts. Rectangular supply and use tables are the appropriate database for deflating GDP on a product basis if information on product prices is available at the more disaggregate level.

The supply table

The supply table shows the supply of goods and services by type of product of an economy for a given period of time. In practice, supply tables are large rectangular tables with many more products than industries. It distinguishes between the output of domestic industries and imports. The valuation matrices for trade and transport margins and taxes less subsidies on products allow a transformation of supply from basic prices to purchasers' prices.

The supply table comprises three major parts: domestic output at basic prices with specific information on primary and secondary activities, imports of goods and services and valuation matrix with vectors for trade and transport margins, value added tax other taxes on products and subsidies on products.In a first step total domestic output at basic prices and imports CIF are aggregated to total supply at basic prices. CIF (cost insurance, freight) values include the insurance and freight charges incurred between the exporter's frontier and that of the importer. In a second step the valuation vectors of trade and transport margins and taxes less subsidies on products are added to total supply at basic prices to arrive at total supply at purchasers' prices.

The use table

A use table shows the use of goods and services by product and by type of use for intermediate consumption by industry, final consumption expenditure, gross capital formation or exports. The use table also shows the components of gross value added by industry for compensation of employees, other taxes less subsidies on production, consumption of fixed capital, and net operating surplus. The use table contains three main matrices: intermediate consumption at purchasers' prices, final uses at purchasers' prices and value added at basic prices.

The use table has two main objectives. Firstly, it reveals by column the input structure of each industry. Secondly, it describes in the rows the use of different products and primary inputs (labour and capital). The costs of production are shown in the columns of the use table for each industry. The total output of an industry at basic prices corresponds to the total output of an industry as reported in the supply table. If the industry output is given and the intermediate consumption of products determined in the use table, value added of an industry can be estimated as a residual variable. However, if the main components of value added (compensation of employees, other net taxes on production, consumption of fixed capital) are known, net operating surplus is treated as the final residual variable. In the use table, total output and value added are recorded at basic prices.

The valuation matrices

In the supply and use system, the valuation concepts constitute an important element. Transactions are valued at the actual prices agreed upon by the seller and buyer. Market prices are thus the basic reference for valuation in the supply and use system. In the absence of market transactions, valuation is made according to costs incurred (non-market services produced by government) or by reference to market prices for analogous goods and services (services of owner-occupied dwellings).

The SNA 2008 distinguishes two main valuation concepts of the flows of goods and services: basic prices and purchasers' prices. The difference between these two basic valuation concepts relates to trade and transport margins on the one hand, and to taxes less subsidies on products on the other. Producers' prices were the main valuation concept in the former system of national accounts. When we also introduce the concept of producers' prices, the difference between these two valuation concepts can be attributed to the two factors.

The relationship between the different types of prices is the following:

Purchasers' prices (excluding any deductible VAT)
- Non-deductible VAT
- Trade and transport margins
= Producers' prices
- Taxes on products (excl. VAT)
+ Subsidies on products
= Basic prices

Trade and transport margins are the difference between purchasers' and producers' prices, and taxes less subsidies on products are the difference between producers' prices and basic prices. Even if the concept of producers' prices is not any more part of the system, the relationship still holds. However, the basic data which are used to compile the supply and use tables have different valuations. Production and output data are usually valued at basic prices, d Data on intermediate consumption and final use are usually valued at purchasers' prices, imports are valued at CIF-prices and exports are valued at FOB-prices. Without separating the different valuation components of the product flows, a supply and use framework cannot be balanced and made consistent. It is thus the task of the valuation matrices to bridge the differences between the valuation at purchasers' prices and the valuation at basic prices.

The valuation matrices comprise all flows that are related to the supply and use of trade and transport margins and the supply and use of taxes less subsidies on products. On the supply-side, valuation matrices are needed to transform supply from basic prices to supply at purchasers' prices and thus to be able to balance supply and use at purchasers' prices. On the use-side, valuation matrices are required to transform the use data from purchasers' prices into basic prices.

Balancing supply and use

One central feature of the System of National accounts is the balancing process of the system. Balanced macroeconomic data can be derived on a more aggregate level by applying the production, income and expenditure approaches. However, the enhanced option is to balance the system at the same time for the institutional sector accounts and the supply and use tables at a much deeper level of products and industries, both at current prices and at constant prices.

The balancing starts with compiling preliminary estimates of all required inputs to assemble a first assessment of the supply and use table at purchasers' prices and at basic prices. The required information contains production matrix at basic prices, use table at purchasers' prices, use table of imports at basic prices and the valuation matrices. Ideally this information is calculated for rectangular matrices with many more products than industries. At the start of balancing an estimate is available for every entry of the supply table, the use table and the valuation matrices. However, in spite of all efforts it has to be expected that inconsistencies in the estimates remain. The differences can be caused by inaccuracies and inappropriate margins in the preceding estimates, errors in the specification items, non-observed changes of inventories and simply calculation errors.

The conclusions for compiling supply and use tables as an integral part in the production of national accounts are the following:

- Supply and use tables are the most efficient way to incorporate all basic data – aggregated or detailed – into the national accounts framework in a systematic way.
- Supply and use tables effectively ensure the consistency of results at a detailed level and thereby improve the overall quality of the national accounts.
- Input-output tables mainly serve analytical purposes and are used as database for macroeconomic models.

4. TRANSFORMATION OF SUPPLY AND USE TABLES TO SYMMETRIC INPUT-OUTPUT TABLES

In a first step rectangular supply and use tables are aggregated to symmetric supply and use tables. In a second step these symmetric supply and use tables are transformed to symmetric input-output tables. Symmetric supply and use tables at basic prices constitute the data base which is required for the transformation of supply and use tables to symmetric input-output tables.

The database for the transformation includes the following tables: Supply Table at basic prices, Use Table at basic prices, Use Table for domestic output at basic prices and Use Tables for imports at basic prices.

Transformation models

Four basic models can be used for the transformation of supply and use tables to symmetric input-output tables. The transformation models are discussed in detail in Chapter 11 of the Eurostat Manual of Supply, use and Input-Output Tables (Eurostat, 2008).

Model A and Model B generate product by product input-output tables which are based on technology assumptions. The input-output tables show products and primary inputs in the rows as they were reported in the use table. However, a transformation is required for industries to homogenous units of production. Model C and Model D result in industry by industry input-output tables which are based on fixed sales structures assumptions. In these cases product rows have to be transformed to bundles of products (product groups) which are provided by industries, while the columns remain as in the use table.

The four basic transformation models are based on the following assumptions:

- Product technology assumption (Model A) Each product is produced in its own specific way, irrespective of the industry where it is produced.
- Industry technology assumption (Model B) Each industry has its own specific way of production, irrespective of its product mix.
- Fixed industry sales structure assumption (Model C) Each industry has its own specific sales structure, irrespective of its product mix.
- Fixed product sales structure assumption (Model D) Each product has its own specific sales structure, irrespective of the industry where it is produced.

The most frequently applied method for compiling product by product input-output tables is Model A. If countries decide to compile industry by industry input-output tables, most of the time Model D is applied. The implementation of Model A bears the risk of running into negatives. Model D avoids negative transactions but has only a loose link to product statistics.

5. SYSTEM OF ENVIRONMENTAL-ECONOMIC ACCOUNTING

Production and consumption activities of the economic system produce goods and services being used for consumption and capital formation and built up labour and man-made capital. However, at the same time these activities also produce joint products in form of residuals affecting the ecological systems. The ecological systems are endangered if the production of residuals is larger than the natural regeneration capacity of the ecological systems. With a rapidly growing world population recycling of residuals becomes more and more important. A full understanding of this process demands a complete understanding of the physical dimension of the economy.

Physical Supply and Use Tables (PSUTs) and Environmental-Extended Input-Output Tables (EE-IOTs) are used to describe the magnitude (measured by tonnes or other physical measuring units) and the nature of materials and products flowing in the economy, within the economy and between the economy and nature. They show how the natural resources (natural inputs) enter, are processed and subsequently, as products, are moved around the economy, used and finally returned to the natural environment in the form of residuals (emissions, waste, waste water, etc.). The exchange of products between the domestic economy and the rest of the world is also described in the PSUTs and EE-IOTs.

The System of Environmental-Economic Accounting – SEEA (United Nations et al., 2014) is fully integrated into SNA 2008. It represents a consistent set of accounts for the economic system and the environment. Contained are the internationally agreed standard concepts, definitions, classifications, accounting rules and tables for producing internationally comparable statistics on the environment and its relationship with the economy. The SEEA Central Framework encompasses measurement in three main areas: the physical flows of materials and energy within the economy and between the economy and the environment; the stocks of environmental assets and changes in these stocks; and economic activity and transactions related to the environment.

The measurement of the physical flows of materials and energy that enter and leave the economy and flow within the economy itself is a key focus of the SEEA. These physical flows are recorded in physical supply and use tables which are complements of the monetary supply and use tables of the SNA. Reported are natural inputs from the environment to the economy, for example, minerals, timber, water, and others. Physical flows within the economy are products, for example, agricultural goods, manufactured goods, electricity, and other goods. Flows from the economy to the environment are residuals of all kind, for example, waste, air emissions, sewage, etc.

Physical supply and use tables

The PSUTs follows as close as possible the structure and classifications of the SUTs. An important difference between the SUTs and PSUTs is that the PSUTs also include the physical flows not having a monetary value. The starting point for the PSUTs is the balanced SUTs as compiled as part of the National Accounts. With the physical flows having a monetary value, the flows not having a monetary value are added like natural inputs, waste and emissions. The integration of all physical flows in the PSUTs generates a consistent and coherent set of data which is also consistent with the monetary information contained in the monetary SUTs of the SNA.

Extended input-output tables

Extended Input-Output Tables (EIOTs) and models have become a powerful tool in supporting environmental and economic analyses and policies. If, for example, IOTs are extended to include environmental information, then this allows for establishing a solid foundation for environmental policy analysis.

Extended Input-Output Tables (EIOTs) comprise useful information of satellite systems which are integrated into the National Accounts. They often include information on investment, capital and labour. However information on energy, air emissions, waste, sewage and water is also needed and should be added to the tables.

The EIOT of Germany for the year 2009 incorporate seven additional satellite systems:

1. Input-Output Table (Mio. Euro)
2. Gross fixed capital formation (Mio. Euro)
3. Capital stock (Mio. Euro)
4. Employment (1.000 Persons)
5. Energy use (Petajoule)
6. Air emissions (1.000 tons)
7. Global warming, acid deposition and tropospheric ozone formation (1.000 tons)
8. Waste, sewage and water (1.000 tons, Mio. cbm)

The first part of the EEIOTs contains the traditional part of the National Accounts. This includes the domestic production of goods and services, taxes *less* subsidies on products and GVA at basic prices. The next sets of matrices below the IOTs are derived from satellite systems which are integrated into the input-output framework. These matrices provide useful information for the various industries on investment (for example, machinery and buildings), capital stock (for example, machinery and buildings) and employment (for example, number of wage and salary earners and self-employed). The following set of matrices of the satellite system contain information on energy consumption, emissions of gases and other residuals (for example, waste, sewage) by production and consumption activities. The first three matrices (IOTs, investment and capital stock) contain values, the next matrix for employment shows persons and the last four matrices represent quantities (labour, energy, emissions, waste, sewage and water).

REFERENCES

Beutel, Joerg (2017): Chapter 3 (The supply and use framework in the national accounts), in: Thijs ten Raa (ed.): Handbook of Input-Output Analysis, Edward Elgar Publishing Ltd, forthcoming.

Eurostat (2008): Eurostat Manual of Supply, Use and Input-Output Tables – 2008 edition, European Communities, Luxembourg.

United Nations et al. (2009): System of National Accounts 2008 - SNA 2008, New York 2009.

United Nations et al. (2014): System of Environmental-Economic Accounting 2012 – Central Framework (SEEA – CF), New York.

United Nations (2017): Handbook on Supply, Use and Input-Output Tables with Extensions and Applications, New York (forthcoming)

Processing Trade Activities: Measure and Contribution to Domestic Economy

M. Yusof Saari
Institute of Agricultural and Food Policy Studies, Universiti Putra Malaysia, Selangor, Malaysia – mysaari@gmail.com

Nur Adilah Hamid
Faculty of Economics and Management, Universiti Putra Malaysia, Selangor, Malaysia – nuradilahhamid@gmail.com

Chakrin Utit*
Institute of Agricultural and Food Policy Studies, Universiti Putra Malaysia, Selangor, Malaysia – uchakrin@gmail.com

Ibrahim Kabiru Maji
Faculty of Economics and Management, Universiti Putra Malaysia, Selangor, Malaysia – keibimaji@gmail.com

Abstract

Increased international outsourcings have spurred trade in intermediate goods, which is termed as processing trade. Malaysia is one of the countries that actively participating in the processing trade activities in the Asian region. While empirical evidences in other countries clearly indicate the unfavorable effects on local economies, empirical works to quantify the impact of increased international outsourcing in Malaysia are clearly lagged behind. The fact is that the conventional national accounts framework does not separate the processing trade activities in the calculation of gross domestic products (GDP). In this paper, we further improve the GDP accounting by separating the manufacturing sectors according to domestic and processing-trade sectors, and integrating them consistently in a so-called 'dualistic' input-output table. Results clearly show that the processing-trade sectors do not only associated with higher foreign content but also have lower growth linkages to the domestic economy. From the overall exports of processing-trade sectors, only 39.64% are benefited to the domestic value added while another 60.35% are attributed to foreign content.

Keywords: Processing trade; Input-output; Free industrial zones; Domestic content.

JEL Classification: F17; O24

1. INTRODUCTION

Rapid globalization has transformed the current international trade practices. Most of the goods nowadays are not produced in a single location, whereby the production activities have been fragmented and outsourced globally. From the design of the product to the manufacture of the components, assembly, processing, and packaging activities are distributed from one country to another country (Cadarso et al., 2008). As results, trade in parts, components and intermediate goods has grown at a faster pace as compared to the trade in final product (IDE-JETRO and WTO, 2011). In the past 14 years world exports of intermediate goods have tremendously increased by 93% from 1995 to 2009 (US$2,774 billion to US$5,373 billion), with an average growth rate of 4.8% per annum. In fact, Malaysia is ranked number four among the major Asian traders who own huge share of intermediate goods in the country trade account, constituted 68% of the country total exports and 72% of its total imports for non-fuel goods in 2009 (IDE-JETRO and WTO, 2011).

Widespread practices on international outsourcing, trade fragmentation and processing trade have raised concerns and debates on proper accounting approach of the trade as wells the impacts towards the value added, firms productivity, society, and environment (see Koopman, 2012; Zhang et al., 2012; Xia et. Al., 2015; Yu and Tian, 2012; Wang and Yu, 2012). This is because today's trade seems to be deviated from the normal trade practice, moving from "trade in goods" towards to the "trade in task" which encourage specialization of different economies in particular task that adds value along the production chain (IDE-JETRO and WTO, 2011). Therefore, the traditional practice in accounting for exports must be improved to reduce bias and misleading in interpreting the contribution of exports to the gross value of the goods instead of the amount that the country actually adds or contributes to the products (Zhang et al., 2012).

Given the considerable growth in outsourcing activities, it is not surprising that a sizeable amount of research has now been devoted to understand the causes and consequences of this disintegration of production. Empirical works across developing countries clearly show that international outsourcing is unlikely to have substantial contribution to the generation of gross domestic products (see Koopman et al., 2012) but rather enlarges income inequality (see Ho et al., 2005), lower productivity spillover (see Görg et al., 2008) and increases CO2 emission (see Su et al., 2013). In the context of employment, it generates less employment as compare to the normal export industries, mostly hires low skilled labors, and pays lowest wages to the workers (Feenstra and Hong, 2010; Chen et al., 2012; Wang and Yu, 2012; and Yu and Tian, 2012).

Malaysia is one of the Asian countries actively participate in processing trade activities with strong support from the government. To date, Malaysia has developed 35 areas of Free Trade Zones (FTZ) which are divided into 18 Free Industrial Zones (FIZ) and 17 Free Commercial Zones (FCZs). However, the impacts of processing-trade activities on the Malaysian economy cannot be completely accessible. This is because the current national accounts framework for the estimation of gross domestic product (GDP) does not make distinctions for export-oriented sectors into processing-trade activities and normal-trade activities. This lacking has two implications: (i) we do not know exactly the relative size and contribution of processing-trade activities on macroeconomic indicators, and (ii) we do not know the magnitude of spillover effects of processing-trade activities into the local economy. Thus, this paper develops a so-called 'dualistic' input-output table that separate an export-oriented sector into processing-trade and normal-trade activities. Based on the dualistic input-output table, we analyze the structural characteristics of the processing-trade activities and their contribution to the domestic economy.

2. EMPIRICAL EVIDENCES ON IMPACTS OF PROCESSING TRADE

The key idea behind the processing-trade is that countries increasingly link sequentially to produce goods. Processing-trade refers to final goods that produced using parts, components, accessories and packaging materials from abroad with free of duty. More formally, processing-trade occurs when: i) a good is produced in two or more sequential stages; ii) two or more countries provide value added during the production of the goods; and iii) at least one country must use imported product in its stage of the production process, and some of the resulting output must be exported. Note that processing-trade involves both import side and export side. On the import side, processing-trade is essentially a subset of intermediate goods trade. While all trades of intermediate goods are consistent with (i) and (ii), only the subset of intermediate goods imports that become embodied in exported goods is consistent with the third condition. On the export side, processing-trade can involve either intermediate goods or final goods.

This paper estimate and create a new dataset by disaggregating export into processing-trade and normal-trade sectors with the goal of assessing the value added content of processing-trade to the Malaysian economy. Thus, the global paradigm shift in international trade from trade driven by comparative advantage to trade driven by value chain and trade in task that revolutionize international trade worldwide is major motivation of this paper. In particular, trade in task has emerged to dominate processing-trade. Hence processing-trade is conceptualized as a production system where multinational companies are allocated in free industrial zones to import raw materials from abroad and further process it through re-packaging, re-assembling and incorporate additional value before exporting to other countries. This has encouraged specialization of various economies to different processing tasks called trade in tasks.

International outsourcing is now a growing phenomenon that led to increase in trade flows in intermediate inputs. This has contributed to the doubling of global export between 1995-2009. More interestingly, Asian exports of intermediate goods grew much faster at 7.2% compared to world average of 4.8%. As such, Asian region is a key player in international outsourcing with Malaysia is one the countries that actively participated. While empirical evidence has clearly revealed the effects of processing-trade on other economy, the empirical evidence on the impact of processing-trade on the Malaysian economy is still vague.

Most empirical findings suggest that processing-trade products contain low domestic value added and employment than the normal-trade does due to over reliant on foreign intermediate inputs (see for example, Wang and Yu, 2012; Zhang et al., 2012; Xing, 2012). For example, Koopman et al. (2012) reveal that processing exports in China dominate 50% of total exports in 2007 but contribute only 37% to domestic value added. Besides the increasing tendencies that processing trade activities may not show considerable contribution to the domestic value added, there is also the tendency that processing trade would increase pollution in the domestic economy. China's active participation in the global trade fragmentation activities has contributed the larger carbon dioxide emissions for industrial production which exceeded the consumption based carbon dioxide by 18.8% in 2005 (Lin and Sun, 2010).

Despite that processing-trade is consider beneficial in reducing unemployment and improves income re-distribution of workers, the processing firms in China are found to be least effective because they employ low skilled labors with lower wages. For instance Wang and Yu (2012), Yu and Tian (2012), Chen et al. (2012) estimated that \$1,000 of ordinary (normal) exports from China leads to 0.70 of workers employment per year, and \$1,000 of processing exports leads to 0.06 of workers employment per year. In fact, Feenstra and Hong (2010) and Chen et al. (2012) find that employment generated by processing-trade is much lower and three times smaller than employment by domestic sectors. As such, Wang and Yu (2012) conclude that processing-trade may help to increase demand and employment of unskilled labor in the domestic economy but the trade activities are specialized in low value added production segment despite of the various incentives given by the government.

3. DUALISTIC INPUT-OUTPUT TABLE

3.1. Accounting and processing trade activities

Input-output model has been widely used in the context of international trade analyses where the interactions between sectors within a single country as well as in multiple countries are explicitly taken into account (see for example, Koopman et al., 2008; Xing, 2012). In relation to the processing-trade, input-output table is considerably the most appropriate approach to account for value added and domestic content induced by processing exports. It does not only link explicitly the processing-trade industries with normal-trade industries but also confirms that all economic flows fulfill the accounting and adding-up constraints and thus ensuring the consistency of the table.

In standard input-output table, economic flows of processing-trade activities are "hidden" and consolidated in the "average" sectors. For example, export vector includes both exports by processing-trade sectors and normal-trade sectors. Specific economic flows on consumption of domestic intermediate inputs, imported intermediate inputs and value added generated by the processing-trade sectors are unknown. To account the economic flows of processing-trade sectors, an extended input-output with processing-trade is proposed. Table 1 provide the simplified framework of the extended input-output table.

Table 1. Simplified Framework of The Extended Input-Output Table with Processing-Trade

		Domestic production (D)	Processing trade (P)	Final use		Gross output
		1 . . n	1 . . n	Domestic (f)	Foreign (e)	
Production for domestic use and normal-trade (D)	1 . . n	$\mathbf{Z}^{DD}$	$\mathbf{Z}^{DP}$	$\mathbf{f}^{T}$	$\mathbf{e}^{D}$	$\mathbf{x}^{D}$
Production for processing trade (P)	1 . . n	0	0	0	$\mathbf{e}^{P}$	$\mathbf{x}^{P}$
Imported input		$\mathbf{m}^{MD}$	$\mathbf{m}^{MP}$	$\mathbf{f}^{M}$	$\mathbf{e}^{R}$	m
Value added		$\mathbf{v}^{D}$	$\mathbf{v}^{P}$			
Gross input		$\mathbf{x}'^{D}$	$\mathbf{x}'^{P}$			

Notes: gray color indicates flows that are not revealed in the standard input-output and they must be estimated.

There are two main features of this extended input-output table that differs from the standard input-output table. First, the "hidden" flows of processing-trade activities are now explicitly revealed. For example, we separate the domestic intermediate input (Z^T) into two users: for production of normal-trade (Z^{DD}) and for processing-trade (Z^{DP}). Similarly, imported intermediate inputs (m^T) are distributed into that used by production of normal-trade (m^{MD}) and that of processing-trade (m^{MP}). Second, this framework provides an estimation of value added generated by the processing-trade activities (v^P). In contrast, the standard input-output table "averages" the contribution of processing-trade activities in to the calculation of GDP.

Our framework of disaggregating processing-trade and normal-trade activities follows closely the framework of Koopman et al. (2012) and Chen et al. (2012). Accordingly, all production output that takes place in the free industrial zones are for exports only. In Malaysia, firms that operate in the free industrial zones allow, to some extents, to sell their products into the domestic economy with the permission of the authority. However, we could not measure outflows from free industrial zones to the domestic economy due to unavailability of data. This limitation is unlikely to affect our estimation because we believe volumes of output sold to domestic economy by the processing-trade activities are marginal.

3.2. Data sources

In Malaysia, processing-trade activities can be traced according to the following industrial locations: Free Industrial Zones (FIZ) and Free Commercial Zones (FCZ). In this paper, we reduce the scope by concentrating on the FIZ for two main reasons. First, dataset that available for the disaggregation of processing trade activities are superior for the FIZ and they can be linked with other databases. For example, data for processing-trade firms that captured in the 2010 Economic Census only cover for the manufacturing sector and these data serve as cross-references for other databases. Second, the manufacturing sector is the key drivers for the Malaysian economy, in particular the Electric and Electronic (E&E) sector and thus, analyzing processing-trade activities within the manufacturing sector would provide new insights on the issue of value added content in exports of the sector.

The main dataset used in the disintegration process of export sectors between normal-trade and processing-trade are derived from two major sources. The first dataset which is the list of firms that operate inside the FIZs are obtained from the Companies Commission of Malaysia (CCM). The dataset provides information about the ownership background, business activity, business address, balance sheet as well as profit and loss account for 1,988 processing-trade firms. However, only 1,670 firms are identified and grouped as processing-trade sectors after the data cleansing process. Second dataset for FIZ manufacturing firms are obtained from the Department of Statistics Malaysia (DOSM). This dataset only covers 50 firms and it provides information regarding the firms' output, input, value added, salary and assets. Using this dataset, validation process is conducted to justify whether the CCM revenue data are able to represent the output of processing trade sectors.

It should be noted that due to the data constraints, our model assumes all output that produced by the processing firms are sold to the foreign market as exports. Based on this assumption, output of processing trade (x^P) is equal to the exports (e^P). The similar assumption is used to estimate exports and output of processing-trade in China (see for example, Koopman et al., 2012) and in Mexico (see for example, De La Cruz et al., 2011). Our dualistic input-output table consists of 248 sectors while the standard input-output table has 124 sectors. Due to data constraints, not all sectors can be separated into domestic and processing-trade sectors. The 'best" separation only can be achieved for the 56 manufacturing sectors.

4. RESULTS AND DISCUSSION

4.1. Domestic value added and foreign contents in exports

In 2010, manufacturing sector contributes the largest amount of export which constitute 74.02% of the total country export, followed by services (17.33%), mining and quarrying (6.7%), and agriculture sectors (1.95%) (See Figure 1). Apparently, manufacturing sector plays an important role as one of the sources of Malaysia export growth as well as GDP at large. As the current trade accounting system is bias, it is crucial to re-evaluate the actual return of Malaysia export by taking into account the calculation of domestic value added content and foreign content embedded in export.

Figure 1. Malaysia Export, 2010

Figure 2 (a) presents the domestic value added content and foreign content in exports. From the total exports, the overall return of export earnings to the domestic value added is only 59.20% with 40.80% of foreign content embedded in the exports. This implies that from the total country export of RM644.53 billion in 2010, only RM381.56 billion is actually earned by the country. When we measure the manufacturing sector separately, we observe a much lower domestic value added share at 51.25% compared to foreign content share at 48.95%. Recall that the manufacturing sector is the biggest contributor to the Malaysian total exports with the value of RM477.08 billion in 2010. However, after properly measure the domestic and foreign contents, the actual gains from the manufacturing exports are only half of the total exports which equivalent to RM244.51 billion.

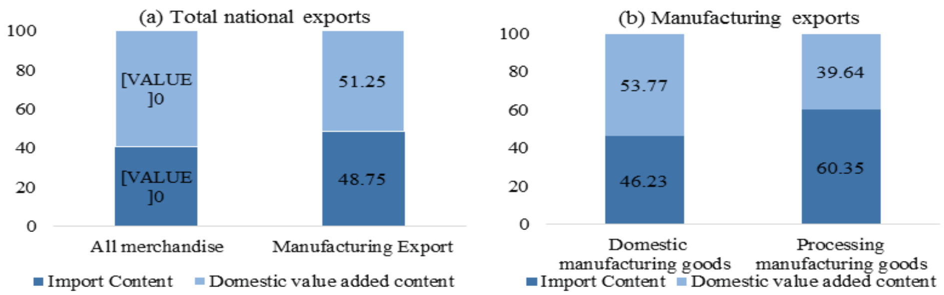


Figure 2. Domestic Value Added and Foreign Content in Exports, 2010 (%)

Considering the dualistic trade structures, the shares of foreign and domestic value added content shows significant differences, as indicated in Figure 2 (b). Manufacturing sectors that categorized under the processing trade regime are given special incentives and facilities for import and export activities. In fact, free duty import treatment encourage processing firms to use more foreign inputs and lower local sourced materials. As the result, foreign content for processing manufacturing sector found to be higher at 60.35% as compared to 46.23% for domestic structure. Hence, value added or export returns generated from the processing exports is much smaller only at 39.64%. Despite of various privileges and incentives provided by the government towards the processing trade firms, the value added or return of exports is smaller than the manufacturing firms operating under the domestic trade regime.

By looking at the export composition based on sectors, it is apparent that Electronic and Electrical (E&E) is the key for the Malaysian exports which contribute 28.37% of the total exports in 2010. But equally important to note, this sector also is the largest import consumer with 25.36%. The other national key export sectors within the manufacturing group are Foods (9.94%), Machineries (8.95%) and Petroleum Refineries (6.66%) sectors. It is also important to note that E&E sector contributes almost half of the export produce by processing sector. This implies that government promotes E&E sectors through the Free Industrial Zones facilities even though the import share of this sectors in processing trade structure is the highest at 60.80%.

4.2. Value added multipliers

The relative contribution of the processing trade sectors also can be assessed by examining the value added multipliers for every 1% increase in the final demand. For the non-processing trade sectors, the highest value added multiplier are sourced from Foods (RM4.13 billion), Electronic & Electrical (RM5.98 billion) and Machineries (RM2.35 billion). Conversely, for processing trade sectors, the highest value added are attained by Electronic & Electrical (RM1.33 billion), Foods (RM0.49 billion), and Chemicals (RM0.47 billion). On overall, results of the value added return show the capacity of processing trade sectors in generating the value added return is lower than the non-processing trade sectors. Our results are in line with the findings in the literature in which processing-trade sectors are found to have lower contribution to the domestic economy.

5. CONCLUSIONS

The development of the dualistic input-output table that separates the domestic and processing-trade sectors has provided a clear implication for the current trade policies. We have shown that the processing-trade sectors have associated with low growth linkages to the domestic sectors and contained higher foreign content. Ignoring the different characteristics of trade structures caused bias in measuring the country GDP as well as in the productivity measures of the economy. To emphasize, this study act an initial step that provides an overview on the export returns of the processing-trade activities that covers the Free Industrial Zones as the sample of the overall processing-trade "population". Due to data constraints, this study is unable to include Licensed Manufacturing Warehouse (LMW) firms that also receive the same privileges like the FIZ firms but more flexible in terms firms locations and located more disperse throughout Malaysia. Therefore, future studies that cover the both the LMW and FIZ may help in providing the bigger picture of processing-trade and international outsourcing in Malaysia.

Based on our analysis, we may argue that the processing-trade sectors or specifically Free Zones policies are less beneficial to the country in the sense that the export gained from this trade structure has limited implication on the generation of value added and limited linkages to the domestic economy. Almost half of the exports produced by Free Industrial Zones firms are generated by Electronic and Electrical (E&E) sector which has significant foreign content and lower domestic return. In fact, E&E under the domestic structure shows much similar trend in the context of domestic value added and foreign content which an arguable why FIZ should promote E&E processing-trade activities.

REFERENCES

Cadarso, M. Á., Gómez, N., López, L. A., & Tobarra, M. Á. (2008). The EU enlargement and the impact of outsourcing on industrial employment in Spain, 1993–2003. *Structural Change and Economic Dynamics*, 19(1), 95–108.

Chen, X., Cheng, L. K., Fung, K. C., Lau, L. J., Sung, Y. W., Zhu, K., Duan, Y. (2012). Domestic value added and employment generated by Chinese exports: A quantitative estimation. *China Economic Review*, 23, 850–864.

De La Cruz, J., Koopman, R. B., Wang, Z., & Wei, S. J. (2011). Estimating foreign value-added in Mexico's manufacturing exports. *US International Trade Commission Working Paper*, (2011-04A).

Feenstra, R. C., & Hong, C. (2010). China's exports and employment. China's Growing Role in World Trade, (March), 167–199.

Görg, H., Hanley, A., & Strobl, E. (2008). Productivity effects of international outsourcing: evidence from plant-level data. *Canadian Journal of Economics*, 41(2), 670-688.

Ho, L. S., Wei, X., and Wong, W. C. (2005). The effect of outward processing trade on wage inequality: the Hong Kong case. *Journal of International Economics*, 67(1), 241-257.

IDE-JETRO, & WTO. (2011). Trade patterns and global value chains in East Asia : IDE-JETRO.

Jayanthakumaran, K. (2003). Benefit–cost appraisals of export processing zones: A survey of the literature. *Development Policy Review*, 21(1), 51-65.

Koopman, R., Wang, Z., & Wei, S.-J. (2008). How Much of Chinese Exports is Really Made In China? Assessing Domestic Value-Added When Processing Trade is Pervasive. *National Bureau of Economic Research Working Paper Series*, No. 14109.

Koopman, R., Wang, Z., & Wei, S.-J. (2012). Estimating domestic content in exports when processing trade is pervasive. *Journal of Development Economics*, 99(1), 178–189.

Lin, B., & Sun, C. (2010). Evaluating carbon dioxide emissions in international trade of China. *Energy Policy*, 38(1), 613-621.

Su, B., Ang, B. W., & Low, M. (2013). Input–output analysis of CO 2 emissions embodied in trade and the driving forces: processing and normal exports. *Ecological Economics*, 88, 119-125

Wang, Z., & Yu, Z. (2012). Trading Partners, Traded Products and Firm Performances of China's Exporter-Importers: Does Processing Trade Make a Difference? *World Economy*, 35, 1795–1824.

Xing, Y. (2012). Processing trade, exchange rates and China's bilateral trade balances. *Journal of Asian Economics*, 23(5), 540–547.

Xia, X. H., Hu, Y., Chen, G. Q., Alsaedi, A., Hayat, T., & Wu, X. D. (2015). Vertical specialization, global trade and energy consumption for an urban economy: A value added export perspective for Beijing. *Ecological Modelling*, 318, 49-58.

Yu, M., & Tian, W. (2012). China's Firm-Level Processing Trade: Trends, Characteristics, and Productivity. Characteristics, and Productivity. China Center for Economic Research.

Zhang, J., Tang, D., & Zhan, Y. (2012). Foreign Value-added in China's Manufactured Exports: Implications for China's Trade Imbalance. *China and World Economy*, 20(1), 27–48.

CPS18:
Survey Sampling & Survey Method

Adaptive Complete Allocation Sampling

Mohammad Salehi M.*
Department of Mathematics, Statistics and Physics, Qatar University, P.O. Box 2713, Doha, Qatar– salehi@qu.edu.qa

Abstract

Adaptive sampling designs are becoming popular for surveying rare and clumped ecological populations. An adaptive sample enjoys modifying the sampling design on basis of obtained information during the survey while it remains in the probability sampling framework. Complete allocation sampling is an efficient and easily implemented adaptive sampling design which targets field effort to rare species and is logistically viable. The population is partitioned into strata and simple random samples are selected from strata in the first wave. If one rare species is observed in any selected unit all its stratum will be observed in the second wave. The condition of observing only one rare species in a unit to survey all units in its stratum might be too little for some populations. Salehi and Seber (2017) developed two-stage complete allocation sampling design in which the condition is to observe any given number of individual rare species rather than only one. The sampling design is also generalized to a two-stage sampling which allows researcher studies just a sample of primary sampling units rather than all of them. In this paper, we show that the Horvitz-Thompson estimator of adaptive complete allocation sampling is the Murthy's estimator of two-stage complete allocation sampling. Using a real-life population of buttercups, we show that how two stage complete allocation sampling broaden the application of adaptive complete allocation sampling design while observe more buttercups and its estimator can be considerably more precise compare to its conventional sampling design counterpart.

Keywords: Adaptive sampling, Horvitz-Thompson Estimator Murthy's Estimator, Rare and Clustered Population.

1. INTRODUCTION

Adaptive sampling designs are becoming popular for surveying rare and clumped ecological populations. Adaptive sampling designs (Thompson and Seber, 1996; Seber and Salehi, 2013) have gained popularity in environmental science for assessing animal and plant abundance when their populations are rare and clumped. They mimic how biologists would like to collect data. If the species is found at any location, biologists would search the vicinity of that location with hope of finding more rare species. In general, adaptive samplings can be categorized into two classes that we define here as: (1) adaptive searching and (2) adaptive allocation. The term adaptive searching refers to the designs such as adaptive cluster sampling (Thompson, 1990). Adaptive cluster sampling begins with an initial sample and, if the species of interest are detected in one of the selected units, then the neighbouring units are sampled. The neighbourhood is typically all units that share a side with the unit containing the species. If the species is further encountered in a unit in the neighbourhood, then the neighbourhood of that unit is also added to the sample, and so on, thus building up a network of units. If the initial sample includes a unit from a clump, then the rest of the clump (network) will generally be sampled. In adaptive cluster sampling, the allocation of extra effort in the immediate neighbourhood of where the species is found. In contrast, in adaptive allocation, extra effort is initiated once a collection of sample units is measured. This collection of sample units may be secondary units within primary units for two-stage sampling, or units within a stratum for stratified sampling. Examples of adaptive allocation sampling include two-phase adaptive stratified sampling (Francis 1984), two-stage sequential sampling (Salehi and Smith 2005), and adaptive two-stage sequential sampling (Brown et al. 2008).

Salehi and Brown (2010) introduced a simplified design to adaptive allocation of stratified sampling, called complete allocation stratified sampling. The proposed design can be considered a mix of both adaptive cluster sampling and adaptive allocation sampling. The appeal of adaptive cluster sampling is that for very rare and highly clustered populations, once a sample unit is found that meets the threshold criteria (for very rare populations, this threshold is usually one indicating the species is present), the biologist is reluctant to leave the sample unit and continue searching according to the fixed sample plan. Instinct is to search the immediate neighbourhood, and the appeal of the adaptive cluster sample design is just this (Brown and Manly, 1998). In effect, the immediate neighbourhood is searched as if it were a complete census. Note that the exactness of whether it is a census depends on the definition and shape of the neighbourhood. The less appealing aspect of adaptive cluster sampling is that it can be very difficult to conduct in the field, with logistical challenges of doing a complete search of a neighbourhood when the size and shape of the network are unknown. In complete allocation stratified sampling, any stratum that has first-phase results that meet a predetermined critical level is completely surveyed in the second phase. As in adaptive cluster sampling, a census is conducted near where the species is found, but unlike adaptive cluster sampling, the neighbourhood is predefined by size and shape of strata. The complete allocation stratified design is attractive because of the logistical simplification and the merging of what we consider the best features of the two classes of adaptive sampling.

Salehi and Seber (2017) has extended the complete allocation stratified sampling to Two-stage complete allocation sampling. They have generalized the complete allocation sampling design to a two-stage complete allocation sampling design. They had to use Murthy's estimator rather than Horvitz-Thompson estimator to derive an unbiased estimator and its variance estimator. The extension has two directions; firstly, the stratified sampling setting was extended to two-stage setting, which provides many more design options for researchers. In Two-Stage Complete Allocation (TSCA) sampling, one should observe all the strata, which results in a large final sample size. Particularly, when there is not enough time or budget to run a Complete Allocation sampling (CAS) design with a large effective sample size, one may use the TSCA design with selection of fewer primary units in the first stage. This should result in more rare species observed and an increase in precision. Secondly, the threshold of selecting all units in the selected primary unit was extended from observing just one rare unit to any number of rare units. This extension will broaden complete allocation sampling to rare and clumped populations that have few individuals of rare species scattered throughout the populations sites.

In this paper, we show that the Horvitz-Thompson of CAS is exactly the Murthy estimator of TSCA when all primary units are selected in the first stage and the threshold of selecting all units in the selected primary unit is just one rare units. We then discuss advantages of TSCAS over CAS.

2. NOTATIONS DESIGN, AND ESTIMATORS

Suppose that we have a population of N units, which are partitioned into M primary units of size N_i, ($i=1,2,\dots,M$), secondary units. A secondary unit may be a plot, or a quadrat, or a fisheries tow. One may preferably choose primary units sizes based on available information about rare clusters such as their sizes, their habitats, and natural restrictions. Let unit (i,j) denote the j^{th} secondary unit in the i^{th} primary unit with an associated measurement or the count of a species of interest of y_{ij}.

Let $\tau_i = \sum_{j=1}^{N_i} y_{ij}$ be the sum of the y values in the i^{th} primary unit, and let $\tau = \sum_{i=1}^{M} \tau_i$ be the total for the whole population. In the first stage of the sampling, we choose a simple random sample of m from the M primary units without replacement. At the second stage, we take an initial simple random sample of n_i secondary units without replacement from primary unit i, $i=1,2,\dots,m$. Let $C=\{y_{ij}|y_{ij}\}>c\}$, where c is a constant, be the condition that if satisfied for at least one unit in primary unit i, then all its units are selected. The population of secondary units in primary unit i, say U_i, is partitioned into two subpopulations according to whether the y-values satisfy the condition C or not. Let two subpopulations in primary i be denoted by $U_{ic}=\{y_{ij}|\ y_{ij}> c\}$ and $U_{ic'}=\{\ y_{ij}\ |\ y_{ij} \leq_c\}$, where $k_i=|U_{ic}|$ and $N_i-k_i=|U_{ic'}|$ are their cardinalities, respectively.

When $m=M$, we have a stratified complete allocation sampling, and if also $c=0$, we shall have the complete allocation sampling design introduced by Salehi and Brown (2010). They used the Horvitz-Thompson estimator to develop an unbiased estimator, but it is not possible to do so for the general condition of $y_{ij}>c$, where c is a constant. Salehi and Seber (2001) showed that one can use Murthy's estimator (1957) to devise unbiased estimators for such a sequential sampling design. Salehi and Seber (2017) derived Murthy's estimator for TSCA sampling design which is given,

$$\hat{\tau} = \frac{m}{M}\sum_{i=1}^{M}\hat{\tau}_i \qquad 1$$

where,

$$\hat{\tau}_i = \begin{cases} \frac{1}{F_i(0)} \sum_{j \in U_{ic}} y_{ij} + \frac{F_i(1)}{F_i(0)} \sum_{j \in U_{ic'}} y_{ij} & |s_i| = N_i \\ N_i \bar{y}_i & |s_i| = n_i \end{cases} \quad , \qquad F_i(x) = 1 - \frac{\binom{N_i - k_i - x}{n_i - x}}{\binom{N_i - x}{n_i - x}} .$$

It variance estimator is given by,

$$\widehat{Var}(\hat{\tau}) = M(M-m)\frac{S_\tau^2}{m} + \frac{M}{m}\sum_{i=1}^{M} \widehat{Var}(\hat{\tau}_i) \qquad 2$$

where, $s_\tau^2 = \sum_{i=1}^{m} (\hat{\tau}_i - \bar{\hat{\tau}}_i)^2 / (m-1)$ and

$$\widehat{Var}(\hat{\tau}_i) = \begin{cases} \frac{k_i}{F_i(0)} \left[(N_i - k_i)(1 - \frac{F_i(1)}{F_i(0)}) + k_i (F_i(2) - (1 - \frac{F_i(1)}{F_i(0)}) \right] \sigma_{ic}^2 + \\ \frac{(N_i - k_i) k_i}{F_i(0)} (1 - \frac{F_i(1)}{F_i(0)}) (\mu_{ic} - \mu_{ic'})^2 + \frac{(N_i - k_i)}{F_i(0)} \left[N_i - k_i \frac{F_i(1)}{F_i(0)} - \frac{(N_i - k_i)}{F_i(0)} \right] \sigma_{ic'}^2 & |s_i| = N_i \\ N_i^2 \left(1 - \frac{N_i}{n_i}\right) \frac{s_i^2}{n_i} & |s_i| = n_i \end{cases}$$

where s_i^2 is the sample variance from primary unit *i* and

$$\mu_{ic} = \frac{\sum_{j \in U_{ic}} y_{ij}}{k_i}, \quad \mu_{ic'} = \frac{\sum_{j \in U_{ic'}} y_{ij}}{N_i - k_i}, \quad \sigma_{ic}^2 = \frac{1}{k_i} \sum_{j \in U_{ic}} (y_{ij} - \mu_{ic})^2, \quad \sigma_{ic'}^2 = \frac{1}{k_i} \sum_{j \in U_{ic}} (y_{ij} - \mu_{ic'})^2$$

In complete allocation sampling, y-values are zero for those strata (primary units) in which the second wave of sampling are not carried out ($|s_i| = n_i$). From (1), we therefore have

$$\hat{\tau}_i = \begin{cases} \frac{1}{F_i(0)} \sum_{j \in U_{ic}} y_{ij} & |s_i| = N_i \\ 0 & |s_i| = n_i \end{cases}$$

Substituting into (1) and considering that *m*=*M* in CAS, we have,

$$\hat{\tau} = \sum_{i=1}^{M} \frac{\sum_{j \in U_{ic}} y_{ij}}{F_i(0)} = \sum_{i=1}^{M} \frac{\tau_i}{1 - \binom{N_i - k_i}{n_i} / \binom{N_i}{n_i}} = \sum_{i=1}^{M} \frac{\tau_i}{\pi_i} \qquad 3$$

Where $\pi_i = 1 - \binom{N_i - k_i}{n_i} / \binom{N_i}{n_i}$. Formula (3) is exactly formula (2) in Salehi and Brown (2010).

3. GAIN IN OBSERVING RARE EVENTS AND SPECIES

In this section, two-stage complete allocation sampling is applied to the Castle Hill buttercup site data (Fig. 1; Brown et al., 2013) to show the gain in observing more species (buttercups). Here the study site was partitioned into 12 primary units. There are 300 quadrates of size 100^2. The counts of buttercups are shown.

								2				3		
						5	2	2	2	1				
							1		1	1				
							3	2	3	1		1	1	
						6	7	1					1	
						1	3				2			
							2	2			2			
						3	5	20	2	2	1			
						3	5	6			2			
						5	13	11						
							16	2		1				
							4	1						
						1	2	2	2					
1														

Figure 1. The Castle Hill Buttercup Site Data. There are 300 Quadrates of Size 1002. The Counts of Buttercups are Shown

The effective sample size for the TSCA sampling is by

$$E_{CA}(\upsilon\ ) = \frac{m}{M}\sum_{i=1}^{M} N\pi_i + n(1-\pi_i)$$

where υ in the final sample size. Here, given a selected primary unit, we either choose all N units with a given probability or n with the complementary probability.

For the buttercups population, the expectation of effective sample sizes of CAS for $n = 2,3,\dots,10$ are given in Table 1.

Table 1. The Expectation of Effective Sample Sizes for Population of Figure 1

n	2	3	4	5	6	7	8	9	10
E(ν)	96.14	121.50	140.39	155.52	168.31	179.53	189.63	198.89	207.49

The minimum possible effective sample size for CAS design is 96.14 which is about one third of population size and it can be too large. If we have limited time or budget we may not able to use complete allocation design. However, the effective sample size can be reduced by 2/25 by taking only two primary units in the first stage. We can then have an effective sample size as small as 7.96 for TSCA sampling which makes the complete allocation practical.

The expected number of observed buttercups (say n_r^*) is given by

$$E_{CA}(n_r^*) = \frac{m}{M}\sum_{i=1}^{M}\pi_i y_{ic}^* + [\pi_i + (1-\pi_i)\frac{n}{N-k_i}]y_{ic\prime\prime}^*$$

Where $y_{ic}^* = \sum_{j\in U_{ic}} y_{ij}$ and $y_{ic\prime}^* = \sum_{j\in U_{ic\prime}} y_{ij}$. In order to study the gain in observing rare species (buttercups), We now compare this expected number with that of the conventional two-stage sampling with the first stage sample of size m and the second stage sample of size $n_1 = E_{CA}(\ \upsilon\ )/m$, which is given by,

$$E_{TS}(n_r^*) = \frac{m}{M}\frac{n_1}{N}\sum_{i=1}^{M}\tau_i = \frac{mn_i}{MN}\tau$$

The ratios of $E_{CA}(n_r^*)/E_{TS}(n_r^*)$ are drawn in Figure 2 for different thresholds, c = 0, 1,..., 5 and the initial sample sizes, n = 2, 3, ...,10. The results show that the highest gain is obtained when the threshold is *c=2* for this population. This means that by generalizing CAS to TSCAS we have chance to design a more

effective sampling design by choosing a right condition to run the second wave of sampling.
For TSCA sampling, the expected number of selected rare units is given by

$$E_{CA}(n_r) = \frac{m}{M}\sum_{i=1}^{M} \pi_i k_i.$$

This is the probability of being a selected primary unit multiplied by the number of rare units times the probability that the whole primary unit is observed, summed over all the primary units. We now compare this expected number with that of the conventional two-stage sampling with the first stage sample of size *m* and the second stage sample of size $n_1 = E_{CA}(\,v\,)/m$ which is given by,

$$E_{TS}(n_r) = \frac{m}{M}\frac{n_1}{N}\sum_{i=1}^{M} k_i.$$

The ratios of $E_{CA}(n_r)/E_{TS}(n_r)$ are drawn in Figure 3 for different thresholds, c = 0, 1, ...,5 and the initial sample sizes, n = 2, 3, ... ,10. The same as the number of buttercups case, the results show that the highest gain is obtained when the threshold is c=2 for this population. We can conclude that we may celebrate finding a clump of buttercups for carrying out the second wave of sampling when we find at least a unit with more than three buttercups in the initial sample.

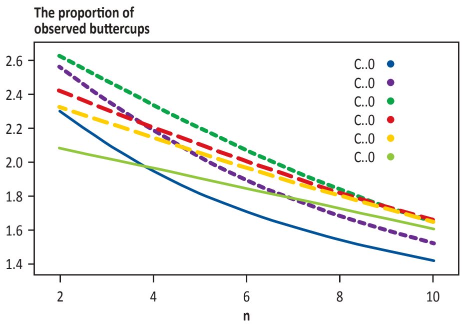


Figure 2. The ratios of expected number of buttercups for different condition and initial sample

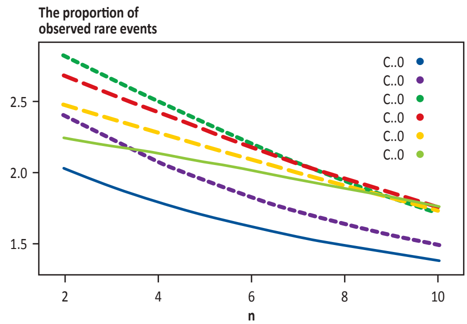


Figure 3. The ratios of expected number of rare events for different condition and initial sample

4. CONCLUSIONS

In complete allocation sampling, we should observe all the strata, which results in a large final sample size. Particularly, when there is not enough time or budget to run a CAS design with a large effective sample size, we may use the TSCA design with selection of fewer primary units in the first stage. This should result in more rare species observed and an increase in precision. Secondly, the threshold of $y_{ij} > 0$ was extended to $y_{ij} > c$ for any value of c. This extension will broaden complete allocation sampling to rare and clumped populations that have few individuals of rare species scattered throughout the populations sites. This also creates a challenge to find an optimal value for c.

REFERENCES

Brown J. A. (2003). Designing an efficient adaptive cluster sample. Environmental and Ecological Statistics, 10, 95–105.

Brown, J. A., Salehi, M. M., Moradi, M., Panahbehagh, B., and Smith, D. R. (2013). Adaptive survey designs for sampling rare and clustered populations. Mathematics and Computers in Simulation, 93, 108–116.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. Sankhy a, 18, 379–390.

Salehi, M. M., and Brown J. A. (2010). Complete allocation sampling: an efficient and easily implemented adaptive sampling design. Population Ecology, 52, 451–456.

Salehi M, Seber GA. (2017) Two-stage complete allocation sampling. Environmetrics.

e2441. https://doi.org/10.1002/env.2441

Salehi, M. M., and Seber, G. A. F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. Australian & New Zealand Journal of Statistics, 43, 901–906.

Seber, G. A. F., and Salehi, M. M. (2013). Adaptive Sampling Designs: Inference for Sparse and Clustered Populations. Heidelberg: Springer,

Salehi MM. and Smith (2005). Two-stage sequential sampling design: a neighborhood-free adaptive sampling procedure. Journal of agricultural, biological and environmental statistics, Volume 10, Number 1, Pages 84–103.

Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85, 1050–1059.

Thompson, S. K., and Seber, G. A. F. (1996). Adaptive Sampling. New York: Wiley.

Household's Balance Sheets Survey: Indonesia Case

Widyastuti Noviandari, Abdul Khalim, Ahmad Rasyid

Abstract

As an effort to maintain price stability and to promote the stability of financial system for sustainable economic growth, Bank Indonesia implementing various strategy, among others by conducting financial stability surveillance to monitor risks that potentially expose financial system stability. The risks may arise from the interconnectedness and interaction between economic players (financial markets, corporate sector, households as well as the banking and nonbank financial industries) as well as the behavior of each economic player.

The household sector plays an important role in the economy for many reasons and has significant exposure to financial stability, as investor/debtor (surplus units) and creditor (deficit units). The pressure on household balance sheets could potentially affect the performance of the financial sector and vice versa. Thus, surveillance of the household sector becomes important to monitor and measure the potential risks for the whole financial system.

Bank Indonesia regularly perform household sector surveillance, supported by various data source, among others Household's Balance Sheet Survey (SNRT). SNRT has been conducted regularly on annual basis since 2007 to provide information on the structure of household balance sheet in Indonesia, establishing useful database for surveillance system, to obtain data on the assets and liabilities of the household sector as an input in constructing financial balance sheet as well as indicators of financial imbalances. This paper discusses the methodology, implementation and results of survey conducted by Bank Indonesia in the year of 2016.

Keywords: Household, Risk, Surveillance, Survey.

Spatial Scan Statistics with a Restricted Likelihood Ratio for Ordinal Outcome Data [a]

Myeonggyun Lee*

Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea - mglee727@yuhs.ac

Inkyung Jung

Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea – ijung@yuhs.ac

Abstract

Spatial scan statistics are widely used as a technique to detect geographical disease clusters for different types of data such as Bernoulli, Poisson, ordinal, normal, and survival. The spatial scan statistic for ordinal data can be used to detect clusters indicating areas with high rates of more serious stages compared with the surrounding areas. It has been pointed out that the Poisson-based spatial scan statistic tends to detect the most likely cluster much larger than the true cluster by absorbing insignificant neighbours with non-elevated risk. We suspect that the spatial scan statistic for ordinal data might also have the similar undesirable phenomena. Tango (2008) proposed to modify the spatial scan statistic using a restricted likelihood ratio for scanning only the regions with elevated risk. The method worked well for preventing over-detection but was evaluated only in the Poisson model. In this paper, we propose to apply a restricted likelihood ratio into two spatial scan statistics to circumvent such a phenomenon in ordinal outcome data. Through a simulation study we compare the performance of the proposed methods with original spatial scan statistics. We calculate sensitivity, positive predicted value (PPV), usual power and bivariate power distribution as performance measures. The simulation study results show that the proposed spatial scan statistics with a restricted likelihood ratio have a reasonable or better performance compared with original ones. The original methods for ordinal data tend to detect larger clusters than the true cluster, and our approach seems to reduce the undesirable property. We illustrate the proposed methods using a real data set of the 2014 Health Screening Program of Korea with the diagnosis results of normal, caution, suspected disease, and diagnosed with disease as an ordinal outcome.

Keywords: Spatial scan statistic; Ordinal data; Cluster detection; Restricted Likelihood Ratio.

1. INTRODUCTION

For geographical surveillance, spatial scan statistics are widely used to detect spatial disease clusters in different types of data. This method proposed by Kulldorff (1997) is based on the likelihood ratio test to find the area with the maximum value of test statistics as a most likely cluster. The spatial scan statistics have been applied, for example, to examine spatial cluster for cancer disease (Gregorio *et al.*, 2006; Alvares *et al.*, 2009; Amin *et al.*, 2014).

However, Tango (2007) pointed out that the Poisson-based spatial scan statistic proposed by Kulldorff (1997) tends to detect an unrealistically larger cluster than the expected true cluster by absorbing adjacent regions with irrelevant risks. To resolve the undesirable trend, Tango (2008) proposed a Poisson-based spatial scan statistic by modifying the likelihood ratio. The method worked well for preventing such undesirable phenomena in detecting the true cluster compared with the original one.

In this paper, we focus on spatial cluster detection for ordinal data which are often obtained in nature such as cancer stage or grade. At this time, we are interested in geographical cluster detection of high rates of more severe categories (e.g., later stage or higher grade). For ordinal outcome data, there are two spatial scan statistics based on likelihood ratio ordering (LRO) and stochastic ordering (STO) proposed by Jung *et al.* (2007) and Jung and Lee (2012), respectively. We suspect that two proposed spatial scan statistics for ordinal data also tend to detect a larger cluster than the true cluster. Thus, we propose to apply a restricted likelihood ratio into two spatial scan statistics to circumvent such a phenomenon in ordinal outcome data.

2. METHODOLOGY

2.1. Spatial scan statistics for ordinal data

Suppose that a study area is composed of I sub-regions and the ordinal outcome variable has K categories. Let C_{ik} be the number of cases in the *i-th* region and the *k-th* category, where $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Since the categories are ordinal scale in nature, for example, a larger k reflects a more severe disease stage. The null hypothesis that there is no clustering can be expressed as $H_0: p_k = q_k$ $(k = 1,\ldots,K)$ for all scanning windows Z. The p_k and q_k mean the probability of being in category k within and outside the scanning window, respectively. Note that $\Sigma_k p_k = 1$ and $\Sigma_k q_k = 1$.

According to the type of order restriction, there are two methods for cluster detection on ordinal data. By Jung *et al.* (2007), an alternative hypothesis was considered as $H_a: \frac{p_1}{q_1} \le \frac{p_2}{q_2} \le \cdots \le \frac{p_K}{q_K}$, with at least one inequality being strict. This order restriction is called likelihood ratio ordering (LRO). Since the LRO does not incorporate all situations in which the probabilities of more severe disease categories are higher, Jung and Lee (2012) proposed another alternative hypothesis called stochastic ordering (STO) as $H_a: \sum_{k=1}^{j} p_k \le \sum_{k=1}^{j} q_k$ $(j = 1, \ldots, K)$, with at least one strict inequality. The LRO hypothesis is a special case of the STO hypothesis.

For the ordinal model, the likelihood function is written as

$$L(Z, p_1, \ldots, p_K, q_1, \ldots, q_K) = \prod_k \left(\prod_{i \in Z} p_k^{c_{ik}} \prod_{i \notin Z} q_k^{c_{ik}} \right),$$

where p_k and q_k are the unknown probabilities. The likelihood ratio test statistic can be expressed as

$$\lambda = \frac{\max_{Z,H_a} L(Z,p_1,\ldots,p_K,q_1,\ldots,q_K)}{\max_{Z,H_0} L(Z,p_1,\ldots,p_K,q_1,\ldots,q_K)} = \frac{\max_Z L(Z)}{L_0} \text{ with } L_0 = \prod_k \prod_i \hat{p}_{0k}^{c_{ik}} = \prod_k \left(\frac{C_k}{C}\right)^{\sum_i c_{ik}} = \prod_k \left(\frac{C_k}{C}\right)^{C_k},$$

where C_k $(= \sum_i c_{ik})$ is the sum of observations in category k, C $(= \sum_k \sum_i c_{ik})$ is the total number of observations in the whole study area and $\hat{p}_{0k}$ $(= \hat{q}_{0k}) = C_k/C$ is the MLE of p_k $(= q_k)$ under the null hypothesis, and with $L(Z) = \prod_k \prod_{i \in Z} \hat{p}_k^{c_{ik}} \prod_{i \notin Z} \hat{q}_k^{c_{ik}}$, where $\hat{p}_k$ and $\hat{q}_k$ are the MLEs of p_k and q_k under the alternative hypothesis in either LRO or STO. Jung *et al.* (2007) and Jung and Lee (2012) explained the details of how to obtain the MLEs under each alternative hypothesis.

2.2. Restricted likelihood ratio test statistics

To circumvent the over-detected phenomenon, Tango (2008) proposed a Poisson-based restricted likelihood ratio test statistic by taking each individual regions's risk into account. The concept of the restricted likelihood ratio is to use the indicator function on the significance of each region as a screening criterion. We also apply the same idea into two spatial scan statistics for ordinal data.

Given the pre-specified significnace level (α_1) for the individual region, we define a restricted likelihood scan statistic as

$$\lambda_{re} = \max_Z \left(\prod_k \prod_{i \in Z} \hat{p}_k^{c_{ik}} \prod_{i \notin Z} \hat{q}_k^{c_{ik}} \right) \left(\prod_k \left(\frac{C_k}{C}\right)^{C_k} \right)^{-1} \prod_{i \in Z} I(p\text{-}value_i < \alpha_1),$$

where the $p\text{-}value_i$ is obtained from the Pearson chi-square (χ^2) test for $H_{i0}: p_i = p_0$.

For the ordinal outcome, we calculate the $p\text{-}value_i$ from the Pearson χ^2 test to compare the proportions of cases ($\mathrm{p}_i = (\hat{p}_{1i}, \hat{p}_{2i}, \ldots, \hat{p}_{Ki})$) in each response category at each region with the whole study area. The Pearson χ^2 test has asymptotically a chi-square distribution with (K-1) degrees of freedom. Under the null hypothesis, the expected frequencies are found by multiplying each region size (c_i) by the proportions specified in the whole study area $\mathrm{p}_0 = (\hat{p}_{10}, \hat{p}_{20}, \ldots, \hat{p}_{K0})$.

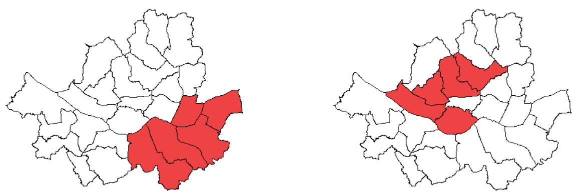
Even though χ^2 cannot completely reflect the ordinal scale, it is possible to distinguish the distinct regions compared with the whole area and thus the ordinal scan statistic is conclusively able to detect clusters. Introducing $I(p\text{-}value_i < \alpha_1)$ as a screening criterion for the ordinal scan statistics does not mean that we are performing multiple hypothesis tests.

3. SIMULATION STUDY

3.1 simulation data and setting

To compare the performances between the original and our proposed spatial scan statistics in the both the LRO and the STO hypotheses, we performed a simulation study under several scenarios. We considered the area of Seoul in South Korea an entire study area, which consists of 25 districts at the "Si-gun-gu" (city-county-district) level. All districts are geographically represented by a centroid coordinate. We assumed two different true cluster models with 5 districts in a circular (A) and an irregular (B) shapes. Figure 1 provides the map of the two true cluster models.

While assumed four disease categories, we considered two different STO and LRO alternative hypotheses each for each of two null hypotheses (see Table 1).



Cluster model A (circular) Cluster model B (irregular)

Figure 1. True Cluster Models A and B in the whole Study Area of Seoul

Table 1. The Scenario Details of Assumed Hypotheses

Null hypothesis	Alternative hypothesis
$H_0 : p = q$ = (0.25, 0.25, 0.25, 0.25)	H_{1a} : p = (0.20, 0.10, 0.40, 0.30)
	H_{1b} : p = (0.15, 0.15, 0.45, 0.25)
	H_{1c} : p = (0.10, 0.20, 0.30, 0.40)
	H_{1d} : p = (0.05, 0.25, 0.25, 0.45)
$H_0 : p = q$ = (0.30, 0.20, 0.30, 0.20)	H_{1a} : p = (0.25, 0.05, 0.45, 0.25)
	H_{1b} : p = (0.25, 0.05, 0.50, 0.20)
	H_{1c} : p = (0.15, 0.20, 0.30, 0.35)
	H_{1d} : p = (0.10, 0.15, 0.25, 0.50)

H_{1a}, H_{1b} : p = STO-based alternative hypotheses and H_{1c}, H_{1d} : LRO-based alternative hypotheses

Since it is difficult to find an asymptotic distribution for spatial scan statistic, we can evaluate statistical significance by using Monte Carlo hypothesis testing. We firstly generated 10,000 random data sets under each null hypothesis to estimate the critical values at a significance level of 0.05. The 500 highest values of the test statistics in the STO and the LRO methods were the critical values in each model. Then, we generated 1,000 random data sets for 8 different alternative hypotheses in each true culster model and searched clusters with high rates of higher categories with a circular scanning window. We set the screening significance levels (α1) at 0.10, 0.20, and 0.40. Based on the critical value at the level of 0.05, we computed the number of rejected data sets out of 1,000 which is the estimated power of the tests called usual power. To assess the accuracy of detected clusters, we considered sensitivity and positive predicted value (PPV) as

$$\text{Sensitivity} = \frac{number\ of\ districts\ correctly\ detected}{number\ of\ districts\ in\ the\ true\ cluster}$$

$$\text{PPV} = \frac{number\ of\ districts\ correctly\ detected}{number\ of\ detected\ districts}.$$

We calculated the average of proportions only for rejected data sets at a significance level of 0.05 in both sensitivity and PPV. A larger value of these measures means that the method is more precise for detecting the true cluster. For example, a lower value of PPV means that the method tends to detect larger clusters than true clusters. In case of a lower sensitivity, the method may miss more regions in detecting the true cluster.

Meanwhile, because the usual power estimates rejected data sets under the null hypothesis of no clustering, it does not reflect the precision of correctly detecting a cluster when data sets are rejected. Tango and Takahashi (2005) proposed a bivariate power distribution based on Monte Carlo simulations in order to compare the power performance of spatial scan statistics. The bivariate power distribution of $P(l,s)$, which is defined by the length l of the significant clusters and the number s of the regions identified out of the assumed s^* regions in a true cluster, can be expressed by

$$P(l,s) = \frac{\#\{significant\ clusters\ have\ length\ l\ and\ include\ s\ true\ regions\}}{trials\ for\ each\ simulation}$$

Where $1 \leq l, 0 \leq s \leq \min\{l, s^*\}$. The bivariate power distribution can be a good measure to compare the performance, especially in terms of the over-detection problem. For instance, if the number of regions in true cluster is 3, $P(l > 3, s = 3)$ indicates over-detection. For each simulation with the use of 1,000 replications at significance levels of 0.05, we presented the bivariate power distribution $P(l,s) \times 1000$ for a more intuitive understanding.

3.2. Simulation Results

Tables 2 through 4 show the estimated power, sensitivity, and PPV in each cluster model with different hypotheses. Also, the bivariate power distribution $P(l,s) \times 1000$ in one scenario for each model is shown in Table 5 as an example.

As we expected, the STO-based method seems to perform better when compared with the LRO-based method under the STO hypothesis, while both methods have similar capacities of sensitivity, PPV and usual power under the LRO hypothesis. The overall pattern showed similar results to Jung and Lee (2012). However, as seen in Table 4, the original scan statistics always have lower PPV than the restricted scan statistics regardless of their conditions, in particular, in the irregular true cluster model B. This can be interpreted that the original spatial scan statistics tend to detect larger clusters than the true cluster. But, our restricted spatial scan statistic tends to have a little lower sensitivity and power. For more details, in Table 5, the original methods show that the estimated bivariate power distribution have a long tail which is an undesirable phenomenon. For example, there are many cases with nine or more regions in the detected cluster using the original methods, while the true cluster consists of only five regions.

Table 2. The Estimated Power of STO-based and LRO-based Methods in Cluster Models

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
A	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	100.00	99.70	100.00	100.00	100.00	98.80	99.90	100.00
		H_{1b}	100.00	100.00	100.00	100.00	100.00	99.10	99.90	100.00
		H_{1c}	100.00	99.70	100.00	100.00	100.00	100.00	100.00	100.00
		H_{1d}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	100.00	100.00	100.00	100.00	100.00	99.30	100.00	100.00
		H_{1b}	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00
		H_{1c}	100.00	97.90	99.60	100.00	100.00	99.50	100.00	100.00
		H_{1d}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
B	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	100.00	99.30	100.00	100.00	100.00	95.70	98.90	99.70
		H_{1b}	100.00	99.40	99.80	100.00	100.00	97.20	98.60	99.50
		H_{1c}	100.00	99.60	100.00	100.00	100.00	100.00	100.00	100.00
		H_{1d}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	100.00	100.00	100.00	100.00	100.00	99.30	99.70	100.00
		H_{1b}	100.00	100.00	100.00	100.00	100.00	99.30	99.90	100.00
		H_{1c}	100.00	94.40	99.40	99.70	100.00	98.80	99.90	100.00
		H_{1d}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 3. The Estimated Sensitivity of STO-based and LRO-based Methods in Cluster Models

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)		0.4	Original (%)	Restricted (%)		0.4
				0.1	0.2			0.1	0.2	
A	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	97.92	67.04	82.76	93.56	96.70	64.92	80.98	92.02
		H_{1b}	98.26	67.30	80.82	92.64	97.44	64.00	77.84	91.04
		H_{1c}	98.40	68.00	82.72	94.08	98.40	69.76	83.56	94.34
		H_{1d}	99.74	94.38	98.44	99.64	99.74	94.72	98.48	99.64
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	98.94	77.92	90.32	96.98	97.30	73.43	87.74	95.10
		H_{1b}	99.14	84.26	93.94	98.40	97.40	80.32	91.46	96.38
		H_{1c}	97.12	56.63	73.82	88.00	97.24	58.15	75.20	88.98
		H_{1d}	99.92	98.26	99.20	99.82	99.92	98.30	99.20	99.84
B	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	76.40	53.90	60.42	63.28	77.16	49.63	57.01	61.00
		H_{1b}	84.10	53.02	58.36	64.10	81.28	50.45	54.83	60.96
		H_{1c}	78.26	54.78	60.32	63.42	78.78	58.72	63.36	66.44
		H_{1d}	80.48	80.46	79.92	79.08	83.08	86.48	86.14	84.00
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	77.58	62.96	67.20	68.34	77.96	53.64	57.97	61.24
		H_{1b}	78.56	66.18	67.92	69.34	79.34	56.92	60.46	63.56
		H_{1c}	77.28	46.25	53.82	60.54	77.76	49.33	57.76	63.74
		H_{1d}	91.66	89.74	89.62	89.20	93.80	93.14	93.22	92.68

Table 4. The Estimated PPV of STO-based and LRO-based Methods in Cluster Models

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)		0.4	Original (%)	Restricted (%)		0.4
				0.1	0.2			0.1	0.2	
A	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	94.91	98.90	98.62	98.07	98.18	99.63	99.50	99.12
		H_{1b}	92.39	98.78	98.43	97.74	98.32	99.23	99.13	98.99
		H_{1c}	98.87	99.63	99.60	99.35	98.93	99.74	99.66	99.41
		H_{1d}	99.62	99.76	99.81	99.72	99.86	99.97	99.93	99.89
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	91.26	98.63	98.63	98.06	97.78	99.67	99.58	99.09
		H_{1b}	80.19	98.17	97.83	96.57	97.64	99.12	98.94	98.68
		H_{1c}	98.63	99.52	99.54	99.18	98.75	99.81	99.76	99.33
		H_{1d}	99.80	99.83	99.81	99.81	99.85	99.93	99.91	99.90
B	p=q =(0.25,0.25,0.25,0.25)	H_{1a}	80.08	98.57	98.19	96.69	78.54	99.13	98.60	96.88
		H_{1b}	72.27	98.33	97.87	94.78	74.50	99.25	98.30	95.26
		H_{1c}	80.06	99.10	99.04	97.24	80.04	99.44	99.23	97.39
		H_{1d}	86.53	99.84	99.61	98.57	86.87	99.89	99.64	98.63
	p=q =(0.30,0.20,0.30,0.20)	H_{1a}	80.84	98.49	97.77	96.25	77.64	99.26	98.73	96.69
		H_{1b}	79.29	97.72	97.23	94.30	77.16	99.01	98.19	95.52
		H_{1c}	78.33	98.97	98.50	96.61	78.64	99.17	98.81	97.15
		H_{1d}	74.70	99.55	99.05	95.43	74.98	99.70	99.15	95.68

Table 5. Estimated Bivariate Power Distributions $P(l, s)\times 1000$ of STO-based Methods for H_0 : p = q = (0.25,0.25,0.25,0.25) and H_1 : p = (0.15,0.15,0.45,0.25) in Cluster Model B

STO-based original scan statistic						
Length	Include strue regions					
l	0	1	2	3	4	5
1-4	0	2	23	275	0	0
5	0	0	0	5	15	0
6	0	0	0	12	0	0
7	0	0	0	4	85	0
8	0	0	0	2	7	2
9	0	0	0	1	0	353
10	0	0	0	0	2	137
11	0	0	0	0	3	37
12	0	0	0	0	1	9
13	0	0	0	0	1	8
14	0	0	0	0	1	5
15-18	0	0	0	0	1	7
usual power = 1.000						

STO-based restricted scan statistic							
α_1	Length	Include strue regions					
	l	0	1	2	3	4	5
0.10	1-2	0	95	0	0	0	0
	3	1	4	251	0	0	1
	4	0	0	6	491	0	0
	5	0	0	5	18	108	0
	6	0	0	0	6	7	0
usual power = 0.994							
0.20	1-2	0	44	0	0	0	0
	3	0	0	159	0	0	0
	4	0	0	4	575	0	0
	6	0	1	2	23	154	0
	7	0	0	2	10	13	0
	8	0	0	0	4	4	0
usual power = 0.998							

4. APPLICATION

We applied two approaches, the original and our proposed spatial scan statistics for the ordinal data, to real data in the 2014 Health Screening Program by the National Health Insurance Servie (NHIS) of South Korea. To improve regional health and medical service, NHIS annually provides dataset including general health screening (NHIS, 2014). We used the data set of results on first diagnosis based on general health screening by district for male in 2014 as an ordinal data (see Table 6).

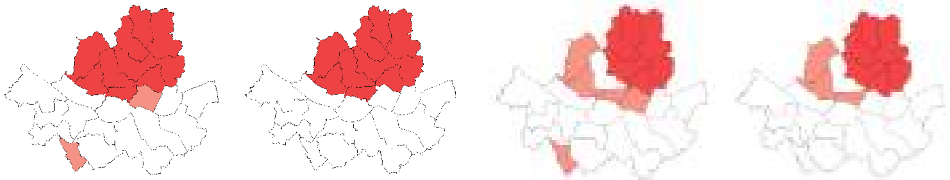
Table 6. Data on the diagnosis of general health screening for male in Seoul (2014)

	Level of diagnosis	N	%
1	Normal	55,891	4.97
2	Caution	369,425	32.84
3	Suspected disease	467,957	41.60
4	Diagnosed with diseases	231,599	20.59
	Total	**1,124,872**	**100.00**

The maximum size of the scanning windows of each location in this study is set to include 50% of the total observations in Seoul. We compare the results of original spatial scan statistics with that of the proposed spatial scan statistics using the value of $\alpha_1 = 0.10$. We evaluate the statistical significance for clusters via 9,999 replications for Monte Carlo simulations at 0.05 of significance level.

Figure 2 shows the result map for spatial cluster detection on the level of first diagnoses on the general health screening in Seoul. We found that the original cluster detection methods detected larger clusters than our proposed methods, in particular, on the most likely cluster. While the original scan statistic for ordinal data detected a large cluster in the north area, our method detected two or three small clusters excluding "Jonglo-gu". We think that the original method might have detected larger clusters by absorbing adjacent regions with irrelevant risks, "Jonglo-gu" in this case. STO-approach additionally detected "Sungdong-gu" and "Keumchun-gu", while they were not detected by LRO-approach due to order restriction.

We presented the detailed information about all statistically significant clusters in Table 7. The most likely cluster detected using the original method consists of ten districts centered on "Gangbuk-gu", while only six districts belong to the most likely cluster detected by the proposed method.



(a) STO cluster map (b) LRO cluster map (c) Restricted STO cluster map (d) Restricted LRO cluster map

Figure 2. Spatial Cluster Detection Results for the General Health Screening for Male in Seoul

Table 7. Spatial Clusters of High Rates of the Diagnoses in Figure 2 (a) and (c)

	Cluster Level		Centroid (Gu)	No. of Districts	LLR	p-value	Total cases	Percent cases in area ([1],[2],[3],[4])
STO	Primary Secondary	1	Gangbuk	10	1239.14	0.0001	399,525	(4.41,31.09,41.65,22.84)
		2	Keumchun	1	13.18	0.0007	33,399	(5.03,31.81,42.35,20.81)
		3	Sungdong	1	8.37	0.0228	30,761	(5.24,31.75,43.03,19.99)
Restricted STO	Primary Secondary	1	Nowon	6	851.60	0.0001	281,474	(4.45,30.80,41.79,22.96)
		2	Eunpyung	2	188.19	0.0001	89,643	(4.32,31.70,41.07,22.90)
		3	Joong	1	28.89	0.0001	13,185	(4.19,30.73,42.81,22.26)
		4	Keumchun	1	13.18	0.0001	30,761	(5.24,31.75,43.03,19.99)
		5	Sungdong	1	8.37	0.0067	33,399	(5.03,31.81,42.35,20.81)

5. CONCLUSION

The purpose of this study was to propose modified spatial scan statistics for ordinal outcome data by considering a restricted likelihood ratio in order to resolve the undesirable over-detection phenomenon. Our simulation study results showed that the original spatial scan statistics tended to detect clusters larger than the true clusters on ordinal outcome data. Proposed spatial scan statistics seemed to relieve that undesirable property; they have a good performance with a high value of PPV compared with the original method. In our application, we discovered the similar patterns with simulation results.

Even if our restricted spatial scan statistics tends to have lower sensitivity and power, we can resolve this by adjusting the screening value ($\alpha 1$). In other words, sensitivity and power can be higher when we have the appropriate screening value and this can be advantageous in our method. Even though our proposed method tends to overlook some true regions when the screening value is very low, it scans only the regions with elevated risk.

REFERENCES

Alvarez, P. L. L., Guerra-Yi, M. E., Faes, C., Alvarez, Y. G., & Molenberghs, G. (2009). Spatial analysis of breast and cervical cancer incidence in small geographical areas in Cuba, 1999–2003. *European Journal of Cancer Prevention*, *18*(5), 395-403.

Amin, R., Hendryx, M., Shull, M., & Bohnert, A. (2014). A cluster analysis of pediatric cancer incidence rates in Florida: 2000–2010. *Statistics and Public Policy*, *1*(1), 69-77.

Gregorio, D. I., Samociuk, H., DeChello, L., & Swede, H. (2006). Effects of study area size on geographic characterizations of health events: prostate cancer incidence in Southern New England, USA, 1994–1998. *International Journal of Health Geographics*, *5*(1), 1.

Jung, I., Kulldorff, M., & Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in medicine*, *26*(7), 1594-1607.

Jung, I., & Lee, H. (2012). Spatial cluster detection for ordinal outcome data. *Statistics in medicine*, *31*(29), 4040-4048.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, *26*(6), 1481-1496.

National Health Insurance Service. (2014). The 2014 National Health Screening Statistical Yearbook. (http://www.nhis.or.kr/menu/boardRetriveMenuSet.xx? menuId=F3328)

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, *4*(1), 1.

Tango, T. (2007). A spatial scan statistic scanning only the regions with elevated risk. *Advances in Disease Surveillance*, *4*, 117.

Tango, T. (2008). A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics*, *29*(2), 75-95.

Efficient Stratification Method for Socio - Economic Survey in Remote Areas

Adhi Kurniawan
Directorate of Census and Survey Methodology Development, BPS - Statistics Indonesia, Jakarta, Indonesia – adhikurnia22@gmail.com

Atika Nashirah Hasyyati*
Directorate of Finance, IT, and Tourism Statistics, BPS - Statistics Indonesia, Jakarta, Indonesia – atikanashirah@gmail.com

Abstract

The problems that exist in implementing a sampling design for socio-economic surveys in remote areas in Indonesia are high cost of the survey, low response rate, and less accurate. Therefore, the sampling design needs to be developed, one of which is to improve the efficiency of the stratification procedure. Stratification of census block in remote areas can be developed by combining the strata of welfare concentration and the strata of geographic difficulty by simulating the various alternatives number of strata and the various alternatives sample allocation. The strata of welfare concentration and the strata of geographic difficulty are constructed by Polychoric Principal Component Analysis. The strata of welfare concentration aim to improve statistical efficiency, while the strata of geographic difficulty are used to improve cost efficiency. The estimation procedure is performed at the domain level and population level. The simulation study focus on Papua Province by using the 2010 Population Census data and the 2011 Village Potency data. Some sampling scenarios can be categorized into four quadrants, the first quadrant with small sampling variance and low cost, the second quadrant with big sampling variance and low cost, the third quadrant with big sampling variance and high cost, and the fourth quadrant with small sampling variance and high cost. Based on these simulation results, several alternative scenarios of efficient stratification with small sampling variance and low cost of the survey are obtained.

Keywords: Statistical efficiency; Cost efficiency; Strata of welfare concentration; Strata of geographic difficulty

JEL Classification: C83, D63, I31

1. INTRODUCTION

Survey design has the important role in influencing the quality of data. There are two aspects that have to be considered in designing survey methods, namely statistical efficiency and the cost efficiency. Statistical efficiency means that the survey design must ensure that the estimator provided by the survey has the appropriate result regarding the reliability criteria. Practically, it is common to use the variance estimation to measure this criteria. On the other hand, cost efficiency also must be considered by the survey designer because usually cost is the commonly important constraint when conducting survey data collection. Therefore, the sample distribution have to be designed by considering two aspects of efficiency as in Kish (2004) stated that variance and cost factors are best to be viewed together.

In order to make the good statistical inference and fulfill the term of representativeness, the sample distribution must be spread out over the areas of population, including the remote areas. In this case, the remote areas issues in Indonesia must be concerned due to the heterogenous characteristics of the areas. Previous experiences in the social economic survey show that there are several problems regarding the samples existence in the remote area, such as the unpunctuality of collecting data, time schedule, accessibility problems, low response rate, and the high cost to access the areas. Therefore, it is very necessarry to develop alternative sampling method that can enhance statistical efficiency and cost efficiency in the domains which consist of the large number of remote area.

One of the most important point that can be improved in sampling design in regard with increasing statistical efficiency and cost efficiency is stratification methods. Stratification means that the population of N units is divided into subpopulations or known as strata (Cochran, 1965). The most common reason of stratification implementation in probability sampling is that stratification can increase efficiency (Rossi, Wright, and Anderson, 2013). Stratification methods can be used to decrease the variances so that there will be homogeneity within strata (Kish, 1965). Several factors that influence stratified sampling efficiency of the estimator of population parameters are the choice of stratification variable, number of strata, determination of strata boundaries, and allocation of sample sizes to the different strata (Verma, Joorel, and Agnihotri, 2012).

This paper propose alternative methods of stratification, especially for the socio-economic survey, by combining the concentration welfare stratum and geographic difficulty stratum and simulating the various alternatives the number of stratum. Stratification is constructed by classifying census block into some categories or stratum based on wealth index and geographic difficulty index. By these procedures, elements in the same stratum are expected to have the similar characteristics so that the variation within stratum will be minimum. In regard to cost efficiency, there is a possibility in stratification procedures to design the appropriate sample allocation to each stratum in order to achieve the possible sample that has the minimum cost.

2. GEOGRAPHIC DIFFICULTY INDEX

Data source that is used to construct geographic difficulty stratum is Village Potency 2011. This data is a village census data in Indonesia that contains various characteristics regarding the condition of the village. In order to construct the geographic difficulty stratum, there are 7 categorical variables used, such as location of village (var1), slope of land (var2), village location of forest area (var3), transportation between villages (var4), road accessibility between village (v5), distance from village to the centre of district (v6), and road accessibility of four-wheels motor vehicles within village (v7). Rescoring must be done in each variables to create the ordinal scale.

Table 1. Scoring of Geographic Variables

Code	Name of variable	Score
var1	Location of village	Peak of mountain Mountainside Valley Plain
var2	Slope of land	Steep Medium Sloping
var3	Village location of forest area	Inside the forest area In the edge of the forest area Outside the forest area
var4	Transportation between villages	Water transportation Land trasportation and the road surface is ground Land transportation and the road surface is hardened by gravel/stones Both land transportation and water transportation, land trasportation and the road surface is asphalt
var5	Road accessibility between villages	Water transportation Sometimes can not be accessed Always can be acceessed
var6	Distance from village to the centre of district	Far Medium Near
var7	Road accessibility of four-wheel motor vehicles within village	Always can be accessed Can be accessed, except if it rains, erosion, etc Can be accessed, except during the rainy season Can not be accessed

The correlation matrix of those variables at national level is shown below.

Table 2. The Correlation Matrix of Geographic Variables at National Level

	var1	var2	var3	var4	var5	var6	var7
v1	1						
v2	0.5137	1					
v3	0.3233	0.2307	1				
v4	0.1751	0.1134	0.1980	1			
v5	0.1420	0.0750	0.2108	0.6214	1		
v6	0.1600	0.1458	0.2539	0.3225	0.2949	1	
v7	0.1967	0.1140	0.2423	0.5398	0.8023	0.3324	1

From the correlation matrix above, it is known that the correlation coefficient among the variables are always positive so that those variables are appropriate to be processed in the next steps. To construct the composite indicator of geographic difficulty, first of all, the weight of each variables must be determined. Because the geographical variables utilized are in the categorical scale, Polychoric Principal Component Analysis is chosen as a method in weighting process to determine the weight of each variables. Weight of each variables is derived from the first eigen vector produced by Polychoric Principal Component Analysis Process because the first eigen vector explains the largest variability of data. The result of Polychoric Principal Component Analysis Process is shown in Table 3.

Table 3. Weight Resulted From Polychoric Principal Component Analysis

	Weight			
	Score1	**Score2**	**Score3**	**Score4**
var1	-0.756	-0.385	-0.215	0.142
var2	-0.515	-0.231	0.133	.
var3	-0.707	-0.369	0.131	.
var4	-0.953	-0.613	-0.328	0.217
var5	-1.069	-0.657	0.121	.
var6	-0.395	-0.002	0.392	.
var7	-0.803	-0.545	-0.455	0.160

The statistic formulation to construct index of geographic difficulty is

$$I_c = \sum_{a=1}^{7} \sum_{b=1}^{B_a} \gamma_{ab}\, x_{abc}$$

I_c : geographic difficulty index for the c-th village

γ_{ab} : weight for the a-th variable and the b-th score

x_{abc} : dummy variable ($x_{abc} = 1$ if the c-th village has the a-th variable and the b-th score, $x_{abc} = 0$ if the others).

In order to be convenient for interpretation, min-max normalization procedure is implemented so that the geographic difficulty index has the range scale from 0 (low difficulty) to 100 (high difficulty).

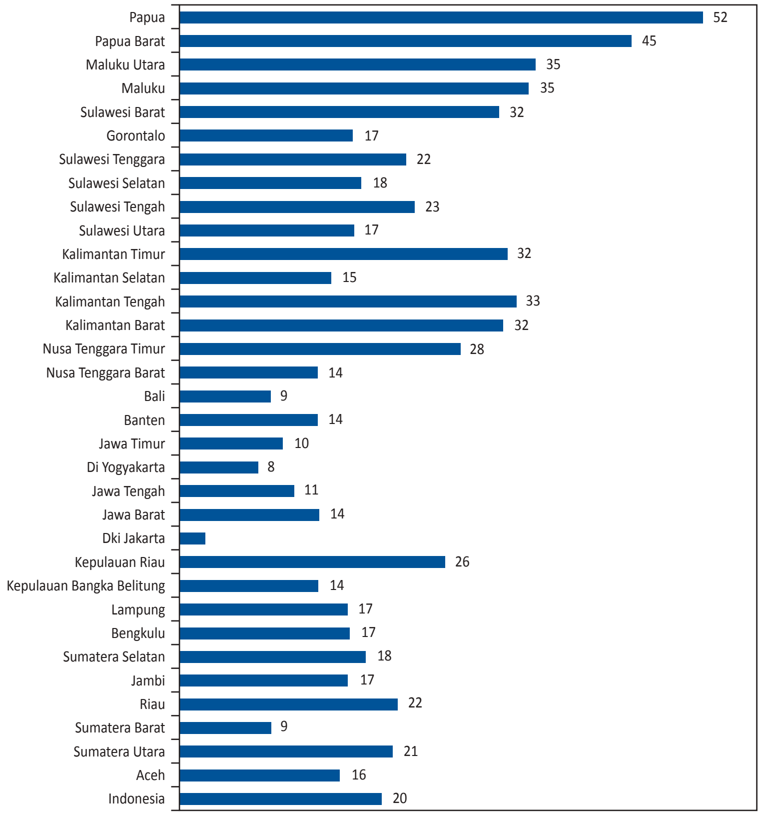


Figure 1. Average Index Of Geographic Difficulty, Split By Provinces

The result of geographic difficulty index can be summarized over provinces so that we can perceive the comparison of geographic difficulty level among provinces. As it can be seen in Figure 1, the average geographic difficulty index at national level is 20. The province that has the lowest average index is DKI Jakarta, while Papua has the highest level of difficulty (the average indexs are 2 and 52, respectively). Based on this results, this paper focusses the analysis in Papua Province as the domain estimation of simulation study in terms of stratification method, statistical efficiency, and cost efficiency issues.

3. WEALTH INDEX

Population Census 2010 is used as data source to construct wealth index, including 9 operational variables such as level of education of household head, type of floor, main source of lighting, the primary fuel for cooking, the main source of drinking water, type of toilet, septic tank, telephone ownership, and internet ownership. The method used to determine weight is the same with that of Geographic Difficulty Index, namely Polychoric Principal Component Analysis. Firstly, wealth index is calculated in household level. Then, wealth index concentration is constructed in the census block level (census block is usually used as primary sampling unit). The statistic formulation used are as seen below.

$$I_r = \sum_{p=1}^{9} \sum_{q=1}^{Q_p} \gamma_{pq}\, x_{pqr}$$

I_r : wealth index for the r-th household

γ_{pq} : weight for the p-th variable and the q-th score

x_{pqr} : dummy variable ($x_{pqr} = 1$ if the r-th household has the p-th variable and the q-th score, $x_{pqr} = 0$ if the others).

Normalization wealth index in household level is calculated by:

$$I_r^{(norm)} = \left(\frac{I_r - \min(I_r)}{\max(I_r) - \min(I_r)}\right) \times 100$$

Then, the wealth concentration index in census block level is calculated by:

$$I_s = \frac{1}{R_s}\sum_{r=1}^{R_s} I_r^{(norm)}$$

Table 4. The Summary of Wealth Index Concentration in Papua Province

Mean	Std. Dev.	Min	Max	Skewness	Kurtosis
38.5756	17.9029	9.8031	88.4714	0.6920	2.0428

Scenarios for Stratification

Stratification is constructed based on the wealth consentration index and geographic difficulty index. There are 15 scenarios of stratification simulated which the determination of strata boundaries is based on cumulative root frequency method. Moreover, optimum allocation is implemented so that the sampling variance for population can be calculated to measure the statistical efficiency. On the other hand, in terms of cost efficiency, there is an approximation by using the average of difficulty index of these scenarios. In the other words, it means that if the scenario has the low average of difficulty index, this result indicate that this scenario tends to have the low level of difficulty. As a consequence, it is expected that the level of cost efficiency can be improved if this scenario is implemented.

Based on the simulation study of Papua Province, the correlation coefficient between geographic difficulty index and wealth concentration index is very strong, approximately -0,7983. It means that the higher level of geographic difficulty, the lower level of prosperity. This high correlation has statistical advantage in terms of stratum construction, because one of the most important aspects that influences the efficiency of stratification is the high correlation between variables used in stratification and research variables.

Table below shows the value of variance according to number of strata in which it shows that difference combination of strata resulted in difference value of variances. In the case that strata of wealth index are not used, the higher number of geographic difficulty strata, the lower value of variance will be, but the value of variance increase when the strata of geographic difficulty is set to be 4. Meanwhile, in the case that strata of geographic difficulty are not used, the higher number of wealth index strata, the lower the variance will be. The smallest value of variance will be obtained in the case that each of wealth index strata and geographic difficulty strata are set to be 4.

Table 5. Value of Variance According to Number of Wealth Index Strata and Geographic Difficulty Strata

	Strata of Geographic Difficulty			
Strata of Wealth Index	**1**	**2**	**3**	**4**
1		0.074884	0.059585	0.059765
2	0.059843	0.044727	0.039180	0.039472
3	0.030705	0.029911	0.028058	0.028277
4	0.023989	0.023083	0.022918	0.022916

According to simulations, the less difficult (the lowest cost) is strata 13 or by using three strata of geographic difficulty. In addition, based on statistical efficiency, the smallest variances are strata 43 and 44. Furthermore, some sampling scenarios can be categorized into four quadrants, the first quadrant with small sampling variance and low cost, the second quadrant with big sampling variance and low cost, the third quadrant with big sampling variance and high cost, and the fourth quadrant with small sampling variance and high cost. The value of in Papua after using stratification are 37 to 39, compared to 52 in mean of population. It shows that by using stratification, the geographic difficulty is decreasing. According to 15 scenarios used, all values of difficulty are less than population mean, this is the advantage of stratification. The high correlation between wealth index and geographic difficulty index provides advantage to the mean of geographic difficulty index based on the selected samples when using optimum allocation.

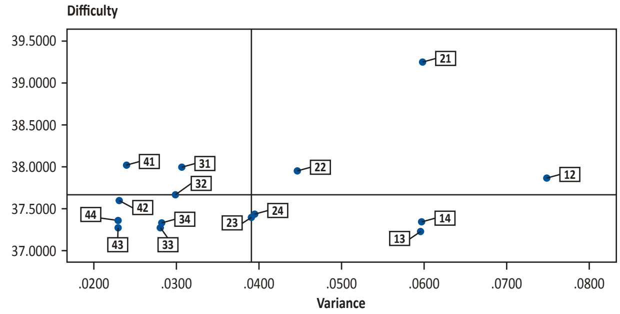


Figure 2. Sampling Scenarios Based on Variance and Difficulty

5. CONCLUSIONS

The correlation coefficient between geographic difficulty index and wealth concentration index is very strong which means that the higher level of geographic difficulty, the lower level of prosperity. This strong correlation has statistical advantage in terms of stratum construction. Simulation results illustrate that several alternative scenarios of efficient stratification with small sampling variance and low cost of the survey can be obtained. Stratification in remote area impacts on better statistical efficiency and cost efficiency.

REFERENCES

Cochran, William G. (1953). Sampling Techniques. New York, John Wiley & Sons.

Kish, Leslie. (2004). Statistical Design for Research. New York, John Wiley & Sons.

_________. (1965). Survey Sampling. New York, John Wiley & Sons.

Rossi, Peter H., James D. Wright, and Andy B. Anderson [Ed.]. (2013). Handbook of Survey Research. New York, Academic Press.

Verma, Med Ram, Jadish Prasad Singh Joorel, and Rajeev Kumar Agnihotri. (2012). Minimum Variance Stratification for Compromise Allocation, Rev. Bras. Biom., Sao Paulo, 30 (2), pp. 278-291.

This page intentionally left blank

CPS19:
Macrofinancial Statistics (3)

Prediction of Mortality Rates Using Latent Factors

Chon Sern Tan*
Universiti Tunku Abdul Rahman, Kajang, Malaysia – tcsern@utar.edu.my

Ah Hin Pooi
Sunway University, Petaling Jaya, Malaysia – ahhinp@sunway.edu.my

Abstract

Forecasting future mortality rates is essential to insurance companies as the increasing life expectancy of human population leads to longevity risks for insurers who provide retirement benefits. In earlier research, mortality data of the United States from the years 1933 till 2000 were fitted with a time series model based on multivariate power-normal distribution to predict the future age-specific mortality rates from the years 2001 till 2010. In this paper, the multivariate power-normal distribution is fitted to the latent factors extracted from a large set of age groups' mortality rates. Aside from having good ability of covering the observed future mortality rates, the prediction intervals attained based on the latent factors tend to have distinctly shorter lengths of intervals as well.

Keywords: Death rate; Power-normal distribution; Prediction interval; Time series model.
JEL Classification: C22, C53, J11

1. INTRODUCTION

Mortality rates are steadily decreasing in this modern era due to healthier lifestyle and improving medical advancement. Increasing lifespan not only renders individuals' personal savings inadequate as financial support for retirement age, it may also affect the long-term solvency of insurers which offer longevity-based policies such as life annuities because their premiums may risk being underestimated due to initial assumptions of mortality rates that are higher than the actual rates. Thus, an appropriate and efficient model is significant for insurers to forecast mortality rates.

A pioneering time series model was established by Lee & Carter (1992) to predict long-run age-specific mortality rates. First, the logs of central mortality rates are modeled as a linear combination of an age-specific constant and the multiplication of a time-varying level-of-mortality index and another age-specific constant. Fitting the historical data with this model, the age-specific constants are obtained using singular value decomposition while the time-varying index is fitted as a stochastic time series from which the future index is predicted. Lastly, these predicted indexes and age-specific constants are used to forecast the future age-specific central mortality rates.

Subsequently, similar models have been proposed by extending the Lee-Carter method. The Renshaw & Haberman (2006) model is among the first stochastic models for population mortality that captures the mortality differences observed between individuals of different cohorts, by adding on a parameter for random cohort effect which is a function of the approximate year of birth. In addition to the cohort effect parameter, the two-factor CBD model by Cairns, Blake & Dowd (2006) incorporates two time-varying indexes instead of only one unique component previously. Considering the different objectives and underlying shapes of mortality structures, a general class of flexible multivariate time series model was proposed and a specific model is chosen where the best criterion is assessed by determining the parameters to keep or remove. In fact, Chan, Li & Li (2014) argued that the CBD model uses the most suitable time-dependent model parameters as indexes to describe the levels of longevity risk at different time-points. Moreover, the ability to incorporate model-parameters uncertainty into simulations is vital to assess the effect of their estimation errors (Cairns, Blake & Dowd, 2008). Using the two-factor model above to investigate this concern in Cairns et al (2006) and Dowd et al (2010), including the parameter uncertainty is observed to have a notable impact on the uncertainty levels in the predicted mortality rates and expected future lifetimes especially in the long term.

The following are some extensions of the Lee-Carter method that are more recent. Brockett et al (2013) obtained a better performance than the original model by applying the Lee-Carter method on the difference in log mortality rates as well as a Levy process and the Normal Inverse Gaussian distribution on the mortality index. In the algorithm called Multiple Lee-Carter Panel Sieve (MLCPS) by D'Amato et al (2012), the forecasts from the Lee-Carter model are combined with a bootstrap procedure for dependent data so that the historical parametric structure and the intra-group error correlation structure are retained. To estimate the relationship between populations that have similar socioeconomic conditions, this algorithm is used and shown through the results to work well even with the existence of the dependence structures. In the paper by Hunt & Blake (2014), the problems of over-parameterization of model and unjustifiable adding of model terms were mitigated by establishing a general procedure in constructing mortality models using a toolkit of functions and expert judgment. Through these models, each significant demographic feature in the data can be distinguished sequentially and then given a parametric structural form, consequently obtaining a relatively parsimonious model with good fit to the mortality data of United Kingdom (UK). Other than that, in the paper by Cairns et al (2011), a general framework was proposed to simultaneously describe the dynamics of mortality rates of two related populations by using a mean-reverting autoregressive process to model the difference in the stochastic factors between the two populations. The modeling of the stochastic factors is further improved in the paper by Zhou et al (2014), where a vector error correction model is used for its advantage of helping to eliminate the need to make assumption in identifying the dominant population.

In the paper by Pooi et al (2014), the prediction intervals for future age-specific mortality rates of the United States (US) were obtained by applying a time series model based on multivariate power-normal (MPN) distribution. Fitting a MPN distribution to the US mortality data from the years 1933 till 2000, the prediction intervals were obtained for future total death rates of the years 2001 till 2010. This methodology was subsequently implemented (i) on data that were augmented to include the death rates of the populations of Canada and UK (Tan & Pooi, 2016a) as well as (ii) using a model with time-varying stochastic parameters (Tan & Pooi, 2016b). The prediction intervals obtained in both papers perform well because they managed to cover all the observed future death rates and also tend to have interval lengths that are increasingly shorter than those by Pooi et al (2014). Besides that, the methodology was also applied to the incomplete US mortality data which contain missing values (Tan & Pooi, 2015). Even though the lengths of prediction intervals may become longer, they were shown to have good ability of covering the observed future death rates.

In this paper, we fit the multivariate power-normal distribution to the latent factors extracted from a large set of age groups' mortality rates. Aside from having good ability of covering the observed future death rates, the prediction intervals attained based on the latent factors tend to have comparatively shorter lengths of intervals than those by Pooi et al (2014). The following describes the layout of this paper. In Section 2, the multivariate time series given by Pooi (2012) is summarized and some results on the forecast of deaths rates given by Pooi et al (2014) are emphasized. Section 3 presents the results of predictions attained based on the latent factors. Lastly, Section 4 concludes the paper.

2. A TIME SERIES MODEL BASED ON MULTIVARIATE POWER-NORMAL DISTRIBUTION

Pooi (2012) proposed a multivariate time series model that uses the following power-normal transformation given by Yeo & Johnson (2000). The random variable $\tilde{\varepsilon}$ has a power-normal distribution with parameters λ^+ and λ^- if

$$\tilde{\varepsilon}=\psi(\lambda^+,\lambda^-,z)=\begin{cases}\dfrac{(z+1)^{\lambda^+}-1}{\lambda^+} & if\,(z\geq 0,\lambda^+\neq 0)\\ \log(z+1) & if\,(z\geq 0,\lambda^+=0)\\ \dfrac{-\left[(-z+1)^{\lambda^-}-1\right]}{\lambda^-} & if\,(z<0,\lambda^-\neq 0)\\ -\log(-z+1) & if\,(z<0,\lambda^-=0)\end{cases} \tag{1}$$

where z is standard normally distributed.

Using the univariate power-normal distribution, a multivariate power-normal (MPN) distribution may be found for a vector y comprising of k random variables that are correlated. This vector y has a k-dimensional power-normal distribution (Pooi, 2012) with parameters $\mu, H, \lambda_i^+, \lambda_i^-, \sigma_i, 1 \le i \le k$ if $y = \mu + H\varepsilon$, where $\mu = E(y)$, H is an orthogonal matrix which consists of the eigenvectors from the variance-covariance matrix of y, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ are uncorrelated,

$$\varepsilon_i = \sigma_i \left[\tilde{\varepsilon}_i - E\left(\tilde{\varepsilon}_i\right) \right] / \left\{ \mathrm{var}\left(\tilde{\varepsilon}_i\right) \right\}^{1/2}, \tag{2}$$

$\sigma_i > 0$ is a constant, and $\tilde{\varepsilon}_i$ has a power-normal distribution with parameters λ_i^+ and λ_i^-.

Through the MPN distribution, we may find a multivariate time series model for a vector $x(t)$ that consists of n_c recorded observations at time t. Letting Δt be a small increment of time after t, an $n_c(l+1)$- dimensional power-normal distribution is obtained for the vector $\boldsymbol{x}^{(1)} = [\boldsymbol{x}(t-(\ell-1)\Delta t), \ldots, \boldsymbol{x}(t-\Delta t), \boldsymbol{x}(t), \boldsymbol{x}(t+\Delta t)]$. Using the above $nc(l+1)$-dimensional power-normal distribution, we may obtain an nc-dimensional conditional distribution of $x(t+\Delta t)$ that will next determine a lag-(l-1) multivariate time series model for the vector of nc time-dependent correlated observations.

With the assumption of stationarity for the multivariate time series, the vector $\boldsymbol{x}^{(d)} = [\boldsymbol{x}(t+(d-\ell)\Delta t), \ldots, \boldsymbol{x}(t+(d-2)\Delta t), \boldsymbol{x}(t+(d-1)\Delta t), \boldsymbol{x}(t+d\Delta t)]$ for $d \ge 2$ may be considered to have the same distribution as the vector $x^{(1)}$. Hence, given the value of $[\boldsymbol{x}(t+(d'-\ell)\Delta t), \ldots, \boldsymbol{x}(t+(d'-2)\Delta t), \boldsymbol{x}(t+(d'-1)\Delta t)]$, a conditional distribution may be obtained for $\boldsymbol{x}(t+d'\Delta t)$, and then a value for $\boldsymbol{x}(t+d'\Delta t)$ may be generated for $d' = 2, 3, \ldots, d$. In this manner, a value of $\boldsymbol{x}(t+d\Delta t)$ may be generated for $d' = d$. This procedure of generating $\boldsymbol{x}(t+d\Delta t)$ may be iterated for a large number of times, and from these generated values of $\boldsymbol{x}(t+d\Delta t)$, a marginal distribution may be found for the j-th $(1 \le j \le n_c)$ component of $\boldsymbol{x}(t+d\Delta t)$. To predict the value of the j-th component of $\boldsymbol{x}(t+d\Delta t)$, we may use the prediction intervals whose end points are given respectively by the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentage points of the marginal distribution.

In Pooi et al (2014), a lag-0 model and a lag-1 model for a vector of 19 age groups' total death rates were constructed to find the nominally 95% prediction intervals for the United States' total death rate of the age group 60-64. The prediction intervals are displayed in Table 1, where the intervals' lower and upper limits are denoted by L_d and U_d respectively for the death rates d years after the year 2000.

Table 1. The nominally 95% prediction intervals for future total death rates of the age group 60-64

d	Observed Values, O_d	Lag-0			Lag-1		
		Lower Limit, L_d	Upper Limit, U_d	Interval Length	Lower Limit, L_d	Upper Limit, U_d	Interval Length
1	0.012049	0.011200	0.012500	0.001300	0.011400	0.012000	0.000600
2	0.011876	0.010800	0.012800	0.002000	0.010500	0.011700	0.001200
3	0.011774	0.010500	0.012900	0.002400	0.009790	0.013000	0.003210
4	0.011319	0.009910	0.013300	0.003390	0.007460	0.015400	0.007940
5	0.011262	0.007840	0.016700	0.008860	0.003430	0.022600	0.019170
6	0.010813	0.003670	0.021400	0.017730	0.000443	0.032300	0.031857
7	0.010625	0.000572	0.022800	0.022228	0.000000	0.038500	0.038500
8	0.010505	0.000000	0.033700	0.033700	0.000794	0.038800	0.038006
9	0.010258	0.000000	0.040300	0.040300	0.001810	0.041100	0.039290
10	0.010063	0.000000	0.054200	0.054200	0.000000	0.039100	0.039100

The interval lengths from Table 1 are observed to be comparable to those found by Jarner & Dengsoe (2009), who used the Lee-Carter method. Therefore, these interval lengths will be the benchmark to be compared with the prediction interval lengths outcome from this paper. Improvement in the predictive ability of the model is indicated by prediction intervals that have shorter interval lengths, and at the same time can still cover the observed future values.

3. PREDICTION OF US MORTALITY RATES USING LATENT FACTORS

In this section, we will fit the multivariate power-normal distribution to the latent factors extracted from a large set of age groups' mortality rates. The available mortality data consist of the total central death rates (male and female combined) from the year 1933 till 2010 (78 years) for the 19 age groups $15-19, 20-24, \ldots, 60-64, \ldots, 105-109$.

Suppose we want to predict the future death rate of the 10th age group 60-64. Let $\boldsymbol{m}_t = \left[m_{t1}\, m_{t2} \cdots m_{tN}\right]^T$ be an N-dimensional vector consisting of N = 18 remaining age groups' time-t mortality rates. The vector m_t may be described by a factor model with its static representation given by $\boldsymbol{m_t} = \Lambda \boldsymbol{F_t} + \boldsymbol{e_t}$ where Λ is an $N \times r$ matrix of factor loadings, F_t is an $r \times 1$ vector of common latent factors underlying m_t, and e_t is an $N \times 1$ vector of random errors.

Using the mortality data that span over T time points, a table consisting of T rows may be formed where the i-th row represents the values of the N age groups' mortality rates observed at the i-th time point. Then, the jw-th sub-table may be formed comprising of the jw-th row till the $\left(j_W + n_t - 1\right)$-th row from the original table of T rows.

Next, the principle components analysis is performed on the N columns of observed age groups' mortality rates in the jw-th sub-table. From the initial set of N principal components, r principal components (r < N) are obtained consisting of $f_1, f_2, \ldots, f_r$ where f_i has the i-th largest variance.

Let f_{ij} be the value of f_i extracted from the j-th row of mortality data in the jw-th sub-table. Now, the jw-th window may be constructed using the f_{ij} extracted from the jw-th sub-table such that a row of the jw-th window is of the form:

$$\left[\left(f_{1\,j-\ell+1} f_{2\,j-\ell+1} \cdots f_{r\;j-\ell+1}, m_{j^*-\ell+1,10}\right) \cdots \left(f_{1j} f_{2j} \cdots f_{rj}, m_{j^*,10}\right), m_{j^*+d,10}\right] \tag{3}$$

where $1 \le j - \ell + 1 \le n_W + 1$, $n_W = n_t - \ell$ and $m_{j^*+d,10}$ is the observed value of the total death rate for the ten-th age group 60-64 d years in the future from the current year $j^* = j_W - 1 + j$.

The data in the first n_W rows of the j_W-th window are then fitted with an $\left[\ell(r+1)+1\right]$-dimensional power-normal distribution. When the values of the first $\ell(r+1)$ entries in the final row of the j_W-th window are given, the fitted multivariate power-normal distribution is used to obtain a conditional distribution for the future death rate $m_{j^*+d,10}$ of the age group 60-64.

We take the $100(\alpha/2)$ % and $100(1-\alpha/2)$ % points of the conditional distribution as the lower limits and upper limits respectively of the nominally $100(1-\alpha)$ % out-of-sample prediction interval for the future death rate $m_{j^*+d,10}$ of the age group 60-64. If its estimated coverage probability is close to the target value of $1-\alpha$ and its average length of interval is short, the above prediction interval is then considered to be satisfactory.

The length of the prediction interval will tend to decrease as the lag value $(\ell - 1)$ and the number of latent factors r increase. However, in order to obtain a parsimonious model, the values of ℓ and r should be kept as small as possible while achieving reasonably small values of the length of the prediction interval. It is found that the value 3 or 4 for r and the value of ℓ which is less than or equal to 3 would be fairly good choices.

Figure 1 displays the lengths of the 95% out-of-sample prediction intervals for the death rate in the year $1981 + \ell + j_w$ based on the lag-($\ell - 1$) model with $r = 4$ latent factors fitted to the data in the j_w-th window where $1 \le j_w \le 29 - \ell$. From Figure 1, it can be observed that the interval lengths tend to decrease as the lag increases. This reaffirms the intuition that the additional information contributed by higher lags may improve the predictive ability of the model.

Figure 2 displays the lower limits and upper limits of the prediction intervals from Figure 1 for the lag-0, lag-1 and lag-2 models respectively, where it is observed that the prediction intervals have covered a large proportion of the observed future total death rates.

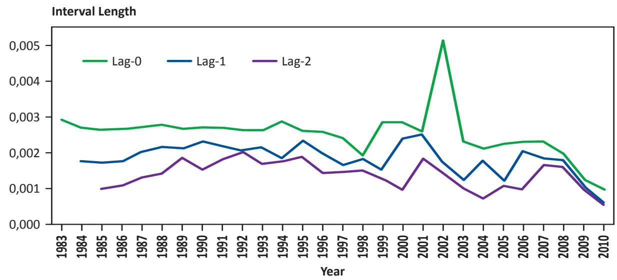


Figure 1. Lengths of the Nominally 95% Prediction Intervals based on $r=4$ Latent Factors ($d=10$, $\alpha=0.05$, $n_W=40$) for Future Total Death Rates $m_{j^*+d,10}$ of the Age Group 60-64

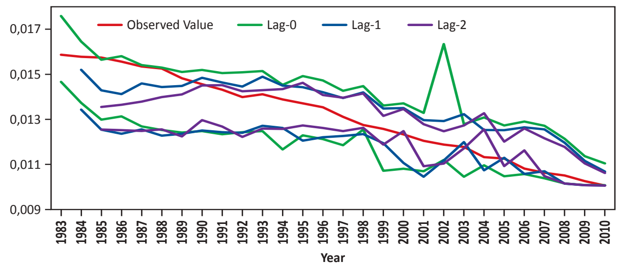


Figure 2. Lower Limits and Upper Limits of the Prediction Intervals based on Latent Factors for the Lag-0, Lag-1 and Lag-2 Models

Table 2 shows the comparison of prediction intervals for the death rate of the age group 60-64 in 2010. From Table 2, it can be observed for $\ell=1,2,3$ that the prediction interval from the model based on the $(29-\ell)$-th window is much shorter than those presented in the literature.

Table 2. Comparison of prediction intervals for *d* = 10

d	Interval Lengths						
	From Table 1		UK-Canada-US from Tan & Pooi (2016a)	Model with Stochastic Parameters from Tan & Pooi (2016b)	Based on Latent Factors		
	Lag-0	Lag-1			Lag-0	Lag-1	Lag-2
10	0.054200	0.039100	0.003805	0.002900	0.000974	0.000616	0.000556

4. CONCLUSIONS

This paper presents a promising methodology that makes use of a multivariate time series model on the United States mortality data where future mortality rates are forecasted based on latent factors. This method has improved the predictive ability of the model as shown by the encouraging results of the prediction intervals' good coverage of the future observed mortality rates and interval lengths that are distinctly shorter. As a further work, we may investigate the model using latent factors extracted from the US mortality data that are augmented with those from countries such as Canada and United Kingdom.

REFERENCES

Brockett, P., Chuang, S., Deng, Y., and MacMinn, R. (2013). "Incorporating Longevity Risk and Medical Information into Life Settlement Pricing", *Journal Risk and Insurance*, 80, 799-826.

Cairns, A., Blake, D., and Dowd, K. (2006). "A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration", *Journal of Risk & Insurance*, 73, 687-718.

Cairns, A., Blake, D., and Dowd, K. (2008). "Modelling and management of mortality risk: a review.", *Scandinavian Actuarial Journal*, 2, 79-113.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., and Khalaf-Allah, M. (2011). "Bayesian Stochastic Mortality Modelling for Two Populations", *Astin Bulletin*, 41, 29-59.

Chan, W., Li, J., and Li, J. (2014). "The CBD Mortality Indexes: Modeling and Applications ", *North American Actuarial Journal*, 18, 38-58.

D'Amato, V., Haberman, S., Piscopo, G., and Russolillo, M. (2012). "Modelling Dependent Data for Longevity Projections ", *Insurance: Mathematics And Economics*, 51, 694-701.

Dowd, K., Blake, D., and Cairns, A. (2010). "Facing up to Uncertain Life Expectancy: The Longevity Fan Charts", *Demography*, 47, 67-78.

Hunt, A., and Blake, D. (2014). "A General Procedure for Constructing Mortality Models", *North American Actuarial Journal*, 18, 116-138.

Jarner, S. F., and Dengsøe, C. (2009). Uncertainty and coherence of mortality projections. Working Paper, *The Danish Labour Market Supplementary Pension Fund (ATP)*.

Lee, R., and Carter, L. (1992). "Modeling and Forecasting US Mortality", *Journal of The American Statistical Association*, 87, 659-671.

Pooi, A. H. (2012). "A Model for Time Series Analysis", *Applied Mathematical Sciences*, 6, 5735-5748.

Pooi, A. H., Pan, W. Y., and Wong, Y. C. (2014). "Prediction Intervals for Future Mortality Rates", *Applied Mathematical Sciences*, 8, 5039-5051.

Renshaw, A., and Haberman, S. (2006). "A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors", *Insurance: Mathematics and Economics*, 38, 556-570.

Tan, C. S., and Pooi, A. H. (2015), "Prediction of Mortality Rates in the Presence of Missing Values" in *Proceedings of the 2nd Innovation and Analytics Conference & Exhibition*, Vol. 1691, pp. 050027.

Tan, C. S., and Pooi, A. H. (2016a). "Prediction of Mortality Rates using Augmented Data", *Jurnal Teknologi*, 78, 4-4.

Tan, C. S., and Pooi, A. H. (2016b), "Prediction of Mortality Rates using a Model with Stochastic Parameters" in *Proceedings of the 4th International Conference on Quantitative Sciences and Its Applications*, Vol. 1782, pp. 050016.

Yeo, I. K., and Johnson, R. A. (2000). "A New Family of Power Transformations to Improve Normality or Symmetry", *Biometrika*, 87, 954-959.

Zhou, R., Wang, Y., Kaufhold, K., Li, J., and Tan, K. (2014). "Modeling Period Effects in Multi-Population Mortality Models: Applications to Solvency II ", *North American Actuarial Journal*, 18, 150-167.

The GARCH-EVT-Copula Model and Simulation in Scenario based Asset Allocation

Gary David Sharp*
NMMU, Port Elizabeth, South Africa – gary.sharp@nmmu.ac.za

Peter GF McEwan
NMMU, Port Elizabeth, South Africa – gareth.mcewan@quintiles.com

Abstract

Financial market integration, in particular, portfolio allocations from advanced economies to South African markets, continues to strengthen volatility linkages and quicken volatility transmissions between participating markets. Largely as a result, South African portfolios are net recipients of returns and volatility shocks emanating from major world markets. In light of these, and other, sources of risk, this study proposes a methodology to improve risk management systems in funds by building a contemporary asset allocation framework that offers practitioners an opportunity to explicitly model combinations of hypothesised global risks and the effects on their investments. Monte Carlo simulations were generated from a fitted MV meta t distribution on which an out-of- sample test was conducted. The 2014-to-2015 horizon served to proxy as an out-of-sample, forward - looking scenario for a set of key risk factors against which a hypothetical, diversified portfolio was optimised. Traditional mean-variance and contemporary mean-CVaR optimisation techniques were used and their results compared.

Keywords: GARCH, Extreme Value Theory, Copula, Simulation, Conditional Value-at-Risk, Portfolio Optimisation.

1. INTRODUCTION

An overarching characteristic of global financial markets is the trend towards unification of financial markets in advanced economies with those in emerging market economies. Factors such as increasing globalisation, financial and exchange rate liberalisation and financial innovation are key contributors strengthening the linkages between international markets (Lane & Milesi-Ferretti, 2008).

Emerging market (EM) economies have been, partly as a result, absorbing a much larger share of outward portfolio investment from advanced economies than was the case prior to the 2008 financial crisis. For example, the International Monetary Fund [IMF] (2014) observed that, between 2002 and 2012, equity portfolio allocations to EM economies from advanced economies increased from 7% of the total stock of advanced economy portfolio investment to almost 20%. Similarly for bond portfolio allocations, from 4% of the total stock of outward portfolio investment from advanced economy markets in 2002 to almost 10% in 2012. These portfolio allocations are principally directed towards only a few destination countries. Of the portfolio allocations to EM equities in 2012, 80% was invested in 12 of the 190 emerging market economies. Similarly for EM fixed income, with 75% directed to the same 12 economies. South Africa was among the 12 destination markets. Concomitantly, a significant degree of portfolio concentration to EM economies hails from only a handful of advanced economies (viz., the United States, United Kingdom, Hong Kong SAR and Singapore). One consequence is an increasing synchronisation in asset price movements and volatilities between these dualistic market-types.

It is well documented (e.g., Fenn, et al., 2011; IMF, 2015; J.P. Morgan, 2011) that there has been a substantial rise in correlations between asset markets in advanced and EM economies. The major asset classes considered generally fall into the categories of international equities, government bonds, corporate bonds, exchange rates, hard commodities and soft commodities, with further differentiation along emerging market and developed market lines.

South African (SA) financial markets are subsumed in the momentum of unification, becoming more synchronised both laterally with EM and vertically with developed markets. The global financial crisis of 2007-2008 was followed by significant losses in portfolios across the globe and, indeed, in portfolios of South African investors. It ushered in a protracted period of market volatility, with traditional risk management models failing because of increased correlation among all asset classes (Stefanova, 2015). The pendulum of investors' attention has since swung away from portfolio returns as the singular measure of success to a more holistic domain focusing on risk management practices.

Diversification, as a means to reduce portfolio risks and extract risk premia, is an effective concept over long-term investment horizons. It provides investors with the best reward per unit risk through an efficient combination of individual assets. The demand for greater diversification is, in fact, one of four of the drivers of globalisation, as investors increasingly trade their "home bias" for exposure outside their home markets (Boston Consulting Group, 2014; Institute of International Finance, 2014; Phillips, 2014). However, short-term, systemic crashes destroy diversification benefits as asset class correlations converge and shocks rapidly propagate through global financial networks. The traditional Market Portfolio Theory (MPT) framework does not protect portfolios against severe losses, or wealth destruction, over such periods. In "normal" markets, the distributional setting in MPT may serve investors well, but financial crises punctuate this state far more often than the normal distribution suggests (Xiong, 2010 and references therein). In "non-normal" markets, current risk management approaches underestimate downside risks, supporting the continued investigation of new approaches to risk management using valid statistical methods.

2. THEORETICAL CONSIDERATIONS

The following facts are typical for univariate returns:

Returns are not independent and identically distributed (i.i.d.). They may also show low absolute values for a first-order autocorrelation coefficient.

Serial correlation in returns is not significant, whereas corresponding absolute or squared returns are mostly autocorrelated (i.e., they exhibit volatility persistence or "long memory").

Return volatilities exhibit conditional heteroskedasticity (i.e., time-varying volatility).

Return distributions are leptokurtic, reflecting fatter tails and "peakedness" around (mostly) positive arithmetic means (i.e., higher probabilities for extreme events and greater central tendency, respectively, than what the normal distribution would generate).

Returns exhibit skewness, reflecting asymmetry in the tails. Distributions are generally negatively skewed, implying higher probabilities of negative returns.

Extreme returns are observed closely in time (i.e., volatility clustering); the clustering itself generates excess volatility, or fat tails.

The approach in this paper builds on research by Inanoglu and Ulman (2009), Mashal and Zeevi (2002), Nyström and Skoglund (2002a), Wang, Sullivan and Ge (2012) and Xiong and Idzorek (2011). This paper proposes a technically advanced, yet tractable, quantitative decision support framework designed to work in lockstep with economists' and experts' opinions on core (statistically or feasibly meaningful) risk drivers to portfolio asset classes. Non-normal stylised facts of univariate financial data are explicitly modelled in the framework. The resulting model is an empirically-and theoretically-consistent simulation engine customised to a portfolio of asset classes and corresponding set of risk drivers.

3. THE ANALYSIS DATA

All data in this paper were prepared using Microsoft Excel and all results produced using the statistical package R (R Development Core Team, 2014), with various R packages used in subsets of analysis. The computer used to implement codes is a Lenovo laptop with Intel Core i5-4200M processor with a clock speed of 2.50 GHz and 8 GB memory, running on a 64-bit Windows 7 Operating System.

This study uses 18 variables in the model: 6 asset class variables and 12 key sector and asset class return driver variables. Asset classes are selected based on a representative, diversified portfolio. Asset classes (e.g., equities) were decomposed into sector and sub-sector levels e.g., industrial sector with sub-sectors: aerospace and defence, electronic and electrical equipment, general industrials, industrial engineering, industrial transportation, etc.

The data were obtained from INET BFA. End-of-month observations, spanning 20 years, from the period 31 March 1994 to 30 April 2014 were used to fit the model. End-of-month observations for the period 30 April 2014 to 30 April 2015 were used to evaluate the model out-of-sample.

Indices were used to represent portfolio asset classes and portfolio return drivers. For portfolio asset classes: FTSE/JSE All Share Total Return Index (ALSI) for South African equities, FTSE/JSE All Bond Index (ALBI) for South African fixed income, MSCI World Index (MSCI.WRLD.ZAR) for international equities, Alexander Forbes money-market index (GMC1) for cash, JP Morgan Global Government Bond Index (GLOUS) for international bonds and the FTSE/JSE Listed Property Index (J253T) for South African real estate. For a list of the portfolio drivers refer McEwan (2016). One example would be the domestic currency (USD/ZAR). Figure 1 illustrates heterogeneity across the JSE all-share index (ALSI) based on the empirical monthly log returns.

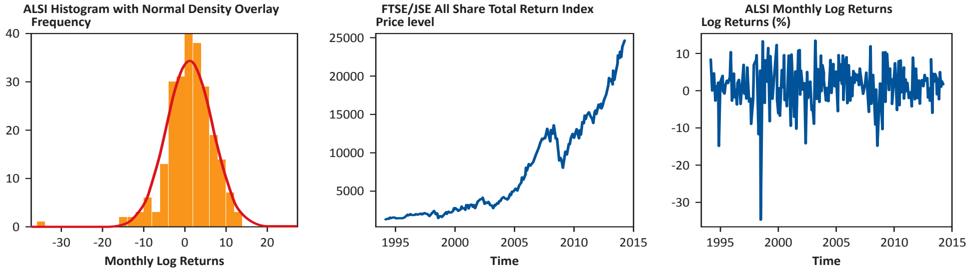


Figure 1. ALSI Return Data Represented Graphically

4. MODELLING METHODOLOGY AND RESULTS

All 18 variables were evaluated in the GARCH modelling framework. In general a GARCH(1, 1) model provided adequate fits based on estimation and goodness of fit tests. Results of the ALSI data fitted for a GARCH (1, 1) model are shown in Table 1. The results provide adequate support for the fitted model.

Table 1. Parameter estimates for GARCH (1, 1) model for ALSI data

Parameter	Estimate	p-value (std errors)	p-value (robust errors)
Mu	0.0159***	0.0000	0.0000
Omega	0.0003	0.2028	0.2243
Alpha1	0.2342	0.1109	0.2666
Beta1	0.6839***	0.0001	0.0031
Shape	7.2137***	0.0092	0.0374

The residuals from each model were standardized with mean zero and unit variance. Figure 2 is a histogram of the standardized residuals providing evidence that for simulation modelling requirements it was necessary to obtain a suitable methodology for generating residuals mirroring the empirical data asymmetry. The residuals were sorted and then using empirical semi-parametric EVT modelled by partitioning the sorted data.

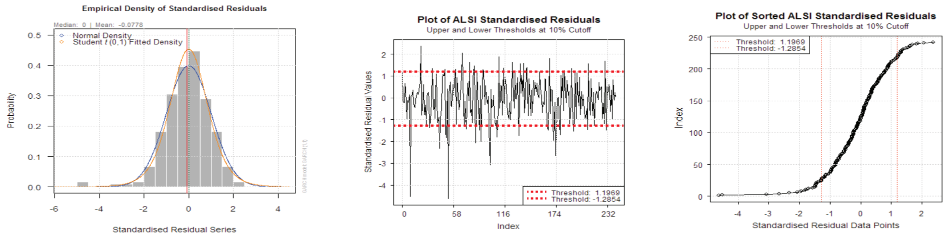


Figure 2: ALSI Residual Data Illustrated Graphically

Each variable is represented by three components, applying the peaks over threshold (PoT) methodology. To ensure sufficient data in the tails, a 10% threshold limit was selected. The lower tail was fitted with generalised Pareto density estimates for shape and scale, the interior represented by density estimates along an equally-spaced grid spanning the range of the data and an upper tail fitted with generalised Pareto density estimates for shape and scale. Figure 3 shows the combined non-parametric interior distribution and parametric tail GPDs for the ALSI standardised residual series superimposed over their empirical counterparts.

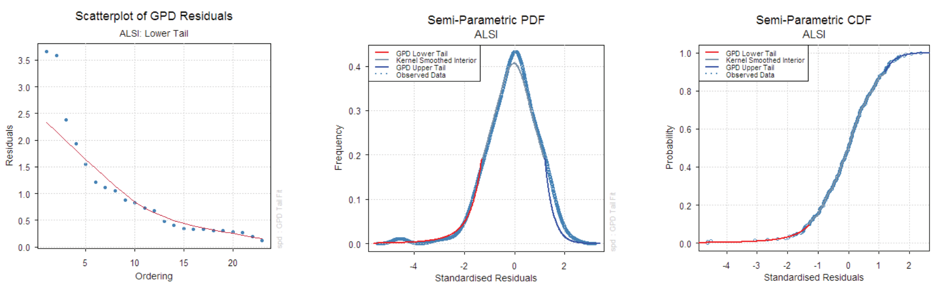


Figure 3. Semi-parametric Representation of Fitted ALSI Residuals

Given the dependence requirement between variables, a copula modelling approach was adopted. This study used a meta-t copula as the density weighting function in a multivariate structure to provide dependence on the residuals. An illustration of the residuals of two indices, the ALSI and ALBI is seen in Figure 4. The meta-t copula captures all dependence between all variables,

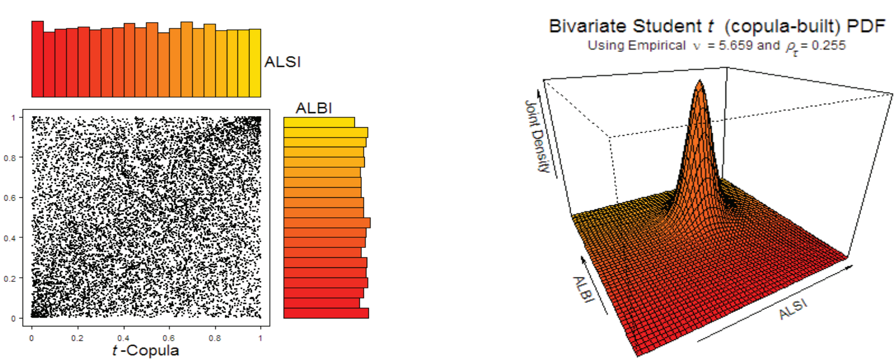


Figure 4. Two Dimensional Representation of ALSI and ALBI Copula Model

Given this methodology to obtain dependence for the asymmetric residuals, data can now be generated to assess opinions of experts in a scenario based simulation. A simple example is provided, more details of the portfolio optimization is available in McEwan (2016). In Figure 5, a scenario is postulated for several variables based on expert opinions. The more volatile the scenario, the better the estimation methodology. An input variable for a portfolio may postulate the rand depreciating, how this and other variables influence the output variables, in this example four are considered.

INPUT VARIABLES (i.e., Return Drivers)	OUTPUT VARIABLES (i.e., Asset Classes)
Range forecasts over subsequent period	Asset classes to be conditioned and optimised
Rand depreciates (+10% to +15%) Dollar strengthens (-15% to -30%) SA 10-yr yields fall (0% to -10%) Oil price falls (-35% to -55%) Commodities prices fall (-30% to -50%) Gold price falls (0% to -20%) Platinum price falls (-15% to -30%) EM equities stable (-5% to +15%) S&P 500 advances (+5% to +15%)	FTSE/JSE All Share Index FTSE/JSE All Bond Index MSCI World Equity Index Domestic Money-market Cash Index

Figure 5. Variables Used as illustrative Example of One Scenario

The results of an annualized return for the ALSI is shown in Figure 6. The density functions illustrate the differences between the empirical results for the year 2014-2015 and those given the scenario based approach with simulated data. The volatility in the left tail is emphasized with the conditional density functions, and provides a better risk assessment method as a result of the inter dependency between variables. Based on the scenario, the expected returns are 4% less for the conditional simulated data. This tool provides a useful measure for portfolio managers to reduce risk in their portfolios.

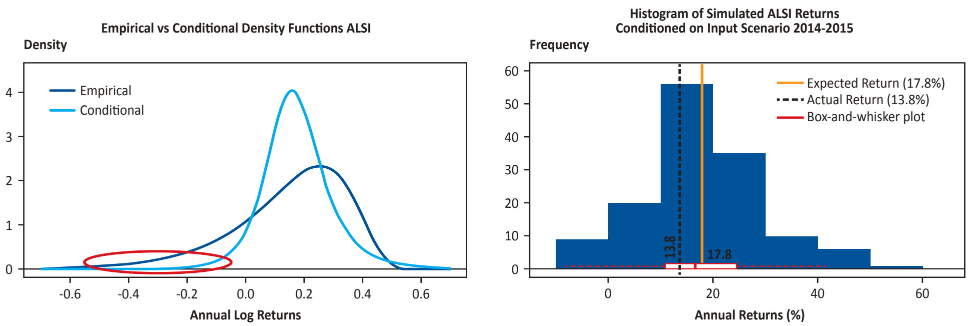


Figure 6. ALSI Results based on Scenario Setting

5. CONCLUSIONS

We acknowledge that global financial markets are trending towards unification, strengthening volatility linkages and quickening volatility transmissions through present-day financial networks. As it relates to South African fund managers, domestic portfolios are net recipients of returns and volatility shocks from major world markets. Financial crises continue to highlight the damaging effects of deep-tail, extreme events on portfolios. Therefore, this study proposes a methodology to improve risk management systems in funds by building an asset allocation framework that offers practitioners an opportunity to explicitly model combinations of hypothesised global risks and the effects on their investments. The goal is to improve fund performances, particularly during periods of market stress.

Opportunity for improvement comes from pooling expert opinions on key near-term portfolio risk drivers. Aggregated opinions on key variables give rise to range forecasts used in a framework where any number of important variables can be combined to construct forward-looking scenarios, or anticipated states of the world. The scenario-building option offers practitioners and investors a pivotal mechanism through which many different effects on fund holdings can be scrutinised. The mechanism fits top-down, macro views onto a flexible, bottom-up, quantitative model.

This paper is a very small part of a dissertation which brings together new technologies available to jointly model portfolio assets and their risk factor threats and elaborated on the requisite theoretical background. At the centre of the model is a dualistic simulation engine constructed to represent the multivariate and univariate data generating processes as accurately as possible.

REFERENCES

Bao, Y. (2015). Should we Demean the Data? *Annals of Economics and Finance,* 16 (1):163-175.

Boston Consulting Group (2014). *Global Asset Management 2014: Steering the Course to Growth* [Online]. Retrieved 10 June 2015 from https://www.bcgperspectives.com/content/articles/ financial_institutions_global_asset_management_2014_steering_course_growth

Fenn, D.J., Porter, M.A., McDonald, M., Williams, S., Johnson, N.F. and Jones, N.S. (2011). Temporal Evolution of Financial-market Correlations. *Physical Review E,* 84 (2):026109.

Ghalanos, A. (2014a). *Introduction to the rugarch package.* R package version 1.3-1.

Ghalanos, A. (2014b). *rugarch: Univariate GARCH Models.* R package version 1.3-4.

Inanoglu, H. and Ulman, S. (2009). *Revisiting Copula Dependency Modeling: A Case for Conservatism, Federal Deposit Insurance Corporation* [Online]. Retrieved 12 July 2014 from https://www.fdic.gov/bank/analytical/cfr/2009/Seminar2009/seminar_series_2009.html

Institute of International Finance (2014). *Financial Globalization: Maximizing Benefits, Containing Risks* [Online]. Retrieved 11 March 2015 from https://www.iif.com/file/7020/download?token=EiCOCQDk

International Monetary Fund (2014). *Global Financial Stability Report, October 2014* [Online]. Retrieved 10 February 2015 from http://www.imf.org/external/pubs/cat/longres.aspx?sk= 41631

International Monetary Fund (2015). *Global Financial Stability Report, April 2015* [Online]. Retrieved 10 February 2015 from http://www.imf.org/external/pubs/cat/longres.aspx?sk=42422.0

J.P. Morgan (2011). Rise of Cross-asset Correlations. *Global Equity Derivatives & Delta One Strategy.* New York, NY: J.P. Morgan.

Lane, P.R. and Milesi-Ferretti, G.M. (2008). The Drivers of Financial Globalization. *American Economic Review,* 98 (2):327-232.

Mashal, R. and Zeevi, A. (2002). Beyond Correlation: Extreme Co-movements Between Financial Assets. (*Unpublished manuscript*), Columbia University: New York, NY.

McEwan P.G.F. (2016). The GARCH-EVT-Copula model and simulation in scenario based asset allocation. Unpublished masters dissertation, NMMU, Port Elizabeth.

Nyström, K. and Skoglund, J. (2002a). A Framework for Scenario based Risk Management. *Swedbank, Group Financial Risk Control*.

Philips, C.B. (2014). Global Equities: Balancing Home Bias and Diversification. *The Vanguard Group, Working Paper*.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing (Version 3.1.0) [Software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Stefanova, K. (2015). *How to Disrupt the Investing Business with Katina Stefanova (Ex-Bridgewater Management Committee) / Interviewer: D. Teten* [Online]. Retrieved 28 April 2015 from http://disruptinvesting.com/2015/04/how-to-disrupt-the-investing-business-with-katina-stefanova-ex-bridgewater-management-committee

Wang, P., Sullivan, R.N. and Ge, Y. (2012). Risk-Based Dynamic Asset Allocation with Extreme Tails and Correlations. *Journal of Portfolio Management,* 38 (4):26-42.

Xiong, J.X. (2010). Nailing Downside Risk. *Ibbotson Associates, Morningstar Advisor*.

Xiong, J.X. and Idzorek, T.M. (2011). The Impact of Skewness and Fat Tails on the Asset Allocation Decision. *Financial Analysts Journal,* 67 (2):23-35.

This page intentionally left blank

CPS20:
Demography & Social Welfare Statistics (3)

The Direct and Indirect Influence of the Urbanization, Unemployment, MYS, Gender Gap in Education and Employment to Indonesian GDP in 2014

Hardius Usman*
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – hardius@stis.ac.id

Fitri Catur Lestari
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – fitricaturlestari@stis.ac.id

Abstract

Indonesia has the relatively high gender inequality. Based on the Gender Gap Index, Indonesia was at 97th place among 142 countries in the world. The rank progress from year to year, it was in 2014, Indonesia was suffered a setback, because Indonesia had previously been at 90 in 2011 and 92 in 2008. Based on the components used to build the index, the gender gap was occurred in the components of economic opportunity and participation, which peaked at 108 in 2014. Meanwhile, the gender gap in education in Indonesia are quite good, namely at the position 78. But the girl participation in education still needed attention.

The first objective of this study was to know the condition of the gender gap in education and employment in Indonesia. The second one was to know the direct and indirect influence of the urbanization, unemployment, Mean Years of Schooling (MYS), gender gap in education and employment to Indonesian GDP in 2014.

The secondary data used were BPS publications (Indikator Kesejahteraan Rakyat, Statistik Ketenagakerjaan, Statistik Indonesia, and Susenas 2014). The graphs and tables were used to describe the condition of the gender gap in Indonesia and the path analysis was used to determine the direct and indirect influence of the urbanization, unemployment, MYS, gender gap in education and employment to Indonesian GDP in 2014.

This first research result was the gender gap in education generally did not occur, contrasting in employment. The second one was Indonesian GDP was influenced by the significantly direct effect of gender gap in education and employment, significantly indirect effect of urbanization and unemployment and unsignificantly indirect effect of MYS.

Keywords: Gender; Education; Employment; Path analysis

1. INTRODUCTION

Gender inequality still persists around the world. The relative status of women is poor in the developing world, compared to developed countries (Dollar and Gatti, 1999). There are different characteristic and measurement of gender inequality in developing and developed country. In many developing countries, the strong evidence that females fare poorly relative to males are: girls receive less education, female mortality is higher than male mortality in many countries, and women often have only tenuous control over land and other productive assets (Mammen and Paxson, 2000).

Indonesia as developing country has the relatively high gender inequality. Based on the Gender Gap Index, Indonesia was at 97th place among 142 countries in the world. The rank progress from year to year, it was in 2014, Indonesia was suffered a setback, because Indonesia had previously been at 90 in 2011 and 92 in 2008. Based on the components used to build the index, the gender gap was occurred in the components of economic opportunity and participation, which peaked at 108 in 2014 (World Economic Forum, 2014). Meanwhile, the gender gap in education in Indonesia are quite good, namely at the position 78. But the girl participation in education still needed attention.

The first objective of this study was to know the condition of the gender gap in education and employment in Indonesia. The second one was to know the direct and indirect influence of the urbanization, unemployment, MYS, gender gap in education and employment to Indonesian GDP in 2014.

2. LITERATURE REVIEW

Gender inequality become a common problem in human development. Although women account for over one-half of the potential talent base throughout the world, as a group they have been marginalised and their economic, social, and environmental contributions go in large part unrealised (OECD, 2008).

When gender inequality is not taken into consideration, economic growth, and development cannot be handled effectively, because poverty leads to gender inequality and gender inequality negatively affects economic development (Canan, 2012). Canan stated that women still suffer from inequalities and concluded that women still suffer from inequalities and human factor is a fundamental determinant of economic development and the major target of economic development is to improve life standards and to provide equal opportunities in health, education, employment for everybody without gender discrimination. Gender equality, the empowerment of women, women's full enjoyment of all human rights and the eradication of poverty are essential to economic and social development as including the achievement of all the Millenium Development Goals (OECD, 2012). World Bank (2003a) also mentioned similiar report:

"There is now a shared understanding within the development community that development policies and actions that fail to take gender inequality into account and fail to address disparities between males and females will have limited effectiveness and serious cost implications."

Women have the important role and can play in economic development and place women in development means legislatively trying to limit discrimination and by promoting their involvement in education and employment (Bradshaw, 2015). Chen's research (2004) provides evidence indicating that gender equality in education is an important contributor to gender equality in employment. Association between gender equality in education and economic development is positive and statistically significant. The ratio of female to male students in primary and secondary schools with economic development also have a statistically significant positive correlation. Increasing educated women relative to men, intuitively more women will be working relative to men. In other words, it likely high positive correlation between gender equality in education and employment. Both in education and employment, gender inequality tends to reduce economic growth (Chen, 2004). Gender inequality in education is bad for economic growth (Dollar and Gatti, 1999). Gender gaps in education and employment considerably reduce economic growth (Klasen and Lamanna, 2009).

In addition, there was some evidence suggesting that gender equality in education measured by female to male ratio of the average or mean years of schooling, also has statistically positive effect on gender equality in employment (Chen, 2004). Educational attainment comparison also used to describe the difference between male and female education phenomenon in Ghana (Amu, 2005).

In a society with gender inequality in the labor market, the majority of the persons not working would be women, and hence it is likely that majority of the new entrants into labor market would be women. The gender equality in terms of labor force participation rates would increase as economic development increasing (Chen, 2004). Beneria and Sen (1981) mentioned in their paper that Boserup's own conclusions on policy emphasized women's education as the major mechanism by which modernization would begin to work to women's advantage. Through education, women can compete more successfully in urban labor markets and gain access to improved agricultural techniques in the rural areas.

In developing country, work opportunity are more available in urban areas rather than in rural ones. In rural areas, people tends to be more conservative and regards to have traditional gender roles views dictated by customs and norms (Chen, 2004). In addition to work opportunity, education opportunity also more available in urban areas. Increased urbanization rates may lead to higher levels of gender equality. Hence urbanization rates have an effect to both employment and education gender equality.

In developing countries that women as secondary breadwinners of the family, women labor force participation rates may actually increase in period of high unemployment if the primary breadwinner is unemployed (Chen, 2004). During it, female workers likely would accept lower wages relative to male workers. Otherwise, high rates of unemployment do not affect to the employment gender equality in developed countries. By the condition, unemployment was supposed to affect the employment gender equality in Indonesia.

Amu's study (2005) mentioned the GLSS4 survey for instance found out that twice as many females as males have never been to school in Ghana. More males have access to education than women. This situation of women in skill and knowledge based industries is low as against the high concentration of women in informal private sector employment and informal self-employment. The study also concluded that women in recent times have increasingly become the backbone of their families as breadwinners. It conforms to traditional conception of women as traders, cooks, and caretakers at home and in the community at large. By reducing the amount of time women and girls must spend in domestic activities, would allow them more time for education or employment (Bradshaw, 2015).

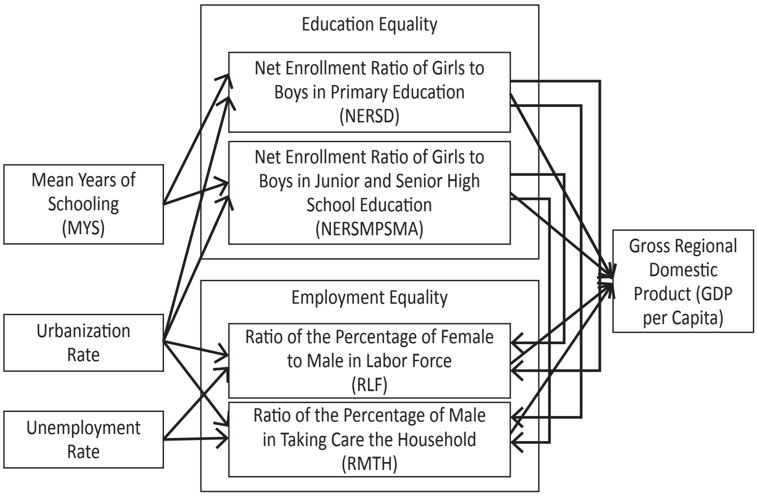


Chart 1. Conceptual Framework

3. METHOD

The secondary data used were BPS publications (Indikator Kesejahteraan Rakyat, Statistik Ketenagakerjaan, Statistik Indonesia, and Susenas 2014). The graphs and tables were used to describe the condition of the gender gap in Indonesia and the path analysis was used to determine the direct and indirect influence of the urbanization, unemployment, MYS, gender gap in education and employment to Indonesian GDP in 2014. The operational definitions of variables were: GDP per capita is Gross Domestic Product (GDP) per capita at current prices (000 Rupiah), NERSD: Ratio Net Enrollment Rate of girls to boys in primary education level, NERSMPSMA: Ratio Net Enrollment Rate of girls to boys in junior and senior high school education, RLF: Ratio of the percentage of female to male in the labor force, MYS: The average length of schooling of the population aged 15 years and over (Year), Urbanization: The percentage of the population living in urban areas (%).

The base models for the path analysis by SPSS software are:

$$\text{GDP per capita} = \beta_0 + \beta_1\,\text{NERSD} + \beta_2\,\text{NERSMPSMA} + \beta_3\,\text{RLF} + \beta_4\,\text{RMTH} + e_i \tag{3.1}$$

$$\text{NERSD} = \beta_0 + \beta_1\,\text{MYS} + \beta_2\,\text{Urbanization} + e_i \tag{3.2}$$

$$\text{NERSMPSMA} = \beta_0 + \beta_1\,\text{MYS} + \beta_2\,\text{Urbanization} + e_i \tag{3.3}$$

$$\text{RLF} = \beta_0 + \beta_1\,\text{NERSD} + \beta_2\,\text{NERSMPSMA} + \beta_3\,\text{Urbanization} + \beta_4\,\text{Unemployment} + e_i \tag{3.4}$$

$$\text{RMTH} = \beta_0 + \beta_1\,\text{NERSD} + \beta_2\,\text{NERSMPSMA} + \beta_3\,\text{Urbanization} + \beta_4\,\text{Unemployment} + e_i \tag{3.5}$$

4. RESULT

The description about GPD per capita, urbanization rate, Net Enrollment Ratio (NER) of girls to boys in primary school, NER of girls to boys in Junior and Senior High School, Mean Years of Schooling (MYS), ratio of the percentage of female to male in labor force, ratio of the percentage of male in taking care the household, unemployment rate of seven islands which consists of many provinces in Indonesia was shown in Chart 2 until 9. Chart 4, 5 and 6 described the condition of gender gap in education. The information about gender gap in employment was shown by Chart 7, 8, and 9.

There was disparity of GDP per capita in seven islands in Indonesia (Chart 2). The highest GDP per capita was in Java and the lowest one is in Kepulauan Maluku. Imbalance value of GDP per capita occurred in the west and the east of Indonesia. The highest urbanization rate was in Java (Chart 3). Analogically based on Chen (2004), there was gender equality in Java.

The deepest girls to boys participation in primary school was happened in Papua and in the secondary level (junior and senior high school) girls to boys participation was reached lowest value in Kepulauan Maluku (Chart 4 and 5). Papua also had the lowest mean years of schooling or educational attaintment comparing with other islands (Chart 6).

Based on Chart 7, the women work involvement had high value in the east respectively: Kepulauan Nusa Tenggara, Papua and Kepulauan Maluku. The most interested result was in Chart 8 which ratio of the percentage of male in taking care the household achieved the top and extreme value in Kepulauan Nusa Tenggara. On the contrary, Kepulauan Nusa Tenggara had the bottom value of unemployment rate (Chart 9). It was not endorsed Chen's research result (2004) that women labor force participation rates may actualy increase in period of high unemployment.

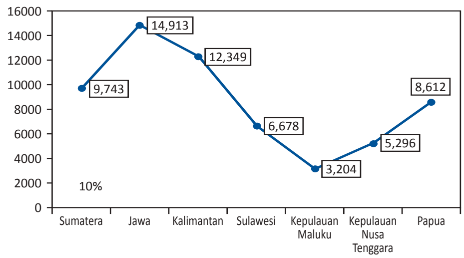


Chart 2. GDP per Capita

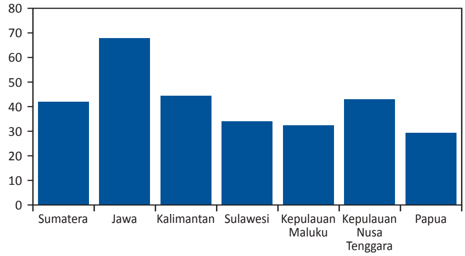


Chart 3. Urbanization Rate

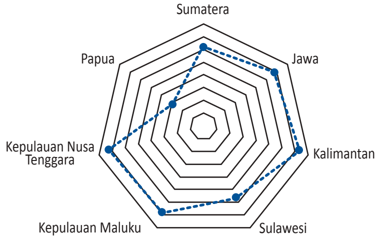


Chart 4. Net Enrollment Ratio of Girls to Boys in Primary Education

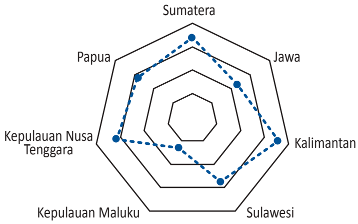


Chart 5. Net Enrollment Ratio of Girls to Boys in Junior and Senior High School Education

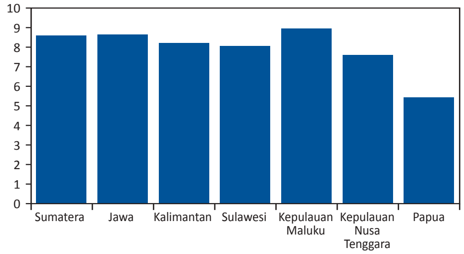


Chart 6. Mean Years of Schooling

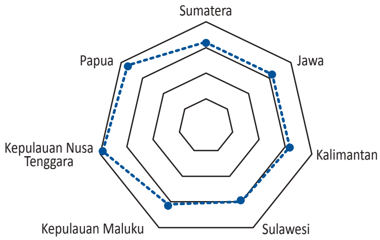


Chart 7. Ratio of the Percentage of Female to Male in Labor Force

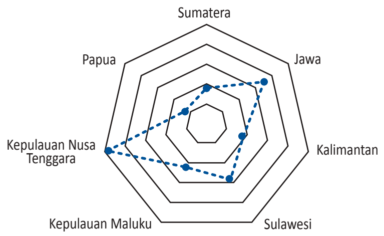


Chart 8. Ratio of the Percentage of Male in Taking Care the Household

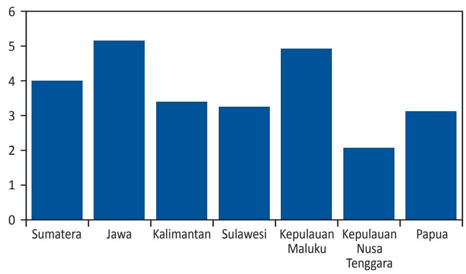


Chart 9. Unemployment Rate

The equations that were produced by Path Analysis (X_1=GDP per capita, X_2=NERSD, X_3=NERSMPSMA, X_4=RLF, X_5=RMTH, X_6=MYS, X_7= Urbanization, and X_8=Unemployment) are:

$$X_1 = -0.138\, X_2 - 0.163\, X_3 - 0.077\, X_4 - 0.143\, X_5 \quad (4.1)$$

$$X_2 = -0.128\, X_6 + 0.044\, X_7 \quad (4.2)$$

$$X_3 = -0.052\, X_6 - 0.247\, X_7 \quad (4.3)$$

$$X_4 = -0.119\, X_2 - 0.073\, X_3 + 0.092\, X_7 - 0.567^*\, X_8 \quad (4.4)$$

$$X_5 = 0.064\, X_2 - 0.276^*\, X_3 + 0.497^*\, X_7 - 0.728^*\, X_8 \quad (4.5)$$

**= insignificant at 10% level*

Generally, the effect of unemployment to gender equality in employment was statsitically insignificant. Urbanization influence the women participation in labor market significantly but the influence of urbanization to men rate of taking care the household was insignificant. It might be an exception for the phenomenon in Kepulauan Nusa Tenggara that was not endorsed Chen's result. The effect of junior and senior high schools girls participation also was insignificant to men rate of taking care the household. Other causality correlations were statistically significant. The visualization of those equations was shown path diagram in chart 10.

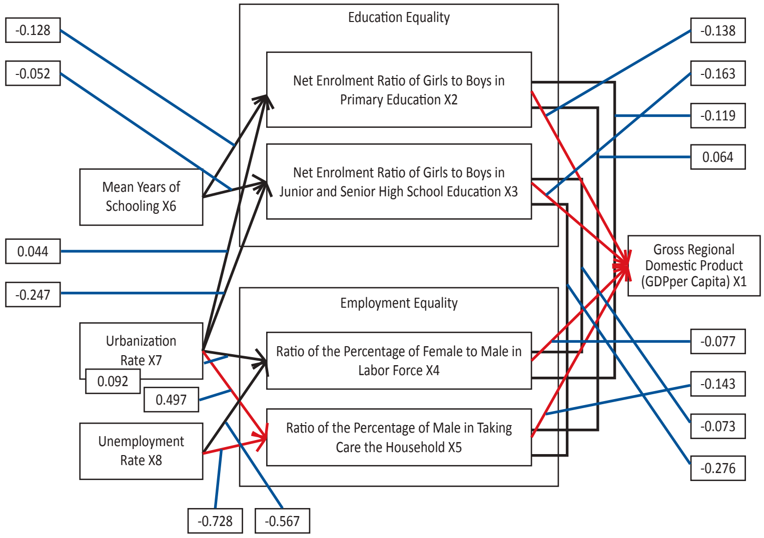


Chart 10. Path Diagram

The direct and indirect effect of variables based on equation 4.1 until 4.5 were presented in Table 1. The effect value less than 0.05 could be neglected. The highest effect reflected the most effective path.

Table 1 The Direct and Indirect Effect

No.	Path	Direct Effect	Indirect Effect	No.	Path	Direct Effect	Indirect. Effect
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	X_2-X_1	0.1380	-	13	X_6-X_3-X_4-X_1	-	0.0003
2	X_2-X_4-X_1	-	0.0092	14	X_6-X_3-X_5-X_1	-	0.0021
3	X_2-X_5-X_1	-	0.0092	15	X_7-X_2-X_1	-	0.0061
4	X_3-X_1	0.1630	-	16	X_7-X_2-X_4-X_1	-	0.0004
5	X_3-X_4-X_1	-	0.0056	17	X_7-X_2-X_5-X_1	-	0.0004
6	X_3-X_5-X_1	-	0.0395	18	X_7-X_3-X_1	-	0.0403
7	X_4-X_1	0.0770	-	19	X_7-X_3-X_4-X_1	-	0.0014
8	X_5-X_1	0.1430	-	20	X_7-X_3-X_5-X_1	-	0.0097
9	X_6-X_2-X_1	-	0.0177	21	X_7-X_4-X_1	-	0.0071
10	X_6-X_2-X_4-X_1	-	0.0012	22	X_7-X_5-X_1	-	0.0711
11	X_6-X_2-X_5-X_1	-	0.0012	23	X_8-X_4-X_1	-	0.0437
12	X_6-X_3-X_1	-	0.0085	24	X_8-X_5-X_1	-	0.1041

The Effect of NERSD (X_2), NERSMPSMA (X_3), RLF (X_4), and RMTH (X_5) to GDP per capita (X_1)

The direct effect of NERSD (X_2) to GDP per capita (X_1) is higher than indirect effect of NERSD (X_2) to GDP per capita (X_1) through RLF (X_4) or RMTH (X_5) {0.1380>0.0092}. Hence NERSD (X_2) was more effective to estimate GDP per capita (X_1) directly. So was NERSMPSMA (X_3) to estimate GDP per capita (X_1). NERSMPSMA (X_3) was more effective influence GDP per capita (X_1) directly (0.1630>0.0056 and 0.1630>0.0395). Both RLF (X_4) and RMTH (X_5) explained GDP per capita (X_1) directly which the effect value was 0.0770 and 0.1430.

The Effect of MYS (X_6), Urbanization (X_7), Unemployment (X_8) to GDP per capita (X_1)

There wasn't direct influence of MYS (X_6) to GDP per capita (X_1). The 9th path to 14th path showed the indirect effect of MYS (X_6) to GDP per capita (X_1). The 9th path was the best path because of the highest effect (0.0177) comparing with The 10th to 14th path although the effect was less than 0.05. It meaned that MYS (X_6) caused GDP per capita (X_1) by way of NERSD (X_2). By the same way, the best indirect effect of urbanization (X_7) to GDP per capita (X_1) was through RMTH (X_5) {0.0711}. RMTH (X_5) was also the best moderator variable to measure indirect effect of unemployment (X_8) to GDP per capita (X_1) {0.1041}.

The Best Direct and Indirect Path

Based on the effect value on table 1 path 1, 4, 7, 8, 22, and 24, it can be concluded that the highest influence to GDP per capita (X_1) to the lowest one respectively came from NERSD (X_2) directly, NERSMPSMA (X_3) directly, RMTH (X_5) directly, RLF (X_4) directly, unemployment (X_8) through RMTH (X_5), urbanization (X_7) through RMTH (X_5). The best path diagram was shown chart 10 with red bold line.

Goodness of fit model was reflected by common determination coefficient (R^2 = 63.4%). It meant that model was able to explain 63.4% GDP per capita's variance. While the remainder value, 36.6% GDP per capita's variance was explained by other variables that were not included in this research.

5. CONCLUSIONS

This first research conclusion was the gender gap in education generally did not occur, contrasting in employment. Whereas it was indicated that there was educational gender inequality in Papua and Kepulauan Maluku. The most interested result was Kepulauan Nusa Tenggara had the highest ratio of the percentage of male in taking care the household but the lowest unemployment rate comparing with other islands.

The second one was Indonesian GDP was influenced by the significantly direct effect of gender gap in education and employment, significantly indirect effect of urbanization and unemployment through men rate of taking care the household and unsignificantly indirect effect of MYS through girls participation in primary education.

6. RECOMMENDATION

The government should attempt gender equality particularly in Papua, Kepulauan Maluku and Kepulauan Nusa Tenggara. The next research should not use means years of schooling as a measurement of gender equality in education.

REFERENCES

Amu, N. J. (2005). The Role of Women in Ghana's Economy. Friedrich Ebert Stiftung. Ghana.

Beneria, L. and Sen, G. (1981). Accumulation, Reproduction, and Women's Role in Economic Development: Boserup Revisited. *Journal of Women in Culture and Society 1981, 7 (2), 279-298*.

Bradshaw, S., Castellino, J. and Diop, B. (2013). Women's role in economic development: Overcoming the constraints. Background paper for the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda. Sustainable Development Solution Network.

Chen, D. H. C. (2004). Gender Equality and Economic Development: The Role of Information Communication Technologies. World Bank Policy Research Working Paper 3285.

Dollar, D and Gatti, R. (1999). Gender Inequality, income, and Growth: Are Good Times Good for Women? Policy Research Report on Gender and Development, Working Paper Series, No. 1, The World Bank, May.

Klasen, S., and Lamanna, F. (2009). The Impact of Gender Inequality in Education and Employment on Economic Growth: New Evidence for a Panel of Countries. *Feminist Economics 15(3), 91–132*.

Mammen, K., and Paxson, C. (2000). Women's Work and Economic Development. *The Journal of Economic Perspectives, Vol. 14, No. 4, pp. 141-164. American Economic Association.*

Organisation for Economic Cooperation and Development (OECD) (2008). Gender and Sustainable Development. Maximising The Economic Social and Environmental Role Of Woman.

World Bank (2003). Gender Equality and the Millennium Development Goals. Washington, D.C., April.

World Economic Forum (2014). The Global Gender Gap Report2014. World Economic Forum

Fertility of Women in the BDHS 2014: Evidence from National Population Based Survey

Shongkour Roy*
Population Council, Dhaka, Bangladesh– sankar1604@gmail.com

Sharif Mohammed Ismail Hossain
Population Council, Dhaka, Bangladesh– sharifhossain@popcouncil.org

Abstract

The aim of this study was to examine the fertility of women age 15 to 49 using data from Bangladesh Demographic and Health Survey (BDHS) 2014- a survey of women born between 1964 to 1999. On an average have 2.3 children of women in BDHS 2014, more than 90% of them give birth to at least one child by aged 49 and average age of first birth 18 years old. Fertility of women strongly differ by education. Women had no children with secondary 50.3% and never attended school 8.4%, those with secondary were six times as likely as those who never attended school to have no children and this pattern being stronger among urban compared to rural women. Fertility was delayed as education increased. Women fertility was also related to religion group, but this factor were not strength than those related to educational attainment.

Keywords: Fertility; BDHS; Women.
JEL Classification: J13 and J12

1. INTRODUCTION

Fertility is the major component of population dynamics that determine the size, structure and composition of populations in any country of the world (NIPROT, 2016). A major changes from the 1975 have affected women decision to have children in Bangladesh. Women childbearing decision affected due to increase the universal provision of family planning services, rates of women educational attainment and urbanization. Differentials in fertility behaviour and fertility levels in different areas and among population strata or characteristics have been among the most pervasive findings in the demography (Cochrane, 1979). Bangladesh has seen substantial improvements in its reproductive health outcomes, such as in the reduction of maternal and infant mortality, the increase in the contraceptive prevalence rate, and health service utilization among married women (NIPROT, 2016). The total fertility rate in the Bangladesh was remarkable change from 6.3 births per women in 1975 to 2.3 births per women in 2014. But total fertility rate was still high compared to that of other developing countries. Ahamed in 1992 were examined differentials of fertility in terms of selected demographic and socio-economic characteristics of the ever married women and their husbands. He showed that age of women has the highest effect on number of children ever born. Ramesh (2010) was also found that demographic, socio-economic, and cultural factors affecting fertility differentials in Nepal. He also exhibited that age at first marriage of women was a strong predictor of children ever born. Desalew Zelalem et al. was conducted a study about the level and patterns of fertility among women in Kersa Demographic Surveillance and Health Research Center. This authors used the follow up data 2008-2012 and found that fertility rate was higher in rural residents and illiterate women than in urban residents and literate women.

Kasey used the NLSY79 to investigate the wage-earning implications of delaying first birth in 2003. This article found that an annual three-percent wage premium existed for each year of delayed motherhood. Delayed childbirth also correlated with high test scores, education, and professional status of the mother. Our study focuses on differences in fertility patterns by educational attainment, religion and residence. The analysis is descriptive and does not attempt to explain why fertility patterns differ across women.

2. DATA SOURCE AND METHODS

This paper used data from the Bangladesh Demographic and Health Survey (BDHS) 2014, consisting of nationally representative sample surveys. Details of the design and data collection methods have previously been published (NIPROT, 2016). Briefly, BDHS 2014 was a cross-sectional study designed to obtain information on key health indicators in Bangladesh i.e. fertility, family planning, infant, child, adult, and maternal mortality, maternal and child health, nutrition, and knowledge of HIV/AIDS and other sexually transmitted infections. The data used in this study is publicly available.

We conducted a secondary data analysis using data of BDHS 2014. Univariate and bivariate analyses were conducted to find out the sample characteristics and evaluate the differences in fertility patterns by educational attainment, religion group, and residence.

3. RESULTS

A total of 17,863 women were participated in the BDHS 2014. A majority of the women lived in the rural areas (71.74%). The participants' higher educational level was low, and 37.39% of the participants had secondary and 24.94% had no education. Utmost one tenth participants were non-Muslim and 10.16% had no child. Among the participants 29.28% had two child and 7.79% had five or more child (Table 1).

Table 1. Sample Characteristics [N=17,863]

Characteristics	Weighted percentage
Education	
No education	24.94
Primary	29.16
Secondary	37.39
Higher	8.51
Religion	
Muslim	90.11
Non-Muslim	9.89
Residence	
Urban	28.26
Rural	71.74
Parity	
No Child	10.16
One child	23.78
Two child	29.28
Three child	19.13
Four child	9.86
Five or more child	7.79

Table 2 shows the fertility outcomes of women from age 15 to age 49, by educational attainment, religion, and residence. Non-Muslim women had lower educational attainment than Muslim women. The minority groups were overrepresented among those with secondary educational attainment and underrepresented among those higher education. Among the women had no child, urban women had higher percentage of no child compared to rural women (116.8% versus 9.56). When we compared this picture with women had five or more children the scenario has changed and rural women had more child than urban (8.95% versus 4.84). The mean number of child were 2.23 in BDHS 2014, those with no education had the highest mean number of child (3.09) and lowest (1.24) in the higher education.

Table 2. Fertility Outcomes of Women from Age 15 to Age 49, by Educational Attainment, Religion and Residence

Characteristics	Percent distribution of people by number of children						Mean number of living Child
	No child	One Child	Two Child	Three Child	Four Child	Five or more Child	
Total, 15 to 49 age	**10.16**	**23.78**	**29.28**	**19.13**	**9.86**	**7.79**	**2.23**
No education	8.40	11.92	20.05	34.04	45.37	56.44	3.09
Primary	21.38	24.38	29.56	34.54	34.46	32.45	2.42
Secondary	50.32	49.95	41.92	28.17	19.27	10.76	1.73
Higher	19.90	13.76	8.47	3.25	0.90	0.35	1.24
Muslim	**10.30**	**23.64**	**28.53**	**19.14**	**10.05**	**8.34**	**2.25**
No education	8.34	11.96	19.17	33.33	45.04	56.74	3.15
Primary	20.73	24.85	30.46	34.98	34.45	32.04	2.45
Secondary	50.77	50.04	42.42	28.49	19.59	10.85	1.74
Higher	20.17	13.15	7.96	3.20	0.91	0.36	1.22
Non-Muslim	**8.88**	**25.09**	**36.06**	**19.07**	**8.16**	**2.73**	**2.01**
No education	9.08	11.54	26.37	40.61	49.11	48.10	2.59
Primary	28.26	20.37	23.14	30.60	34.61	43.68	2.20
Secondary	45.59	49.11	38.31	25.25	15.60	8.22	1.67
Higher	17.07	18.98	12.18	3.53	0.68	0.00	1.39
Urban	**11.68**	**27.83**	**30.50**	**17.38**	**7.77**	**4.84**	**1.96**
No education	6.91	10.76	14.89	27.97	43.05	55.99	2.83
Primary	23.28	21.86	23.65	31.15	31.75	27.86	2.14
Secondary	42.91	45.50	44.18	32.56	23.74	15.24	1.72
Higher	26.90	21.88	17.28	8.32	1.46	0.91	1.34
Rural	**9.56**	**22.19**	**28.80**	**19.82**	**10.69**	**8.95**	**2.67**
No education	9.12	12.50	22.20	36.14	46.04	56.54	3.16
Primary	20.47	25.62	32.03	35.71	35.24	33.43	2.52
Secondary	53.88	52.14	40.97	26.65	17.99	9.80	1.73
Higher	16.53	9.74	4.80	1.50	0.73	0.23	1.23

Table 3 also presented descriptive statistics of women fertility age and birth spacing by number of children, educational attainment, religion group, and residence. Women with higher education had child at later aged. The average age at first birth was 18.98 years for women with one child, 17.98 years for women with two children, and 17.24 years for women with three children. The births of women in the latter group were spaced more closely than those of women with two children. On average, the time between first and second births were 5.69 years for women with two children, compared with 4.68 years for those with three children. When we looked birth spacing with no education by two and three child women, we found that the mean time between first and second births were 6.20 years for women with two children, compared with 4.94 years for those with three children. These pattern were similar by religion and education.

Table 3. Women Age at Birth and Birth Spacing by Number of Children, Educational Attainment, and Residence

Characteristics	All	Education				Religion		Residence	
		No education	Primary	Secondary	Higher	Muslim	Non-Muslim	Urban	Rural
Among those with one Children									
Average age at first birth	18.98	18.92	17.91	18.47	22.27	18.86	20.07	19.69	18.52
Among those with Two Children									
Average age at first birth	17.98	17.49	17.09	17.90	21.94	17.89	18.76	18.59	17.65
Average age at 2nd birth	23.67	23.69	22.76	23.39	27.47	23.57	24.42	24.45	23.23
Year between 1st and 2nd child	5.69	6.20	5.67	5.49	5.53	5.68	5.66	5.86	5.58
Among those with Three Children									
Average age at first birth	17.24	16.97	16.91	17.48	20.86	17.19	17.69	17.52	17.11
Average age at 2nd birth	21.92	21.91	21.41	22.08	25.42	21.87	22.34	22.11	21.82
Average age at 3rd birth	26.77	26.91	26.19	26.88	30.19	26.76	26.84	27.20	26.56
Year between 1st and 2nd child	4.68	4.94	4.50	4.60	4.56	4.68	4.65	4.59	4.71
Year between 2n and 3rd child	4.85	5.00	4.78	4.80	4.77	4.89	4.50	5.09	4.74
Year between 1st and 3rd child	9.53	9.94	9.28	9.40	9.33	9.57	9.15	9.68	9.45

4. CONCLUSION

Fertility patterns mostly differ by educational attainment than others factors i.e. religion group and residence. Fertility rate was higher in rural residents and illiterate women than in urban residents and literate women. Since education was major factor to delay fertility and we need improve our girl's education in among the communities in Bangladesh.

REFERENCES

NIPORT, Mitra and Associates, and ICF International (2016). Bangladesh Demographic and Health Survey 2014, Dhaka, Bangladesh, and Rockville, Maryland, USA.

Cochrane, S. H. (1979). Fertility and education. Baltimore, The John Hopkins University Press.

Ahamed, M. M. U. (1992). Fertility differentials in Bangladesh. Department of Mathematical Science, Ball State University, USA.

Adhikari, R. (2010). Demographic, socio-economic, and cultural factors affecting fertility differentials in Nepal. BMC Pregnancy and Childbirth, 10:19.

Zelalem et al. (2015). The level and patterns of fertility among women in Kersa Demographic Surveillance and Health Research Center (KDS-HRC) Field site, Kersa District, East Ethiopia. Fertility Research and Practice, 1:18.

Kasey, S. B. (2008). Understanding the returns to delayed childbearing for working women. American Economic Review, 98(2), pp. 403–407.

The Impact of Antenatal Care Program to Birth Weight on Pregnant Mother in Indonesia: Instrumental Variable Regression Approach

Choerul Umam*
Master Student of Economics Development Department, Faculty of Economics and Business,
Universitas Gadjah Mada, Yogyakarta, Indonesia – choerul.umam@mail.ugm.ac.id

Abstract

Development is not only about increasing the economic growth, but also increasing the quality of education and health. Maternal and child health is one of the national development goals which already listed as one of the Millennium Development Goals (MDGs) and now has become the Sustainable Development Goals (SDGs). This paper evaluates the impact of antenatal care check-up program - ANC on pregnant mother to birth weight, which it's one of the indicators of the healthy baby born. In this study, characteristics of the recipient of ANC are pregnant mothers who are measured by her weight and her blood pressure, checked of urine, taken a blood sample, received the tetanus injection and the iron tablets. This study using data from The 2012 Indonesia Demographic and Health Survey. This study also uses instrumental variable regression to deal with endogeneity of antenatal care program. The results confirm that antenatal care program (ANC) can increase the birth weight about 366.45 grams. However, the probability of pregnant mother got antenatal care program (ANC) in rural areas 1.22% less than urban areas.

Keywords: Human development; Antenatal care check-up program (ANC); Birth weight; Instrumental variable regression.

1. BACKGROUND

Maternal and child health is one of the national development goals which already listed as one of *the Millennium Development Goals* (MDGs) and now has become *the Sustainable Development Goals* (SDGs). As effort to improve the quality of women and families life, a life-cycle approach becomes important. Infant and child mortality rates also describe of health development, quality of life of the people, and also are used to monitor and evaluate programs or policies of health and population in a country's level (BPS, BKKBN, Kemenkes-MOH, and ICF International, 2013). Health and education as a human capital is also importance on economic development process (Todaro & Smith, 2011).

Health development is directed to raising awareness, willingness and ability of people to live healthy independently so that the achievement of the highest public health degrees can be realized. In practice, health development organized by the principles of humanity, empowerment and self-reliance as well as fair and equitable by emphasizing the beneficial aspects particularly for vulnerable groups such as mothers, infants, children, the elderly and poor families

To improve maternal health status, health center, and a referral hospital hold a variety of health actions, both promotive, preventive, curative and rehabilitative. Such efforts form of health services to pregnant women, deliveries by skilled health personnel, management of complications, family planning and reproductive health. Every expectant mother is expected to run a healthy pregnancy, childbirth safely and give birth to a healthy baby. Therefore, every pregnant woman should be able to access a health facility easily to obtain an appropriate care standards, including the detection of possible problems/diseases that can negatively impact the health of the mother and fetus.

Malnutrition on pregnant women also remains a public health problem that needs special attention. Less intake of iron in women especially pregnant women can lead to anemia which would increase the risk of bleeding and gave birth to a baby with low birth weight. Besides the lack of intake of iron, anemia can also be caused by intestinal worms and malaria. Other nutritional problems is Less Energy Chronic (Kurang Energi Kronik - KEK) and consumption of iodized salt are still low. The results of Basic Health

Research 2007 (Riset Kesehatan Dasar – Riskesdas 2007) mention that fertile women age are at risk of less energy chronic about 13.6% and only 62.3% of households are enough consuming iodized salt (Kemenkes RI, 2010b).

There are some problems that could affect pregnancy, fetal growth and can even cause complications of pregnancy and childbirth which can threatening the life of the mother and the baby also can affect growth and development of the fetus, such as less energy, anemia, lacking iodine, HIV/AIDS, Malaria, tuberculosis, and others.

In fact, the service of antenatal should be implemented in a comprehensive manner, integrated, and qualified so that problems/diseases can be detected and treated early. Antenatal services through integrated, pregnant women will get a more thorough service and integrated, so that the rights of reproduction can be met, ***missed opportunity***[1] can be avoided and the health service can be organized in a more efficient and effective way.

Maternal health is a very important component in reproductive health since the whole other component is strongly influenced by maternal health. If the Mother is healthy, they will giving birth a healthy baby who will become strong generation also creating healthy and happy families.

Based on Indonesia Demographic and Health Survey (SDKI) 2012 results, maternal mortality rate (MMR) in Indonesia reached 359 per 100,000 live births and infant mortality rate (IMR) reached 32 per 1000 live births. Although MMR in Indonesia are still far from the target of the MDGs 2015 (102 per 100,000 live births) and IMR is also still far from the target of the MDGs (17 per 1000 live births) so we still need hard work from all the components to achieve the target (BPS, BKKBN, Kemenkes-MOH, and ICF International, 2013).The results of Riskesdas in 2013 that the prevalence of infants with low birth weight (LBW) reduced from 11.1% in 2010 be 10.2% (Kemenkes RI, 2013b). Although there are improvements but continues to strive to the lower prevalence figures. The trend decline in the MMR and IMR indicating the success of exertion in Indonesia to reach the target of MDG's. However the number of such figures – especially IMR is still high. For its wide range of activities and best practices have been implemented and developed including family planning program (Keluarga Berencana-KB).

One of the risks from decreasing in maternal health is baby born with low birth weight. Babies born with a low birth weight (LBW) is infants who born with weight less than 2500 grams so that affect infant mortality (Kemenkes RI, 2010a). According to the Ministry of Health, the cause of the occurrence of Low Birth Weight may be due to pregnant women with anemia condition, less intake of nutrients content in time, less energy chronic, or less months. Generally, babies born with LBW less able to adapt to a new environment, so may result in fence of their growth and development, it can even interfere with his survival. In addition it will also increase the risk of pain and death of infants due to susceptible to respiratory tract infections down, learning disorders, behavior problems and so forth. In order to improve maternal health status who will ultimately contribute to a decrease in maternal and child mortality. (Kemenkes RI, 2010b).

Based on the background, then wants to know the impact of antenatal care check-up program to birth weight on pregnant women in Indonesia. The question research are how the impact of antenatal care program on pregnant women against the weight of the baby is born and how much the probability of a pregnant woman's getting antenatal care program by regions (rural and java).

2. STUDY LITERATURE

Hidayati (2005) was investigated the chronic lack of energy (KEK) and anemia in pregnant mothers as a risk factor for the incidence of low birth weight (LBW) in Mataram City of West Nusa Tenggara. The results showed that subjects who are exposed KEK has a probability of 4.71 times more likely to give birth to low birth weight and anemia has exposed a probability of 3.70 times more likely to have low birth weight than those not exposed. While the subjects were exposed to KEK and anemia have a higher probability of 5.5 times more likely to have low birth weight than those not exposed to both.

1 *Missed opportunity* was the incident in which pregnant women do not receive the services which should be provided at the time of contact with the health care personnel.

Choirunnisa (2010) was examined the relationship weight, arm circumference differences and hemoglobin levels of pregnant women trimester III was born in Surakarta. Research results show that there is relations between hikes weight, differences arm circumference, and hemoglobin concentration fiber on the third trimester pregnant women genesis birth weight was low.

Hermawan (2012) was researched on growth rate of low birth weight (LBW) and the factors that affected it. With multivariate and bivariate analysis (logistic regression) obtained that gestational age, socioeconomic level, full breast feeding, the mother's level of education and the status of working moms were not proven to be related to better growth speed.

3. DATA AND METHODS

The main source of the micro data used in this research is The Indonesia Demographic and Health Surveys (Survei Demografi dan Kesehatan Indonesia, SDKI) in 2012. The Indonesian DHS (SDKI) is a longitudinal socio-population survey conducted in Indonesia was held in 1987, 1991, 1994, 1997, 2000, 2007, and 2012. SDKI collects data on individual and household respondents covered women and men of reproductive age, and recorded information on childbirth, fertility, marriage and sexual activity, mortality, family planning, and reproductive health (BPS, BKKBN, Kemenkes-MOH, and ICF International, 2013). Data of birth weight used in this research is less 4.000 grams, to reduce the outlier data, by reasoning that the normal of birth weight about 2.500-4.000 grams (Kemenkes RI, 2010a). This research methodology used in this paper is Instrumental Variable (IV) regression to evaluate the effects of antenatal care program on birth weight while IV estimation is used to deal with endogeneity of ANC and also correct for selection bias based on unobservable characteristics. In addition, IV estimates the effect of the treatment on those individuals whose behavior is affected by treatment. That is, IV estimation provides an estimate of the causal effect for those individuals who change the treatment status because of the instrument.

The analysis used are descriptive analysis and econometric analysis using the *instrument variable regression (IV regression)*. There are 8 (eight) of the variables used in this study with variables paramedic as *instrument variables*, with details as follows:

Table 1. The Variables Used in This Research

Variable	Description
ANC	*Dummy* variable of *Antenatal Care* Check-up *Program* 1: if pregnant women was checked her weight and blood pressure, checked urine and blood (Hemoglobin), got iron pills (Fe tablets), and got a TT (Tetanus Toxoid) injection during her last pregnancy, 0: others.
rural	Dummy village 1: rural residence 0: urban residence
Java	Dummy Java 1: reside in Java and Bali 0: reside outside Java and Bali
paramedic	Dummy paramedic (health workers) 1: was checked by doctor/general practitioner/obstetrician specialist/ midwife/ nurse/village midwife (bidan desa) during her last pregnancy; 0: others
age_group	Age groups of pregnant women
educ	The highest level of maternal education
wealth	The level of prosperity (already calculated by DHS)
BBL (berat bayi lahir)	Birth weight of the last baby born (grams)

The equations used in the *instrumental variable regression* is as follows:

First Stage:

$$ANC_i = \beta_0 + \beta_1 \text{ paramedic}_i + \mu_i \qquad (1)$$

Second Stage:

$$BBL_i = \alpha_0 + \alpha_1 \widehat{ANC}_i + \alpha_2 X_i + \varepsilon_i \qquad (2)$$

In the first stage, dummy variables and the dependent becomes the ANC we use *dummy* paramedic (health workers) variable as a variable for the ANC with the instrument in addition to the health workforce as outside category. Explanatory variables as variables then in capturing individual and household characteristics like wealth index, a typology of the area of residence (urban and rural, Java-Bali and Outside Java-Bali), age group of the mother, educational level of the mother. In the second stage, the parameters of which we get from estimated equations (1) and variables, we put as explanatory variables in the regression OLS with variable birth weight (BBL) as a dependent variable.

4. RESULTS

Paramedics used in this study is for pregnant women examined by medical personnel that is checked by a doctor / obstetrician / nurse / midwife / midwives. From table 2 below, it can be seen that most pregnant women have been checked by health care personnel during her pregnancy about 85 percent, and only 15 percent have not checked by paramedics. This indicates that the service of obstetrics in pregnant mother was almost equally, although in urban areas are greater than rural areas. It is also shown that pregnant mother was checked by paramedic in Java and Bali are larger than outside Java and Bali but the differences is not large.

Table 2. Description of Antenatal Care Program in Indonesia (percent)

Description	Category	Urban	Rural	Java & Bali	Outside Java & Bali
Checked by paramedic during last pregnancy	Yes	85,32	84,53	89,08	83,23
	No	14,68	15,47	10,92	16,77
Got full antenatal care program during last pregnancy	Yes	17,68	13,61	21,88	13,16
	No	82,32	86,39	78,12	86,84

Source: Indonesia DHS data 2012 (calculated by author)

Antenatal care program in this research are pregnant women who were checked her weight, blood, urine, and blood pressure (tension), got iron pills (Fe tablets), and got a tetanus injection (*Tetanus Toxoid*) during her last pregnancy. From Table 2 above, it shown that there are still many pregnant women who have not gotten the sixth antenatal care services. It is shows that about 84% of pregnant women are not getting the midwifery service standard during the period of maternity which is rural areas is larger than urban areas. Pregnant women who lived outside Java and Bali more did not get the standard of obstetric care than those who live in Java and Bali with their respective percentages of 87% and 78%.

To correct the endogeneity factor variable, then we use *IV regression* to analyze the impact of antenatal care program to birth weight in Indonesia. Results from the first stage equation *IV regression* is shown at Table 3 below that the chances of pregnant women get the standard obstetric care if checked by paramedics at 17.55% greater if not checked by trained paramedics. While pregnant women who live in rural areas 1.22% less chances than living in urban areas. Pregnant women who live in Java and Bali 7.36% greater chance of getting a standard obstetric care than staying outside Java and Bali.

The higher the age and the more educated a pregnant mother, then the greater chance of getting the standard obstetrics care. This study shows that the more experienced and have enough knowledge about the health of mothers and children so will be greater chances of getting the standard obstetrics care. The interesting thing also on conditions of prosperity of a pregnant mother. The results show that a poor mother could get pregnant midwifery service standard of almost 2%. This shows that government programs to protect the poor already proven in particular in the field of health of pregnant women, although a standard obstetric care services should still be done regardless of socioeconomic status.

Table 3. First Stage Least Square Estimation

Dependent Variable: ANC	Coefficient	Standard Error
[1]	[2]	[3]
Paramedic	.1755041 ***	.0083113
Rural	-.0122071 *	.0066102
Java	.073612 ***	.0069727
Age Group		
20-24	.0369513 **	.0182943
25-29	.0533048 ***	.0178647
30-34	.0578409 ***	.0180474
35-39	.0895052 ***	0184964
40-44	.0478983 **	.0204042
45-49	.06141 *	.0324036
Educ	.0334484 ***	.0050444
Wealth		
Poorer	.0195915 **	.0092336
Middle	-.0042175	.0096938
Richer	-.0068888	.0103552
Richest	.0088632	.0115917
Constanta	-.1289488 ***	.0214029
Observation	14.457	

Note: * significant at level 10%, ** significant at level 5%, and *** significant at level 1%

From Table 4 below, it can be shown that the difference of birth weight between pregnant women who get antenatal care program and did not get ANC program during the pregnancy period about 366.45 grams. This shows that the importance of the standard obstetrics care service standard program for pregnant women, especially in improving the health of mothers and children, as well as reducing the risk of infant death.

Pregnant women living in rural areas gave birth to babies with birth weight 17.86 grams heavier than the pregnant women who live in urban areas. This could indicate that food menu is still fresh and has not been contaminated by the preservative substances that can affect the health of the mother and the fetus. In addition, the condition of the water and the environment in the countryside is still fresh so that it is possible to effect the maternal and child health and also to the baby birth weight (Wang, et al., 1997), (Ritz & Yu, 1999), (Ritz, et al., 2000), (Gouveia, et al., 2004), (Mannes, et al., 2005), (Salam, et al., 2005), (Sram, et al., 2005), (Aguilera, et al., 2009), (Darrow, et al., 2011).

Quite an interesting thing happen to pregnant women who lived in Java & Bali, that the weight of babies who are born smaller 59.27 grams compared pregnant women living outside the island of Java & Bali. This is made possible by the quality of the air and the environment in Java are not conducive where industry and economy activities is still centered in the Java island (more 50% the economic share centered in Java), so that is possible effect on the weight of the baby to be born.

If viewed from the age group of pregnant women, it can be seen that the expectant mother at age 30-34 years gave birth to babies with birth weight 103.08 grams bigger than other age groups. This is due to the mother's experience and knowledge factors in terms of the health of the mother and her child. The higher education and the higher the level of well-being of pregnant women will also have an effect on the weight of the baby was born.

Table 4. Instrumental Variable Regression

Dependent Variable: Birth Weight (BBL)	Coefficient	Standard Error
[1]	[2]	[3]
ANC	366.4535 ***	68.94998
Rural	17.86067 *	9.644393
Java	-59.2687 ***	11.59634
Age Group		
20-24	27.80134	26.70207
25-29	53.18517 **	26.11774
30-34	103.0882 ***	26.42069
35-39	79.90229 ***	27.40586
40-44	80.79105 ***	29.80556
45-49	-4.650496	47.35321
Educ	18.1683 **	7.645456
Wealth		
Poorer	52.43834 ***	13.57785
Middle	60.1368 ***	14.10046
Richer	62.57726 ***	15.0531
Richest	63.1852 ***	16.92066
Constanta	2911.156	29.35336
Test of excluded instrument		
F statistics	12.45	
p-value	0.000	
Sargan Test	0.000	
Observation	14.457	

Note: * significant at level 10%, ** significant at level 5%, and *** significant at level 1%

5. CONCLUSION AND LIMITATIONS

The conclusion is the antenatal care program service on pregnant mother very heavy impact to birth weight, but there are still many areas that have not yet to get this services. So that the Government especially the Ministry of Health gave more attention towards the implementation of these services, such as antenatal care increased obstetric facilities and infrastructure, equitable distribution and increased the personnel of health workers, the existence of incentives for health workers deployed in remote areas. Hoped all pregnant women both in urban and in rural areas, both in Java & Bali and outside Java & Bali gets the same midwifery services in order to improve maternal and child health as well as making the community healthier and Indonesia more prosperous.

In this study there are a few limitations: **first**, a standard obstetric care is observed only 6 (six) component, so that future research can be carried out all components of the antenatal services; **second**, the used area typologies is still limited to rural & urban, as well as Java & Bali and outside Java & Bali. So that in the future could be made more specific, suppose research based on the big island or each province in Indonesia.

REFERENCES

Aguilera, I., Guxens, M., Garcia-Esteban, R., Corbella, T., Nieuwenhuijsen, M., Foradada, C., & Sunyer, J. (2009). Association between GIS-Based Exposure to Urban Air Pollution during Pregnancy and Birth Weight in the INMA Sabadell Cohort. *Environmental Health Perspectives, 117, No. 8 (Aug 2009)*, 1322-1327.

BPS, BKKBN, Kemenkes-MOH, and ICF International. (2013). *Indonesia Demographic and Health Survey 2012.* Jakarta, Indonesia: BPS, BKKBN, Kemenkes, and ICF International.

Choirunnisa, M. L. (2010). *Hubungan Kenaikan Berat Badan, Lingkar Lengan Atas dan Kadar Hemoglobin Ibu Hamil Trimester III dengan Kejadian Berat Badan Lahir Rendah di Kota Surakarta (Bachelor Thesis).* Surakarta, Indonesia: Universitas Sebelas Maret.

Darrow, L., Klein, M., Strickland, M., Mulholland, J., & Tolbert, P. (2011). Ambient Air Pollution and Birth Weight in Full-Term Infants in Atalanta, 1994-2004. *Environmental Health Perspectives, 119, No. 5 (May 2011)*, 731-737.

Gouveia, N., Bremner, S., & Novaes, H. (2004). Association between Ambient Air Pollution and Birth Weight in São Paulo, Brazil. *Journal of Epidemiology and Community Health , 58, No. 1 (Jan 2004)*, 11-17.

Hermawan, K. (2012). *Laju Pertumbuhan Berat Badan Bayi Lahir Rendah (BBLR) dan Faktor-Faktor Yang Mempengaruhinya (Thesis Abstract).* Yogyakarta, Indonesia: Universitas Gadjah Mada.

Hidayati, M. (2005). *Kurang Energi Kronis dan Anemia Ibu Hamil Sebagai Faktor Risiko Kejadian Berat Bayi Lahir Rendah di Kota Mataram Propinsi Nusa Tenggara Barat (Thesis Abstract).* Yogyakarta, Indonesia: Universitas Gadjah Mada.

Kemenkes RI. (2010a). *Buku Saku Pelayanan Kesehatan Neonatal Esensial: Pedoman Teknis Pelayanan Kesehatan Dasar* (Revision ed.). Jakarta, Indonesia: Kementrian Kesehatan RI.

Kemenkes RI. (2010b). *Pedoman Pelayanan Antenatal Terpadu.* Jakarta, Indonesia: Direktorat Jenderal Bina Kesehatan Masyarakat Kemenkes RI.

Kemenkes RI. (2013b). *Riset Kesehatan Dasar (RISKESDAS) 2013.* Jakarta, Indonesia: Badan Penelitian dan Pengembangan Kesehatan, Kemenkes RI.

Mannes, T., Jalaludin, B., Morgan, G., Lincoln, D., Sheppeard, V., & Corbett, S. (2005). Impact of Ambient Air Pollution on Birth Weight in Sydney, Australia. *Occupational and Environmental Medicine, 62, No. 8 (Aug 2005)*, 524-530.

Ritz, B., & Yu, F. (1999). The Effect of Ambient Carbon Monoxide on Low Birth Weight among Children Born in Southern California between 1989 and 1993. *Environmental Health Perspectives, 107, No. 1 (Jan 1999)*, 17-25.

Ritz, B., Yu, F., Chapa, G., & Fruin, S. (2000). Effect of Air Pollution on Preterm Birth among Children Born in Southern California between 1989 and 1993. *Epidemiology, 11, No. 5 (Sep 2000)*, 502-511.

Salam, M., Millstein, J., Li, Y.-F., Lurmann, F., Margolis, H., & Gilliland, F. (2005). Birth Outcomes and Prenatal Exposure to Ozone, Carbon Monoxide, and Particulate Matter: Results from the Children's Health Study. *Environmental Health Perspectives, 113 No. 11 (Nov., 2005)*, 1638-1644.

Sram, R., Binkova, B., Dejmek, J., & Bobak, M. (2005). Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. *Environmental Health Perspectives, 113, No. 4 (Apr 2005)*, 375-382.

Todaro, M., & Smith, S. (2011). *Pembangunan Ekonomi - Economic Development (Bahasa Indonesia Translated)* (Jilid 1 11st ed.). Jakarta, Indonesia: Penerbit Erlangga.

Wang, X., Ding, H., Ryan, L., & Xu, X. (1997). Association between Air Pollution and Low Birth Weight: A Community-Based Study. *Environmental Health Perspectives, 105, No. 5 (May 1997)*, 514-520.

This page intentionally left blank

CPS21: Probability Theory & Statistical Modelling (3)

New Extension of Exponentiated Weibull Distribution with Properties and Application to Survival Data

Mundher Abdullah Khaleel*
Faculty of science, Universiti Putra Malaysia, 43400, Selangor, Malaysia– mun880088@gmail.com

Noor Akma Ibrahim
Faculty of science, Universiti Putra Malaysia, 43400, Selangor, Malaysia.
Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia – nakma@upm.edu.my.

Abstract

In this paper, a new continuous distribution is developed called the Weibull Exponentiated Weibull distribution, which is an extension of many other distributions like the Exponentiated Weibull, Burr type X, Exponential distributions etc. Expression is obtained for the probability density function (pdf). Various structural properties of the new distribution are also derived, such as the included quantile function, the r th moment, moment generating function as well as order statistics. The parameters are estimated using Maximum Likelihood methods. Finally, the flexibility of this distribution is illustrated by using real data set and it is shown that the new model provides a better fit than nested models.

Keywords: Quantile function; Moment; Order Statistics; Estimation.

1. INTRODUCTION

Attempts have been made to develop new models extended from familiar distributions to supply greater flexibility in modeling real-data in many applied areas as well as to improve the goodness-of-fit of the data. These extended models are developed by adding one or more parameters to the baseline model. Eugene et al. (2002) proposed and study a general class of distributions based on the logit of a beta random variable named Beta-G family distribution. They added two shape parameters to the baseline distribution. In the same way, Alzaatrah et al. (2013b) employed a new technique and proposed a general form to generate a new family called the transformed-transformer (T-X) family. More recently, Bourguignon et al. (2014) proposed and studied the generality family of a univariate distribution with two additional parameters using the Weibull generator applied to the odds ratio K(t)/[1- K(t)]. The term "generator" means that for each baseline distribution K there is a different distribution W. Many attempts have been proposed to extend Weibull distribution. Mudholkar, G. S., & Srivastava, D. K. (1993) proposed a modification of the Weibull family named Exponentiated Weibull distribution (EW) by

$$K(t,\delta,\varepsilon,\vartheta) = \left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}, \qquad \delta,\varepsilon,\theta > 0, \tag{1}$$

adding one shape parameter. The cumulative distribution function (cdf) of EW is given as and the probability density function (pdf) of the EW distribution corresponding to the (1) is given as

$$k(t,\delta,\varepsilon,\vartheta) = \varepsilon\,\vartheta\,\delta^{\varepsilon}\,t^{\varepsilon-1}\,e^{-(\delta t)^{\varepsilon}}\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta-1}. \tag{2}$$

The hazard rate $h(t,\delta,\varepsilon,\vartheta)$ of the EW distribution is

$$h(t,\delta,\varepsilon,\vartheta) = \varepsilon\,\vartheta\,\delta^{\varepsilon}\,t^{\varepsilon-1}\,e^{-(\delta t)^{\varepsilon}}\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta-1}\left\{1-\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}\right\}^{-1}. \tag{3}$$

The r th moment for EW distribution is given as

$$E(T^r) = \mu'_r = \frac{\vartheta}{\delta^r}\Gamma\left(\frac{r}{\varepsilon}+1\right)\sum_{m=0}^{\infty}\binom{\vartheta-1}{m}\frac{(-1)^m}{(m+1)^{\frac{r}{\varepsilon}+1}} \tag{4}$$

There are a few sub models as a special case of EW distribution. When ϑ=1 the model reduces to the Weibull distribution. When ε=1 the model reduces to the Exponentiated exponential Gupta, R. D., & Kundu, D. (1999). When ε=2 the model reduces to the Burr type X with two parameter Surles, J.G., & Padgett, W.J. (2001). When ε=2 and ϑ=1 the model reduces to the Rayleigh distribution. When ε=2 and ϑ=1 the model reduces to the Burr type X with one parameter Burr, (1942). When ε=2 and ϑ=1 the model reduces to the exponential distribution.

The motivations for generating new distribution are as follows:

1. To have a distribution which is more flexible than the sub-models.
2. To provide consistently a better fit than sub-models like Weibull and Exponentiated Weibull.
3. To generate distributions with reversed-J, right-skewed, symmetric and left-skewed shaped.

The Weibull-G family is defined as follows: Let $K(t,\varphi)$ and $k(t,\varphi)$ be a cdf and pdf respectively of any random variable with the vector φ of parameter for any baseline random variable. The cdf of the Weibull - G family with two extra parameters ρ and σ is defined as:

$$W(t,\rho,\sigma,\varphi) = \int_0^{\left[\frac{K(t,\varphi)}{[1-K(t,\varphi)]}\right]} \rho\,\sigma\, t^{\sigma-1} e^{-(\rho t)^{\sigma}} = 1 - exp\left(-\rho\left\{\frac{K(t,\varphi)}{[1-K(t,\varphi)]}\right\}^{\sigma}\right). \tag{5}$$

The pdf corresponding to equation (5) can be written as;

$$w(t,\rho,\sigma,\varphi) = \rho\,\sigma\, k(t,\varphi)\, \frac{[K(t,\varphi)]^{\sigma-1}}{[1-K(t,\varphi)]^{\sigma+1}}\, exp\left(-\rho\left\{\frac{K(t,\varphi)}{[1-K(t,\varphi)]}\right\}^{\sigma}\right). \tag{6}$$

In this paper, the Weibull Exponentiated Weibull distribution (WEW) is proposed and evaluated based on equations (5) and (6).

This article aims:

1. To introduce a new distribution called the WEW distribution.
2. To explore and study several mathematical properties of the WEW model.
3. To illustrate the new model is more flexible than its sub-models by applying it to real data via goodness of fit test.

The rest of this paper is arranged as follows: In Section 2, the cumulative function, density function and hazard function of the WEW distribution are defined. The expansion of the density functions is provided in Section 3. In Section 4, some mathematical properties of the new model are studied and discussed, including the quantile function, the rth moment, the moment-generating function and order statistics. The maximum likelihood estimation is proposed to estimate the parameters and is presented in Section 5. In Section 6, a real dataset is used to illustrate the usefulness of the new model. Finally, concluding remarks are presented in Section 7.

2. THE WEIBULL EXPONENTIATED WEIBULL DISTRIBUTION.

In this section, we define the Weibull Exponentiated Weibull model of continuous distributions by using the Weibull-G family as a generator. Inserting (1) into (5) yields WEW model, which the cdf is given as: Here $\delta,\varepsilon,\vartheta,\rho,\sigma>0$ where ρ and σ are two additional shape parameter. The corresponding pdf of WEW model is given as:

$$W(t,\rho,\sigma,\delta,\varepsilon,\vartheta) = 1 - exp\left(-\rho\left\{\frac{\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}}{\left[1-\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}\right]}\right\}^{\sigma}\right) = 1 - exp\left(\frac{-\rho\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta\sigma}}{\left[1-\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}\right]^{\sigma}}\right) \tag{7}$$

$$w(t,\rho,\sigma,\delta,\varepsilon,\vartheta) = \rho\,\sigma\,\varepsilon\,\vartheta\,\delta^{\varepsilon}\, t^{\varepsilon-1}\, e^{-(\delta t)^{\varepsilon}} \frac{\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta\sigma-1}}{\{1-[1-e^{-(\delta t)^{\varepsilon}}]^{\vartheta}\}^{\sigma+1}} * exp\left(-\rho\frac{\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta\sigma}}{\left[1-\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta}\right]^{\sigma}}\right) \tag{8}$$

The hazard function of WEW distribution is:

$$h(t,\rho,\sigma,\delta,\varepsilon,\vartheta) = \rho\,\sigma\,\varepsilon\,\vartheta\,\delta^{\varepsilon}\, t^{\varepsilon-1}\, e^{-(\delta t)^{\varepsilon}} \frac{\left[1-e^{-(\delta t)^{\varepsilon}}\right]^{\vartheta\sigma-1}}{\{1-[1-e^{-(\delta t)^{\varepsilon}}]^{\vartheta}\}^{\sigma+1}} \tag{9}$$

Figures 1 and 2 display some of the possible shapes of pdf and hazard function of WEW distribution for selected values of parameters $\rho,\sigma,\delta,\varepsilon$, and ϑ. The pdf of WEW has various shapes, such as reversed-J, right-skewed, symmetric and left-skewed. Figure 2 shows that the hazard function of WEW distribution can have very flexible shapes, such as increasing, decreasing, and bathtub.

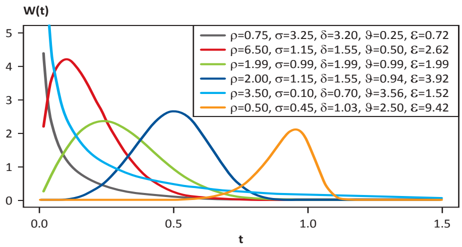


Figure 1. Plot of the WEW Density Function for Different Values of Parameters ρ,σ,δ,ε, and ϑ

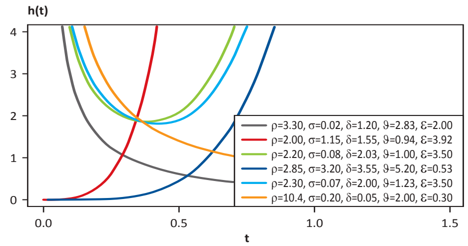


Figure 2. Plot of the WEW Hazard Function for Some Different Values of Parameters ρ,σ,δ,ε, and ϑ

3. EXPANSION FOR PDF.

We established pdf expansion of the WEW distribution to determine several statistical properties. Since the power series for the following expansion can be written as in Jeffrey, A., & Zwillinger, D. (2007) such that

$$\exp\left(-\rho\left\{\frac{K(t,\varphi)}{[1-K(t,\varphi)]}\right\}^{\sigma}\right)=\sum_{m=0}^{\infty}\frac{(-1)^m\rho^m K(t,\varphi)^{\sigma m}}{m![1-K(t,\varphi)]^{\sigma m}} \quad (10)$$

Based on (10) the pdf can be rewritten as:

$$w(t,\rho,\sigma,\delta,\varepsilon,\vartheta)=\rho\,\sigma\,\varepsilon\,\delta\,\vartheta^{\delta}\,t^{\delta-1}e^{-(\vartheta t)^{\delta}}\sum_{m=0}^{\infty}\frac{(-1)^m\rho^m\left[1-e^{-(\vartheta t)^{\delta}}\right]^{\varepsilon\sigma(m+1)-1}}{m!\left[1-\left[1-e^{-(\vartheta t)^{\delta}}\right]^{\varepsilon}\right]^{\sigma(m+1)+1}} \quad (11)$$

Similarly, since the generalized binomial theorem can be written as Jeffrey, A., & Zwillinger, D. (2007)

$$[1-K(t,\varphi)]^{-b}=\sum_{p=0}^{\infty}\frac{(-1)^p\Gamma(b+p)}{p!\,\Gamma(b)}[K(t,\varphi)]^p \quad (12)$$

Substituting (12) into (11) the expansion for the pdf of WEW model can be written as

$$w(t,\rho,\sigma,\delta,\varepsilon,\vartheta)=\rho\,\sigma\,\varepsilon\,\delta\,\vartheta^{\delta}\,t^{\delta-1}e^{-(\vartheta t)^{\delta}}\sum_{p=0}^{\infty}\sum_{m=0}^{\infty}\frac{(-1)^{m+p}\rho^m\,\Gamma(p+\sigma(m+1)+1)}{m!\,p!\,\Gamma(\sigma(m+1)+1)} \times\left[1-e^{-(\vartheta t)^{\delta}}\right]^{\varepsilon[p+\sigma(m+1)]-1} \quad (13)$$

Equation (13) can be rewritten as follows:

$$w(t,\rho,\sigma,\delta,\varepsilon,\vartheta)=\rho\,\sigma\,\varepsilon\,\delta\,\vartheta^{\delta}\,t^{\delta-1}e^{-(\vartheta t)^{\delta}}\sum_{p,m=0}^{\infty}\varepsilon[p+\sigma(m+1)]\,\Upsilon_{m,p} \times\left[1-e^{-(\vartheta t)^{\delta}}\right]^{\varepsilon[p+\sigma(m+1)]-1}, \quad (14)$$

Where

$$\Upsilon_{m,p}=\frac{(-1)^{p+m}\rho^m\,\Gamma(p+\sigma(m+1))}{m!\,p!\,\Gamma(\sigma(m+1)+1)}$$

The pdf of WEW can also be expressed as an infinite linear combination of the exponential- EW (exp-EW) density function if we use the generalized binomial theorem with different power parameter. Furthermore, many mathematical properties of the WEW distribution can be obtained directly from exp-EW, such as, moment, the moment-generating function and order statistics.

4. MATHEMATICAL PROPERTIES

In this section, some important mathematical properties of the WEW distribution are derived, specifically the quantile function, moment, moment generating function, and order statistics.

Quantile function.

The quantile function for Weibull Exponentiated Weibull distribution can be found by inverting equation (5) as follows:

$$Q(u)=\frac{1}{\varepsilon}\left[-\ln\left(1-\left\{\frac{\left[\frac{-\ln(1-u)}{\rho}\right]^{\sigma}}{1-\left[\frac{-\ln(1-u)}{\rho}\right]^{\sigma}}\right\}^{\frac{1}{\vartheta}}\right)\right]^{\frac{1}{\delta}} \tag{15}$$

Equation (15) is important for finding some essential measures, such as Bowley's skewness and Moor's kurtosis. Equation (15) is used also to generate random numbers from the WEW distribution when the parameters ρ,σ,δ,ε and ϑ are known.

Sub Models

It is worth noting that many sub-models can be obtained as special cases of the WEW distribution by fixing the parameters in equation (8) such as Weibull Exponentiated Exponential (WEE), Weibull Burr type X (WBX), Weibull Burr type X with one parameter (WBX1), Weibull Weibull (WW), Exponentiated Weibull (EW) and others.

Moment

The $\mu r'$ moment of random variable T can be obtained from pdf (14) as follows:

$$\mu_r' = \int_0^{\infty} t^r w(t,\rho,\sigma,\delta,\varepsilon,\vartheta)\,dt,$$

Using the the pdf of the WEW distribution in equation (14) with some algebra we can rewrite and find the *r* th moment as follows:

where

$$\mu_r' = \frac{\rho\sigma}{\delta^r}\sum_{m,p,l=0}^{\infty}\Upsilon_{m,p}\,\varpi_l\,\vartheta[\rho+\sigma(m+1)]\frac{\Gamma\left(\frac{r}{\varepsilon}+1\right)}{(l+1)^{\frac{r}{\varepsilon}+1}}, \tag{16}$$

Equation (16) is also important for finding many measures, such as the mean, coefficient of central moments, variance, cumulants, skewness and kurtosis.

Moment generating function

The moment generating function for WEW distribution is given by:

$$M_T(z)=E(e^{zT})=\rho\sigma\sum_{m,p,l=0}^{\infty}\Upsilon_{m,p}\,\varpi_l\frac{z^r\vartheta[\rho+\sigma(m+1)]\Gamma\left(\frac{r}{\varepsilon}+1\right)}{r!\,\delta^r\,(l+1)^{\frac{r}{\varepsilon}+1}}. \tag{17}$$

Order statistics

Let *T1*, *T2*,…, *Tn* be a random sample size *n* from the WEW distribution. The *l* th order statistics of WEW *T1*;*n* is given by

$$w_{l:n}=\sum_{q,j=0}^{\infty}\Upsilon_{q,j}\,w(t;\rho,\sigma,\delta,\varepsilon,\vartheta[\sigma(q+1)+j]), \tag{18}$$

Where

$$\Upsilon_{q,j}=\sum_{m=0}^{n-1}\sum_{p=0}^{\infty}\binom{n-1}{m}\binom{l+m-1}{p}\frac{(-1)^{m+p+q}[\rho(p+1)]^j[\sigma(q+1)+j]!}{q!\,j!\,B(l,n-l+1)[\rho(p+1)]^j} \tag{19}$$

Based on equation (18), many structural properties of *T1;n* can be obtained easily for example, the *r* th moment of *T1;n*.

5. PARAMETER ESTIMATION

The maximum likelihood estimators of the WEW model parameters Φ=(ρ,σ,δ,ε,ϑ) from complete samples are derived. Let *T*1, *T*2,…, *Tn* be a random sample of size *n* from the WEW model with vector of parameters Φ=(ρ,σ,δ,ε,ϑ). The logarithm of likelihood function for the vector of parameters is given by

$$
\begin{aligned}
l = \log L(\Theta) = {} & n\log(\rho\sigma\vartheta\varepsilon\delta^{\varepsilon}) + \sum_{i=1}^{n}(\varepsilon-1)\log(t_i) + (\vartheta\sigma-1)\sum_{i=1}^{n}\log(1-e^{-(\delta t_i)^{\varepsilon}}) \\
& -\sum_{i=1}^{n}(\delta t_i)^{\varepsilon} - (\sigma-1)\sum_{i=1}^{n}\log\left(1-(1-e^{-(\delta t_i)^{\varepsilon}})^{\vartheta}\right) \\
& -\rho\log\left(\sum_{i=0}^{n}\frac{[1-e^{-(\delta t_i)^{\varepsilon}}]^{\vartheta\sigma}}{[1-[1-e^{-(\delta t_i)^{\varepsilon}}]^{\vartheta}]^{\sigma}}\right)
\end{aligned}
\tag{20}
$$

Differentiating with respect to each parameter and setting the result equals to zero, the maximum likelihood estimators will be obtained. The partial derivatives of l with respect to each parameter are given by

$$
\frac{\partial l}{\partial \rho} = \log\sum_{i=1}^{n}(c_i) = 0 \tag{21}
$$

$$
\frac{\partial l}{\partial \sigma} = \frac{n}{\sigma} + \vartheta\sum_{i=1}^{n}\log(a_i) - \sum_{i=1}^{n}\log(b_i) - \rho\log\sum_{i=0}^{n}\frac{\partial c_i}{\partial \sigma} = 0 \tag{22}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \delta} = {} & \frac{n\varepsilon}{\delta} + (\vartheta\sigma-1)\sum_{i=1}^{n}\log\left(\frac{\partial a_i}{\partial \delta}\right) - \sum_{i=1}^{n}\varepsilon\, t_i\,(\delta t_i)^{\varepsilon-1} \\
& -(\sigma-1)\textstyle\sum_{i=1}^{n}\log\left(\frac{\partial b_i}{\partial \delta}\right) - \rho\log\left(\sum_{i=0}^{n}\frac{\partial c_i}{\partial \delta}\right) = 0
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \varepsilon} = {} & \frac{n}{\varepsilon} + \sum_{i=1}^{n}\log(t_i) + (\vartheta\sigma-1)\sum_{i=1}^{n}\log\left(\frac{\partial a_i}{\partial \varepsilon}\right) - \sum_{i=1}^{n}(\delta t_i)^{\varepsilon}\log(\delta t_i) \\
& -(\sigma-1)\sum_{i=1}^{n}\log\left(\frac{\partial b_i}{\partial \varepsilon}\right) - \rho\log\left(\sum_{i=0}^{n}\frac{\partial c_i}{\partial \varepsilon}\right) = 0
\end{aligned}
\tag{24}
$$

$$
\begin{aligned}
& \frac{\partial l}{\partial \vartheta} = \frac{n}{\vartheta} + \sigma\sum_{i=1}^{n}\log\left(\frac{\partial a_i}{\partial \vartheta}\right) - (\sigma-1)\sum_{i=1}^{n}\log\left(\frac{\partial b_i}{\partial \vartheta}\right) - \rho\log\left(\sum_{i=0}^{n}\frac{\partial c_i}{\partial \vartheta}\right) = 0 \\
& a_i = (1-e^{-(\delta t_i)^{\varepsilon}}), \quad b_i = [1-(a_i)^{\vartheta}], \quad c_i = \frac{a_i^{\vartheta\sigma}}{(1-b_i)^{\sigma}}
\end{aligned}
\tag{25}
$$

These equations can solved numerically (example we used the Package in R (AdequecyModel)).

6. APPLICATION.

We consider a real data set to illustrate the flexibility of the new model. We compare the fit of the WEW distribution with its sub-models such as Exponentiated Weibull (EW), Weibull (W), Exponentiated Exponential (EE) and Exponential (E) distributions.

Aarset data set.

We consider the dataset from Aarset, M.V. (1987) on lifetime of 50 components, which have a bathtub shape failure rate property. This data set is also reported in many papers such as in Mudholkar, G. S., & Srivastava, D. K. (1993) ; Mudholkar et al. (1996) and others. We estimate the unknown parameters of the distribution by maximum likelihood method as described in section 5. We use some measures of goodness of fit, including the Akaike information criterion (AIC), corrected Akaike information criterion (AICC), and Bayesian information criterion (BIC) statistics. Smaller values of these statistics indicate a better fit. The results are shown in Table 1.

Table 1. The Value of MLEs, log L (l), AIC, AICC, and BIC for Aarset Data Set

Models	MLEs of parameters	l	AIC	AICC	BIC
WEW	$\hat\rho = 0.5127$ $\hat\sigma = 0.0531$ $\hat\delta = 0.0295$ $\hat\varepsilon = 3.7293$ $\hat\vartheta = 2.7866$	216	441.45	442.81	447.10
EW	$\hat\delta = 0.1115$ $\hat\varepsilon = 4.1326$ $\hat\vartheta = 0.1678$	230	465.89	466.42	471.63
W	$\hat\delta = 0.0223$ $\hat\varepsilon = 0.949$	241	486.09	486.25	489.82
EE	$\hat\delta = 0.0187$ $\hat\vartheta = 0.7798$	240	483.99	484.24	487.81
E	$\hat\delta = 0.0219$	241	484.18	484.26	486.09

Based on the values of the criteria *l*, AIC, AICC, and BIC, we found that WEW distribution is the best fitted model than the sub models EW,W, EE, and E distributions for the Aarset data set. Histogram (Figures 3) shows that the data set is positively skewed. The estimated pdfs and cdfs for the fitted data are displayed in Figures 3.

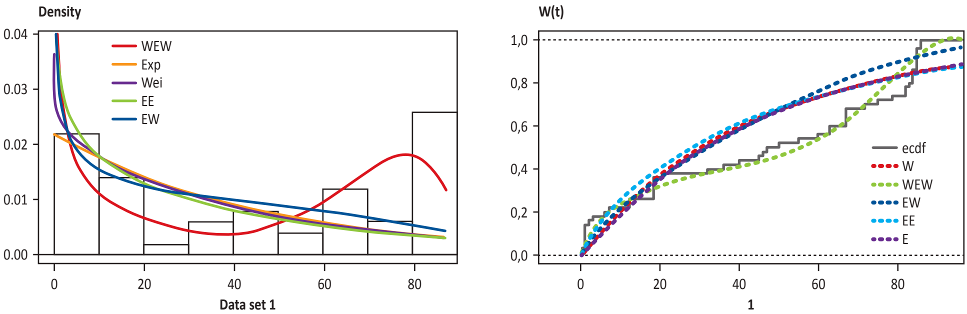


Figure 3. Plots of the Histogram, Estimated Pdfs and Cdfs for Aarset Data Set

7. CONCLUSION

We have introduced and developed a new generalization of Exponentiated Weibull distribution, called the Weibull Exponentiated Weibull (WEW) distribution. The WEW model generalized many baseline models that includes several new distributions. Properties of the WEW distribution are providing include an expansion for the pdf and expressions for the quantile function, the r th moment, moment generating function, and order statistics. The maximum likelihood method is employed to estimate the model parameters. A real data set is used to illustrate the flexibility of the distribution. This model gives a better fit than its nested models. The finding in this paper is useful for the practitioners in various fields of probability, statistics, and applied Sciences.

REFERENCES

Aarset, M. V. (1987). How to identify a bathtub hazard rate. IEEE Transactions on Reliability.

Alzaatreh, A., Lee, C., & Famoye, F. (2013b). A new method for generating families of continuous distributions. Metron.

Bourguignon M., Silva, R. B., & Cordeiro, G. M. (2014). The Weibull–g family of probability distributions. Journal of Data Science.

Burr, I. W. (1942). Cumulative frequency functions. The Annals of Mathematical Statistics.

Eugene, N., Lee, C., & Famoye, F. (2002). Beta-normal distribution and its applications. Communications in Statistics-Theory and methods.

Gupta, R. D., & Kundu, D. (1999). Theory & methods: Generalized exponential distributions. Australian & New Zealand Journal of Statistics.

Jeffrey, A., & Zwillinger, D. (2007). Table of integrals, series, and products. Academic Press.

Mudholkar, G. S., & Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Transactions on Reliability.

Mudholkar, G. S., Srivastava, D. K., & Kollia, G. D. (1996). A generalization of the Weibull distribution with application to the analysis of survival data. Journal of the American Statistical Association.

Surles, J.G., & Padgett, W.J. (2001). Inference for reliability and stress-strength for a scaled burr type x distribution. Lifetime Data Analysis

The Estimation of Parameters Spatial Autoregressive Geographically Weighted Regression (SAR-GWR) By Means Instrumental Variable (IV) Approach

I Gede Nyoman Mindra Jaya*
Universitas Padjadjaran, Bandung, Indonesia – mindra@unpad.ac.id
Faculty of Spatial Sciences, the University of Groningen, the Netherlands

Henk Folmer
Faculty of Spatial Sciences, the University of Groningen, the Netherlands

Budi Nurani Ruchjana
Universitas Padjadjaran, Bandung, Indonesia

Yudhie Andriyana
Universitas Padjadjaran, Bandung, Indonesia

Abstract

The applications of standard regression analysis on spatial data are not appropriate because of the characteristics of the spatial data. Spatial data has two characteristics i.e.: spatial dependence and spatial heterogeneity. The estimation parameter of standard regression leads to biased, inconsistency, and inefficient. Several models were developed to accommodate the characteristics of the spatial data. However, the model generally developed to solve only one problem of data spatial. Three kinds of spatial models for spatial dependence i.e.: spatial autoregressive (SAR), spatial lagged exogenous variables (SLX), and spatial error model (SEM). To accommodate the spatial heterogeneity, geographically weighted regression (GWR) or varying coefficient model (VCM) as the solution. Our research proposed a new model to accommodate two problems of the spatial data. The model is developed based on the combination SAR model and GWR model. We call the model as Spatial Autoregressive GWR (SAR-GWR). We used Instrumental Variables (IV) approach to estimate the parameters model. We have done the simulation study by mean Monte Carlo simulation to test the biased and efficiency the parameter estimates. Our new method provides a better estimate with small bias and RMSE. We also found that our method relative robust to the multicollenierity problem. The application in modeling prevalence rate of TB found that the healthy hose gives serious effect to the increasing of prevalence rate of TB in Bandung.

Keyword: GWR, SAR, Spatial dependence, Spatial heterogeneity, SAR-GWR

1. INTRODUCTION

The regression analysis can be used to model relationship between one dependent variable and several of the independent variables (Sen & Srivastava, 1990). However, for the spatial data, the regression analysis suffers from the spatial dependence and spatial heterogeneity (Brunsdon et al., 1988, Fotheringham et al., 2002), Mindra Jaya, 2016). The use of standard regression for data with spatial dependence and heterogeneity spatial will produce bias standard error and inefficiency of the parameter estimates (Anselin, 1988, Fotheringham et al., 2002).

Several models were developed to accommodate the characteristics of the spatial data. However, the model generally developed to solve only one problem of data spatial. Spatial dependence is accommodated using spatial econometrics approach. Three kinds of spatial models for spatial dependence i.e.: spatial autoregressive (SAR), spatial lagged exogenous variables (SLX), and spatial error model (SEM) (Vega & Elhorst, 2015). However, SAR model is more often applied because of the ease of computing and provide information spillover effect (LeSage & Pace, 2009). Spatial heterogeneity can be seen in two points view, i.e.: spatial heteroscedasticity (non-constant error variance) and spatial instability structure relationship (varying coefficients) (Fotheringham et al., 2002). For the former condition, the standard regression model with robust standard error can be used to be an alternative solution (Hayes & Cai, 2007). The last condition needs a different solution. To accommodate the spatial varying coefficients, the Geographical Weighted Regression (GWR) is commonly used (Fotheringham et al., 2002).

The literature suggests that both features could be considered in the analysis simultaneously (Brunsdon et al., 1998). Despite the possibility of addressing spatial heterogeneity and spatial dependence at the same time, only few studies have been focused on developing parameter estimation for this problem.

In this research, we develop a new approximation to modeling the spatial dependence and spatial heterogeneity simultaneously using Instrumental Variable (IV) approach with Two Stage Weighted Least Square (TSWLS) estimation. The model is developed based on the combination SAR model and GWR model. We call the model as Spatial Autoregressive GWR (SAR-GWR). Instead the analytical approach, we use simulation study Monte Carlo to test the biased and efficiency the parameter estimates of our model. We also check the robustness of our method from multicollinearity problem.

We use SAR-GWR for modeling prevalence rate of Tuberculosis (TB) in Bandung. Bandung is a capital city of west Java. In 2013, Bandung recorded as many as 1507 TB cases, increased significantly to 1872 cases in 2014 and back down in 2015 to 1584 cases (Bandung Health Office, 2013-2015). Various factors that have effect on the prevalence rate of tuberculosis are Healthy Hoses and Cleaning Water (Manalu, 2010).

This paper is organized into four sections. The first part discusses the introduction, the second section discusses the method, the third section discusses the design of the simulation and the fourth section about the conclusions and further research.

2. METHODS

Basic Concept of GWR

In this section we are going to discuss of the basic concept of the GWR. Let's we define the GWR model with K exogenous variables X

$$y_i = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} x_{ki} + \varepsilon_i \; ; i = 1,2,\dots,N \tag{1}$$

where y_i is the dependent variable at location i, x_{ki} is the value of kth independent variable at location i, the β_{ki} is the local regression coefficient of k^{th} independent variable, the β_{0i} is the intercept parameter at location i, and ε_i is the random error at location , which may follow an independent normal with zero mean and homogenous variance (Wheeler & Tiefelsdorf, 2005). The contrast with standard regression model, the GWR allows regression coefficients are varying from location to location. Note that the model (1) has $n(K+1)$ parameters that must be estimated. This number larger than the available degrees of freedom based on the given number of observations. The GWR must solve the lack of degrees of freedom by develop parameters estimation that represent variation of the parameters over space (Leung, Mei, & Zhang, 2000). The parameters in equation (1) are estimated by a weighted least squares (WLS) procedure, making the weighting system dependent on the location in geographical space and, therefore, allowing local rather than global parameters to be estimated.

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^{T}\boldsymbol{\Psi}(i)\mathbf{X}]^{-1}\mathbf{X}^{T}\boldsymbol{\Psi}(i)\mathbf{y} \tag{2}$$

where $\psi(i)=(\omega_1(i), \omega_2(i),\dots, \omega_n(i))$ is the diagonal weights matrix that varies for any prediction location i, X is the matrix exogenous variable with a first column of 1's for intercept, y is the vector dependent variables, and $\hat{\beta}(i) = \left((\hat{\beta}_0(i)\ \hat{\beta}_1(i),\dots,\hat{\beta}_k(i)\right)$ is the vector of $K + 1$ local regression coefficients at location i, where $\psi(i)$ is a matrix of weights specific to location such that observations nearer to are given greater weight nearer to are given greater weight than observations further away.

To estimate the parameters model GWR, first, we have to define weights matrix $\psi(i)$. The weights matrix is specified as a local kernel function that models a distance decay effect from the calibration locations to the prediction location. One of the most commonly used kernel functions, and the one used in this analysis, is the bi-square nearest neighbor function:

$$w(i) = \begin{cases} \left[1 - (d_{ij}/b)^2\right]^2 & if\ j \in \{N_i\} \\ 0 & if\ j \notin \{N_i\} \end{cases} \tag{3}$$

where d_{ij} is the distance between the calibration location j and the prediction location i, b is the threshold distance to the N^{th} nearest neighbor, and the set $\{N_i\}$ contains the observations that are within the distance range of the threshold N^{th} nearest neighbor (see Fotheringham et al. 2002, p 58). The next Kernel function that usually use too is Gaussian function:

$$w(i) = \begin{cases} \exp(-0.5(d_{ij}/b)^2) & if\ j \in \{N_i\} \\ 0 & if\ j \notin \{N_i\} \end{cases} \tag{4}$$

Instrumental Variable Approaches

WLS estimation leads to the bias and inconsistent estimate in case spatial autoregressive structure. In contrast, IV method allows an unbiased and consistent estimation of the parameter spatial Autoregressive model (Anselin, 1988). This concept is accommodating in this research to estimate parameter SAR-GWR. The SAR-GWR model can be modeled as:

$$y_i = \mathbf{X}_i\boldsymbol{\beta}_i + \rho_i\mathbf{W}_i\mathbf{y} + \varepsilon_i \tag{5}$$

where y_i is a dependent variable at i th location, ρ_i is a $(N \times 1)$ vector autoregressive parameter, W_i is a $(N \times 1)$ vector of row i th of spatial weight matrix W, y is a $(N \times 1)$ vector of dependent variable, $X_i = (1, X_{1i}, X_{2i}, X_{1i}, \ldots, X_{ki})$ is a $(1 \times (K + 1))$ vector of independent variable, $\beta_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_k)$ is a $((K + 1) \times 1)$ parameter regression at i th location, and ε_i is a disturbance at i th location which assume independently and identically distribution. Model (5) can be written as a matrix notation using Hadamard Product[1]

$$\mathbf{y} = (\mathbf{X} \circ \boldsymbol{\beta})\boldsymbol{\iota}_p + \rho_i \circ \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{6}$$

where "$\circ$" is the Hadamard Product.

To facilitate estimation, we arrange the X-values and the values of the endogenous lagged variable $W_i y$ in an $(1 \times (p + 1))$ vector Z.

$$\mathbf{Z} = [\mathbf{X} \quad \mathbf{W}\mathbf{y}] \tag{7}$$

and the respective regression coefficients in a $(p + 1) \times 1$ parameter vector θ:

$$\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \rho] \tag{8}$$

with the definitions (7) and (8) the extended spatial lag model (6) reads

$$\begin{aligned} \mathbf{y} &= ([\mathbf{X} \quad \mathbf{W}\mathbf{y}] \circ [\boldsymbol{\beta} \quad \rho])\boldsymbol{\iota}_{p+1} + \boldsymbol{\varepsilon} \\ \mathbf{y} &= (\mathbf{Z} \circ \boldsymbol{\theta})\boldsymbol{\iota}_{p+1} + \boldsymbol{\varepsilon} \end{aligned} \tag{9}$$

An IV estimator is based on the assumption that a set of r instruments, $r \geq p + 1$, where $p = K + 1$ arranged in an instrument matrix Q of size $n \times r$, is asymptotically uncorrelated with the disturbances ε

$$plim\ \frac{1}{n}\mathbf{Q}'\boldsymbol{\varepsilon} = \mathbf{0} \tag{10}$$

but (preferably strongly) correlated with the original variables stored in matrix Z,

$$plim\ \frac{1}{n}\mathbf{Q}'\mathbf{Z} = \mathbf{M}_{QZ} \tag{11}$$

where M_{QZ} is a finite nonsingular moment matrix Z.

A number of instruments $r = p + 1$ is equal to the number of exploratory variables including vector

one and Wy . Suppose at first the numbers of instruments are equal to the number of original exploratory variables $(X_0, X_1, X_2, \ldots, X_K, Wy)$. Then both matrices Z and Q are of size $\times (p + 1)$. As the X-variables are a priori fixed, it is natural to use them as their own instruments, because they fulfill the requirements (10) and (11). Thus only one additional instrument variable is needed which does not belong to the original set of regressors. Such an additional instrument variable must be –at least in large samples –uncorrelated with the error term ε, but the same time should be strongly correlated with Wy. It can, however, not be viewed as a single instrument for the spatially lagged endogenous variable, because information carried by the X-variables will also be used for approximating Wy. While the X-variables are perfect instruments for themselves, the spatially lagged endogenous will be instrumented by all instrument variables in Q.

In order to derive an IV estimator for for the case that the number of instruments are equal to the number of exploratory variables, we premultiply (9) by (1/N)Q:

$$\frac{1}{N}\mathbf{Q}'\mathbf{y} = \frac{1}{N}\mathbf{Q}'(\mathbf{Z} \circ \boldsymbol{\theta})\iota_{p+1} + \frac{1}{N}\mathbf{Q}'\varepsilon \tag{12}$$

For large N ($N \to \infty$) the last term goes in probability to zero. In this case equation (13) reduces to

$$\frac{1}{N}\mathbf{Q}'\mathbf{y} = \frac{1}{N}\mathbf{Q}'(\mathbf{Z} \circ \boldsymbol{\theta})\iota_{p+1}\varepsilon \tag{13}$$

Since Q and Z are both matrices of size $n \times (p + 1)$, the matrix product Q'Z gives a matrix of size $(p + 1) \times (p + 1)$. Because the parameters model is different for each location, we have to estimate the parameters for each location using WLS approach as in (2):

$$\hat{\boldsymbol{\theta}}_i^{IV} = (\mathbf{Q}'\boldsymbol{\Psi}_i\mathbf{Z})^{-1}\mathbf{Q}'\boldsymbol{\Psi}_i\mathbf{y} \tag{14}$$

As (14) converges in probability to θ_i

Proof:
$$\hat{\boldsymbol{\theta}}_i^{IV} = (\mathbf{Q}'\boldsymbol{\Psi}_i\mathbf{Z})^{-1}\mathbf{Q}'\boldsymbol{\Psi}_i((\mathbf{Z} \circ \boldsymbol{\theta}) + \varepsilon)$$
$$\hat{\boldsymbol{\theta}}_i^{IV} = \boldsymbol{\theta}_i + (\mathbf{Q}'\boldsymbol{\Psi}_i\mathbf{Z})^{-1}\mathbf{Q}'\boldsymbol{\Psi}_i\varepsilon$$
$$\operatorname{plim} \hat{\boldsymbol{\theta}}_i^{IV} = \boldsymbol{\theta}_i$$

It has been proved that IV estimator is consistent. However the (14) requires that the number of instruments equals the number of regressors. For over-identified models, the IV estimator can be used, by discarding some of the instruments so that the model is just-identified. But there can be an asymptotic efficiency loss in discarding these instruments. Instead, a common procedure is to use the two-stage least squares (2SLS), estimator.

The idea IV approach to SAR-GWR is creating an instrument variable $\hat{\mathbf{Z}}$ with dimension $(n \times (p+1))$ and the $(p+1)$th column contains the ultimate instrument $\widehat{\mathbf{Wy}}$ for $\mathbf{Wy}$. The first p column of are identical with the first p columns of $\mathbf{Q}$, since the exogenous variables $(\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_K)$ are instruments for themselves with $\mathbf{X}_0 = (1,1,\dots,1)'$. The $(p+1)$th column of contains the ultimate instrument for $\mathbf{Wy}$ which may be constructed by a linear combination of all instruments $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_K, \mathbf{W}\boldsymbol{X}_1, \dots, \mathbf{W}\boldsymbol{X}_K$ in $\mathbf{Q}$. The first stage is obtaining $\hat{\mathbf{Z}} = [\mathbf{X} \quad \widehat{\mathbf{Wy}}]$ by regress the $\boldsymbol{Z}$ on $\mathbf{Q}$ where:

$$\hat{\mathbf{Z}} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Z} \tag{15}$$

and the second stage is obtaining the parameter estimate SAR-GWR

$$\hat{\boldsymbol{\theta}}_i^{IV} = (\hat{\mathbf{Z}}'\boldsymbol{\Psi}_i\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\boldsymbol{\Psi}_i\mathbf{y} \tag{16}$$

Estimation Steps

1. Define matrix $\mathbf{Z} = [\mathbf{X} \quad \mathbf{Wy}]$
2. Define matrix $\mathbf{Q} = [\mathbf{X} \quad \mathbf{WX}]$
3. Obtain instrumental variable $\hat{\mathbf{Z}}$ by regress the $\mathbf{Z}$ on Q where $\hat{\mathbf{Z}} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Z}$
4. Estimate the parameter SAR-GWR using $\hat{\boldsymbol{\theta}}_i^{IV} = (\hat{\mathbf{Z}}'\boldsymbol{\Psi}_i\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\boldsymbol{\Psi}_i\mathbf{y}$
5. Estimate the $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_i^{IV}) = \hat{\sigma}^2(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})'$ where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{n-p}$

3. SIMULATION STUDY

We use Monte Carlo simulation to show that the new method we proposed give good results. We used a map of Bandung as a spatial reference in this simulation study.

Bandung is composed of 30 districts so that we have a 30 point spatial point. Weight matrix for the model based on the autoregressive spatial contiguity Queen. While the weight matrix for modeling based on the weighted Gaussian GWR.

Scheme of simulation

1. Generate variable bivariate normal X_1 and X_2 with correlation $(X_1, X_2) = R$

$$X_1 = \mu_1 + \sigma_1 Z_1$$
$$X_2 = \mu_2 + \sigma_2\left[RZ_1 + Z_2\sqrt{1-R^2}\right]$$

with Z_1 and $Z_2 \sim N(0,1)$, $R = \{0, 0.3, 0.5, 0.7\}$, $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 2$

2. Generate variable dependent y

$$y = (I - \rho W)^{-1}(\beta_0 + \beta_1 \circ X_1 + \beta_1 \circ X_1 + \beta_2 \circ X_2) + (I - \rho W)^{-1}\varepsilon$$

with $\beta_0 = 1 + 0.5 \times Lat + 0.5 \times Long$

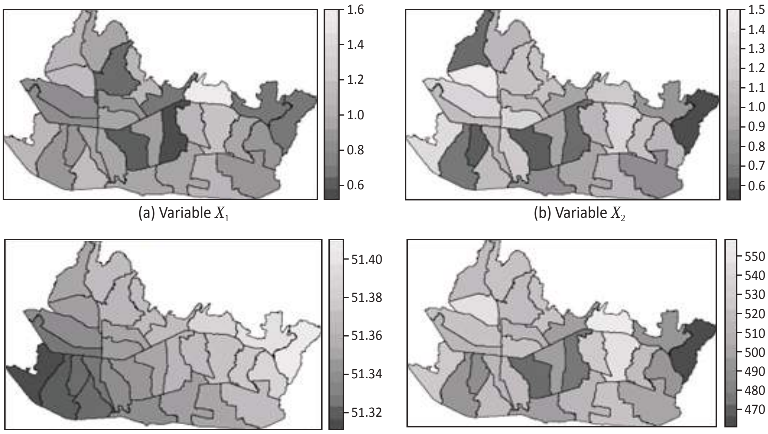
$\beta_1 = 1 + 0.5 \times Lat + 0.5 \times Long$

$\beta_2 = 1 + 0.5 \times Lat + 0.5 \times Long$

$\varepsilon \sim N(0,1)$

$\rho = \{0.1, 0.3, 0.5, 0.7\}$

W is a $n \times n$ standardized weight matrix



(a) Variable X_1 (b) Variable X_2

(c) Regression Parameter $\beta_0=\beta_1=\beta_0$ (d) Variable y

Figure 1. Data Generation Examples ($\mu_1=\mu_2=1$, $\sigma_1=\sigma_2=0.2$, $\sigma_2=1$, $\rho=0.7$, $R=0.3$)

4. RESULTS

Case 1: Simulation Study

Simulation Monte Carlo has been done with 1000 time iteration process. We use Boxplots to visualize the Bias and Root Mean Square Error (RMSE) of the parameters estimate based on standard GRW and SAR-GRW using IV approach.

The Boxplots does not include the parameter intercept, β_0, because the Bias and RMSE for standard GWR which is not incorporating spatial autoregressive is too big compare than β_1 and β_2. This is because the part of spatial autoregressive added to the intercept. Let's refers to the (),

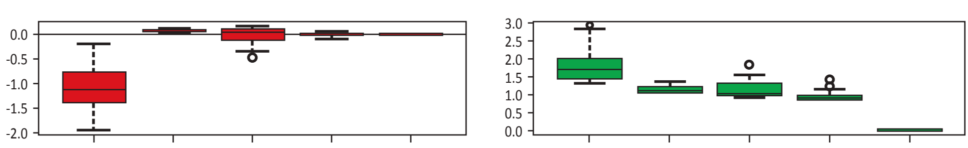
$$\mathbf{y} = (\mathbf{X} \circ \boldsymbol{\beta})\boldsymbol{\iota}_p + \rho\mathbf{W} \quad + \boldsymbol{\varepsilon} \tag{17}$$

and we can re-write the () become:

$$\mathbf{y} = (\tilde{\mathbf{X}} \circ \tilde{\boldsymbol{\beta}})\boldsymbol{\iota}_p + (\boldsymbol{\beta}_0 + \rho\mathbf{W}\mathbf{y}) + \boldsymbol{\varepsilon} \tag{18}$$

$$\mathbf{y} = (\tilde{\mathbf{X}} \circ \tilde{\boldsymbol{\beta}})\boldsymbol{\iota}_p + \tilde{\boldsymbol{\beta}}_0 + \boldsymbol{\varepsilon} \tag{19}$$

Where $\tilde{\mathbf{X}}$ is a (n x K) dependent variables without vectors one's, $\tilde{\boldsymbol{\beta}}$ is a (n x K) regression parameters without intercept. For standard GWR that ignore $\rho\mathbf{W}\mathbf{y}$ will have intercept $\tilde{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0 + \rho\mathbf{W}\mathbf{y}$.



(a) BIAS: R = 0, ρ = 0.1 (b) RMSE: R = 0, ρ = 0.1

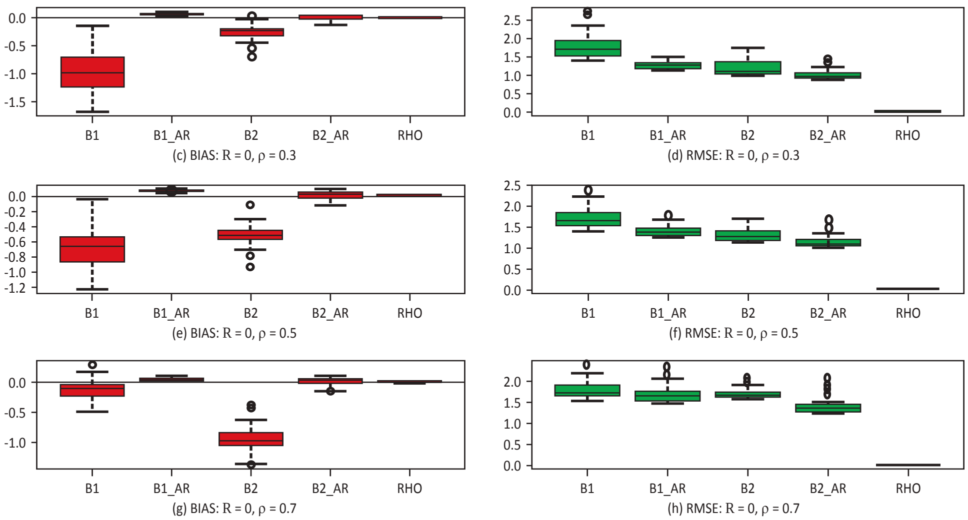


Figure 2. Bias and RMSE of (β_1 and β_2) for R=0 and ρ={0.1, 0.3, 0.5, 0.7} Model GWR and SAR-GWR

Figure 2 show that the Bias and RMSE of regression parameters (β_1 and β_2) for zero multicollinearity. The Bias and RMSE of SAR-GWR regression parameters closed to zero, while the regression parameter estimates (β_1 and β_2) for standard GWR have higher Bias and RMSE. It means ignoring the spatial autoregressive caused Bias estimate and the RMSE increase aligned with the increasing spatial autoregressive.

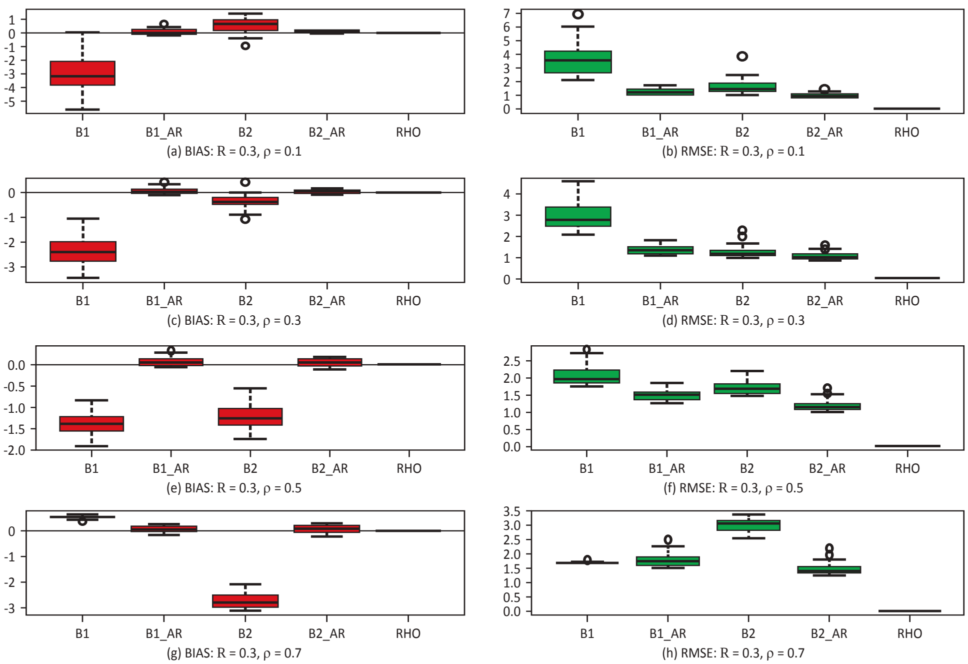
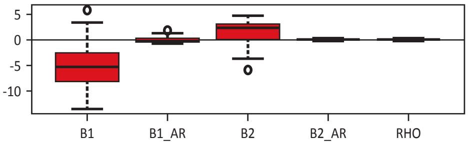
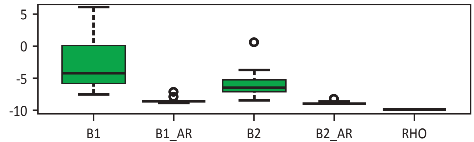


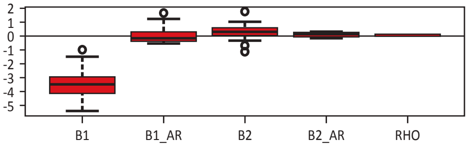
Figure 3. Bias and RMSE of (β_1 and β_2) for R=0 and R= 0.3, and ρ={0.1, 0.3, 0.5, 0.7} Model GWR and SAR-GWR



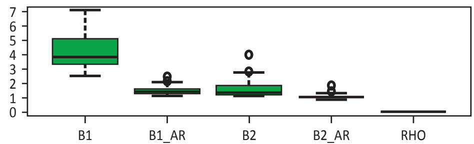
(a) BIAS: R = 0.5, ρ = 0.1



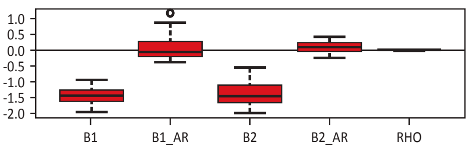
(b) RMSE: R = 0.5, ρ = 0.1



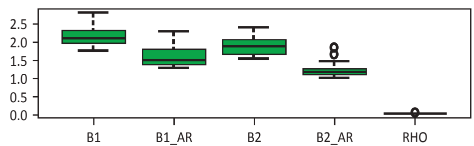
(c) BIAS: R = 0.5, ρ = 0.3



(d) RMSE: R = 0.5, ρ = 0.3



(e) BIAS: R = 0.5, ρ = 0.5



(f) RMSE: R = 0.5, ρ = 0.5

Utilization of a Known Coefficient of Variation in the Linear Combination of Normal Variance Interval Estimation Procedure

Sirima Suwan*
Statistics Department, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand - sirimasuwan@hotmail.com

Abstract

In this article we study the performance of a variety of linear combinations of independent normal population variances interval estimators that utilize the known coefficient of variation by using the well- known method of variance estimates recovery and its application to a general linear function of parameters. We do not only consider confidence intervals derived from the unbiased normal mean, the biased normal mean and the efficiency biased normal mean, respectively but also evaluate the performance of all confidence limits by comparing their empirical coverage probabilities and empirical confidence average widths.

Keywords: Minimum mean squared error, Confidence interval approximation, The MOVER.

1. INTRODUCTION

The coefficient of variation (*CV*) is the ratio of the standard deviation to the mean (average) which is defined as σ/μ. It describes the dispersion of the variable in a way that does not depend on the variable's measurement unit. The CV is particularly useful and found to have practical importance for use in many areas such as in clinical and diagnostics, chemical experiments, finance, biological and industrial applications, etc. However, as the coefficient of variation is known one should be use it with care since its unavailable information was often assumed.

In the existence of an earlier knowledge of the coefficient of variation, many methods are obtainable for finding efficient normal variance estimators. Searls and Intarapanich (1990) create the MMSE (Minimum Mean Square Error Estimator) of the normal population variance. Another MMSE of the normal population variance which has smaller bias and mean squared error than Searls and Intarapanich's estimator, when its known coefficient of variation is lesser than 0.6 has been constructed by Winston et al.(2010) The attractiveness of each kind of estimators is line in the fact that even though they are biased, each has smaller MSE than the usual unbiased estimator S^2. Consequently, the construction of confidence intervals for the variance linear combination generated from each kind of the variance that mentioned above in which the coefficient of variation is known then have been proposed by using the well known unconditional MOVER methods of combining separate confidence bounds for more than one variance to form a confidence interval for their combination.

The first prediction interval is derived by making use of the normal mean confidence interval in its procedure. The rest two predictors are derived from the minimum mean-square error estimator (MMSE) of the normal population variance, given in Searls and Intarapanich (1990) and the optimal estimator of the normal population variance, given in Winston et al. (2010), respectively. Additionally, all confidence intervals constructed are defined in Section 2. Section 3 reports all results and evaluates the performance of those three types of confidence interval estimators with the exact normal variance confidence limit by means of an empirical simulation study with respect to the coverage probabilities and the average interval widths. A brief conclusion is presented in the last section of this paper.

2. CONFIDENCE INTERVAL FOR THE LINEAR COMBINATION OF VARIANCE INTERVAL ESTIMATORS

Let $x_1,\ldots, x_n$ represent a random sample of size $n \geq 2$ from a normal population with mean μ ($\mu > 0$) and variance σ^2. The population coefficient of variation CV = σ/μ, $\bar{x} = \sum x_i/n$ be the sample mean and $S^2 = \sum (x_i - \bar{x})^2/(n-1)$ be the sample variance, and then the confidence interval for the linear combination of variance interval estimators are derived.

2.1. The normal variance interval approximation

It is well known that an exact 100(1-α) % confidence interval for σ^2 is

$$CI_{\sigma^2} = [(n-1)S^2/U_1, (n-1)S^2/L_1] \quad (1)$$

where $U_1 = \chi^2_{\alpha/2;n-1}$, $L_1 = \chi^2_{1-\alpha/2;n-1}$ and $\chi^2_{p,df}$ is the point on a central chi-squares distribution with degrees of freedom exceeded with probability p (Bonett,2006).

In the section that follows, we present three methods for making inference about σ^2 using the known *CV*.

2.2. The normal variance confidence interval that utilize the known CV

It is also well known that the usual confidence interval 100(1-α) % for the mean μ (with unknown variance σ^2) is

$$CI_\mu = \left[\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] \quad (2)$$

where $t_{1-\alpha/2,n-1}$ is the 100(1-α/2) percentile of the *t* distribution with *n* - 1 degrees of freedom.

Since the population coefficient of variation (*CV*) of a distribution with mean μ (μ > 0) and variance σ^2 is defined as $CV = \sigma/\mu$ then $CV = \mu = \sigma\ CV^2 = \alpha$ and $CV^2\mu^2 = \sigma^2$, respectively. Hence, we get the 100(1-α) % confidence interval for σ^2 with known *CV* as $CI_A = [L_A, U_A]$, where

$$CI_A = \begin{cases} L_A = a(\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})^2 \\ U_A = a(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})^2 \end{cases} ;if \left|(\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right| < \left|(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right|,$$

or

$$CI_A = \begin{cases} L_A = 0 \\ U_A = a(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})^2 \end{cases} ;if \left|(\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right| = \left|(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right|,$$

$$or \quad CI_A = \begin{cases} L_A = a(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})^2, \\ U_A = a(\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})^2, \end{cases} ;if \left|(\bar{x} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right| > \left|(\bar{x} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}})\right|. \quad (3)$$

In the remainder of this article we try to construct the variance interval estimations generated from the biased estimators of the normal variance with a known *CV*.

2.3. The normal variance interval approximation based on the Seals mean with known *CV*

From a result of Searls (1964) it follows that in case of normal distribution, the Mean Squared Error (MSE) estimator of mean is of the form $\bar{x}_w = nw\bar{x}$, w being an optimal constant value for which the mean squared error (MSE) of $\bar{x}_w$ is least, then $\bar{x}_w = \frac{n\bar{x}}{n+a^2}$ with $MSE(\bar{x}_w) = \frac{\sigma^2}{n+a^2}$. In which $\alpha = CV^2$ is assumed known. Since an approximate determination of the confidence interval is based on the approximation of distribution function of S^2 to the normal distribution. The following 100(1-α) % confidence interval for σ^2 with known *CV* is then given by

$$CI_w = \begin{cases} L_w = a(\bar{x}_w - t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})^2 \\ U_w = a(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})^2 \end{cases} ;if \left|(\bar{x}_w - t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right| < \left|(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right|,$$

$$CI_w = \begin{cases} L_w = 0 \\ U_w = a(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})^2 \end{cases} ;if \left|(\bar{x}_w - t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right| = \left|(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right|,$$

$$or \quad CI_w = \begin{cases} L_w = a(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})^2, \\ U_w = a(\bar{x}_w - t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})^2, \end{cases} ;if \left|(\bar{x}_w - t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right| > \left|(\bar{x}_w + t_{1-\alpha/2,n-1}\frac{S\sqrt{n}}{n+a})\right|. \quad (4)$$

where $Z_{\alpha/2}$ is two-sided critical z-value.

2.4. Confidence interval approximation of the efficient estimator of normal variance using the known *CV*

Winston A.R. et.al.(2010) considered a class of estimator's population variance σ^2 of normal distribution of the form $S_K^2 = K\bar{x}^2$, K is a parameter such that the MSE of S_K is at a minimum value, in which the coefficient of variation is known, as:

$$S_K^2 = \frac{a(1+a/n)(\bar{x})^2}{[3(1+a/n)^2-2]}$$

where $\alpha = CV^2$. The idea behind this estimator was that, the $S_K{}^2$ improves the unbiased estimator of the variance S_U where $S_U^2 = \frac{a\bar{x}^2}{(1+a/n)^2}$. And hence we cleanly verify that,

$$MSE(S_K^2) = \frac{2(a^3/n)(2+a/n)\mu^4}{[3(1+a/n)^2-2]}$$

where $\alpha = CV^2$. The lower and upper confidence limits L_K and U_K for σ^2 with a known CV are:

$$CI_K = \begin{cases} L_K = \dfrac{a\bar{x}^2(1+a/n)}{(1+a/n)^2 + az_{\alpha/2}\sqrt{2(a^3/n)(2+a/n)[3(1+a/n)^2-2)]}} \\ U_K = \dfrac{a\bar{x}^2(1+a/n)}{(1+a/n)^2 - az_{\alpha/2}\sqrt{2(a^3/n)(2+a/n)[3(1+a/n)^2-2)]}} \end{cases}, \quad (5)$$

where $\alpha = CV^2$ and $Z_{\alpha/2}$ is two-sided critical z-value.

2.5. Approximate Intervals for a linear function of variances

2.5.1. The MOVER method and its applications

The MOVER method that will be investigated in this section was used to find the effective confidence intervals for a linear function of independent quantities.

This procedure advantage requirement is only the availability of separate confidence limits that have coverage levels close to nominal, and does not require that the distributions of $\hat{\theta}_i$ (i = 1,2,…,k) follow specific forms or to be symmetric. When the sampling distribution for $\hat{\theta}_i$ (i = 1,2,…,k) are symmetric, it directly shows that the method leads to the conventional confidence intervals (Zou et al., 2009). Zou et al. (2009) have extended the argument of Donner and Zou (2010) to a linear function of parameters by regarding $\theta_1 + \theta_2$ and $\theta_1 - \theta_2$ as $c_1\theta_1 + c_2\theta_2$, where c_1 and c_2 are constants, hence, the interval can be written as

$$L_{c_1\theta_1+c_2\theta_2} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2 - \sqrt{\left[c_1\hat{\theta}_1 - \min(c_1 l_1, c_1 u_1)\right]^2 + \left[c_2\hat{\theta}_2 - \min(c_2 l_2, c_2 u_2)\right]^2},$$

$$U_{c_1\theta_1+c_2\theta_2} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2 + \sqrt{\left[c_1\hat{\theta}_1 - \max(c_1 l_1, c_1 u_1)\right]^2 + \left[c_2\hat{\theta}_2 - \max(c_2 l_2, c_2 u_2)\right]^2}.$$

Their further extension is to use a mathematical induction application in order to derived a generally 100(1-α) % confidence interval for linear functions of parameters $\sum_{i=1}^{k} c_i\theta_i$, i=1, 2, … , k ,where k ≥ 2, θ_i denote any interested parameters and c_1 are coefficient in the linear function, as defined

$$\sum_{i=1}^{k} c_i\hat{\theta}_i - \sqrt{\sum_{i=1}^{k}\left[c_i\hat{\theta}_i - \min(c_i l_i, c_i u_i)\right]^2} \le \sum_{i=1}^{k} c_i\theta_i \le \sum_{i=1}^{k} c_i\hat{\theta}_i + \sqrt{\sum_{i=1}^{k}\left[c_i\hat{\theta}_i - \max(c_i l_i, c_i u_i)\right]^2}. \quad (6)$$

2.5.2. The intervals estimation for a linear function of variances

Let's defined a linear function of variances as $\sum_{i=1}^{k} c_i\sigma_i^2, k \ge 2$ where c_i are known constants. Since there are at least four intervals for a single variance as in equation (1), (3), (4) and (5) and to obtain a confidence interval for linear functions of variances via equation (6) we should have k separate confidence limits for the asymptotic variance estimates σ^2_i, i = 1,2,...,k (i.e., (l_1, u_1),...,(l_k, u_k), thus the four distinct confidence intervals for a linear functions of variances can then be easily computed as they all have closed form solutions. Hence, the 100(1-α)% traditional MOVER and its application confidence interval for a linear function of variances $\sum_{i=1}^{k} c_i\sigma_i^2$ that arise from each equation (1), (3), (4) and (5) are respectively, as follows,

$$CI_{(j)} = \left[\sum_{i=1}^{k} c_i\hat{\sigma}_i^2 - \sqrt{\sum_{i=1}^{k}\left[c_i\hat{\sigma}_i^2 - \min(c_i l_i, c_i u_i)\right]^2}, \sum_{i=1}^{k} c_i\hat{\sigma}_i^2 + \sqrt{\sum_{i=1}^{k}\left[c_i\hat{\sigma}_i^2 - \max(c_i l_i, c_i u_i)\right]^2}\right]$$

where (l_1, u_1),i = 1,2,...,k denote an available (1-α)100% confidence intervals for σ^2_i, i = 1,2,…k given by equation (j; j=1,3,4 and 5).

3. SIMULATION OUTCOMES

3.1. Method

Simulation empirical numerical study is attempted first to estimate the coverage probabilities and average widths of Equation (1), (3), (4), and (5) respectively, using 5,000 Monte Carlo random samples under different sample values of n [n = 15, 25, 50, 100, 200, and 300]. Then considered various normal populations with the same variance σ^2 [σ^2 =1 and 0.5], difference means μ and those various possible values of the coefficient of variation $CV = \sigma/\mu$ [CV =1, 0.7, 0.5, 0.25, 0.1, 0.05, 0.01, and 0.0001, practically] and level of significance, 0.05. All simulation studies presented here were obtained via the R program.

The second simulation study was then carried out to investigate the performance of the 95% confidence limits for a linear function of variances. The 5,000 sets of the normal variance values were randomly sampled. The CI_1, CI_3, CI_4, and CI_5 were used to compute the coverage probabilities (Cp) and the average interval widths (Aw) for each of 50,000 sets of variance values and for various balanced and unbalanced sample sizes. The simulation programs were again written in R (R Development Core Team, 2010) and executed on an Intel computer.

3.2. Outcomes

The table provided some comparisons of the performance of the propose interval estimations as presented in Equation (1), (3), (4), and (5) for a variety of coefficient of variations CV_S with σ^2=1 and 0.5, respectively. The results in both the cases suggest that all the estimators have coverage probabilities (Cp) close to 1-α with an increase in sample size n>50 where as a coefficient of variation CV decreases [CV<1]. Clearly, Eq.(2) is superior to all the estimators investigated, in almost all cases considered as shown in both tables since it always holds its coverage close to the nominal level well regardless of sample sizes or CV. As CV < 0.2, Eq.(2), and Eq.(3) are approximately identical but Eq.(3) gave the narrower average interval width (Aw) values. Note further that in terms of both coverage and width, the discrepancy of Eq.(2), and Eq.(3) is very small as CV decreases. That is, these two estimators are asymptotically coinciding. It also appears that, if Eq.(4) is used in a value of CV<0.1 then its average will be close to the nominal level as an increasing.

Table 1. The 95% Comparative Performance of the Proposed Interval Estimations of the Normal Variance with Selected Coefficient of Variation when s σ^2=1 based on 50,000 Runs

n	Eq.(1)		Eq.(3)		Eq.(4)		Eq.(5)	
	Cps	Aws	Cps	Aws	Cps	Aws	Cps	Aws
cv=1								
15	0.948	1.95	0.949	2.174	0.955	2.038	0	0
25	0.951	1.326	0.95	1.636	0.955	1.573	0.966	4.918
50	0.951	0.856	0.95	1.13	0.952	1.108	0.962	1.668
100	0.95	0.578	0.95	0.792	0.952	0.784	0.956	0.943
200	0.949	0.4	0.951	0.557	0.951	0.554	0.954	0.605
300	0.95	0.325	0.95	0.454	0.951	0.453	0.953	0.479
cv=.7								
15	0.948	1.95	0.949	1.522	0.952	1.474	0.965	3.172
25	0.951	1.326	0.95	1.145	0.953	1.123	0.96	1.64
50	0.951	0.856	0.95	0.792	0.951	0.784	0.957	0.927
100	0.95	0.578	0.95	0.554	0.951	0.552	0.953	0.598
200	0.949	0.4	0.951	0.39	0.951	0.389	0.952	0.404
300	0.95	0.325	0.95	0.318	0.951	0.317	0.952	0.326

Table 1. The 95% Comparative Performance of the Proposed Interval Estimations of the Normal Variance with Selected Coefficient of Variation when s σ^2=1 based on 50,000 Runs

n	Eq.(1)		Eq.(3)		Eq.(4)		Eq.(5)	
	Cps	Aws	Cps	Aws	Cps	Aws	Cps	Aws
cv=.5								
15	0.948	1.95	0.949	1.087	0.951	1.069	0.959	1.405
25	0.951	1.326	0.95	0.818	0.951	0.81	0.955	0.943
50	0.951	0.856	0.95	0.565	0.95	0.563	0.954	0.604
100	0.95	0.578	0.95	0.396	0.951	0.395	0.952	0.409
200	0.949	0.4	0.951	0.278	0.951	0.278	0.952	0.283
300	0.95	0.325	0.95	0.227	0.951	0.227	0.951	0.23
cv=.25								
15	0.948	1.95	0.949	0.544	0.95	0.541	0.953	0.544
25	0.951	1.326	0.95	0.409	0.951	0.408	0.951	0.409
50	0.951	0.856	0.95	0.283	0.95	0.282	0.952	0.283
100	0.95	0.578	0.95	0.198	0.951	0.198	0.951	0.198
200	0.949	0.4	0.951	0.139	0.951	0.139	0.951	0.139
300	0.95	0.325	0.95	0.114	0.95	0.114	0.951	0.114

In constructing confidence intervals for linear functions of variances $\sum_{i=1}^{k} c_i\sigma_i^2$, $k \geq 2$, the performance of those with no difference among variances for a variety of normal distributions are also investigated in terms of coverage probabilities and the average intervals widths for CI_1, CI_3, CI_4, and CI_5 , respectively when various group sizes are balanced and unbalanced designs (not shown the results here). Only the results at the nominal five-percent level of significant were studied. Surprisingly, for a linear function of 2, 3 or 4 variances, regardless of balanced or unbalanced designs and in all cases of considered, the results always show that as groups of observations are moderate or large, for a variety of coefficient of variations CV, the CI_1 performs substantially better than CI_3, CI_4 and CI_5 in terms of holding the mean coverage closest to the nominal level, with narrowest average width in all. For CV~.7, the CI_3 and CI_4 are identically and both have coverage close to nominal but a little bit less than CI_1 even though the sample sizes become large. In addition, we found a few instance where the coverage probabilities of CI_4 seems to perform as well as the CI_1 only at CV~.3 since its coverage are nearly the nominal levels but slightly less coverage and wider width lengths (regardless of equal or unequal variances). The CI_5 is clearly the poorest performer for all cases of our investigation, as the coverage never reaches 95% and in all cases does not maintain its nominal level.

4. CONCLUSION

We used an unconditional MOVER method for investigating confidence intervals for the linear combination of variances that generated from each kind of the normal variance estimators by combining more than one separate confidence interval for individual variances. According to our simulation results, it is suggested that with prior knowledge of the coefficient of variation one can really improve the width performances of the linear combination of variances confidence intervals based on the unbiased normal variance estimator CI_1 outperform than the rest in all cases of intentional. In particular, these estimators adequately controls the expected coverage probabilities around the pre-assigned coverage level, moreover have shortest interval widths.

Finally, we thank the work of Donner and Zou(2010), and the work of Zou, Huang and Zhang (2009).

REFERENCES

Bonett, D. G. "Approximate confidence interval for standard deviation of nonnormal distributions." *Computational Statistics & Data Analysis.* 50 (2006) : 775-782.

Casella, G. and Berger R. L. (2002): Statistical Inference, 2nd e dition,Duxbury, California.

Donner, A. and Zou, Y. "Closed-form confidence intervals for function of the normal Mean and standard deviation." *Statistical Methods in MedicalResearch*. 8(2010): 1-12.

R Development Core Team,(2010). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, Searls, Donald T. (1964): The Utilization of Known Coefficient of Variation in the Estimation Procedure, Journal of the American Statistical Association, 59, 1225-1226.

Searls, D. T., and Intarapanich, P. (1990): A Note on an Estimator for the Variance that Utilized the Kurtosis, The American Statistician, 44, 295-296.

Winston,A. R.et.al.(2010) : On Efficient Iterative Estimation Algorithm Using Sample Counterpart of the Searles'Normal Mean Estimator with Unknown Coefficient of Variation, Department of mathematics and statistics, the Pennsylvania State University, Harrisburg, USA.

Zou, G. Y., Huang, W. and Zhang, X. "A Note On Confidence Interval Estimation for a Linear Function of Binomial Proportions." *Computation Statistics & Data Analysis.* 53 (2009) : 1080 – 1085.

CPS22:
Health & Social Statistics (3)

Spatial and Temporal Distribution of Water Quality in Tropical Canal

Natthaphon Somching*
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand – nthph.sch@gmail.com

Putri Fajriati
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand – putrifajriati@gmail.com

Piamsook Chandaravithoon
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand – foncututor@gmail.com

Abstract

The objective of this study was to investigate the long-term distribution patterns of water quality parameters at the Bang Yai Canal, located in Phuket province of Thailand. The data used was obtained from the Regional Environment Office (15th edition), under the management of the Ministry of Natural Resources and Environment. Water samples were regularly collected from two different stations; Bang Khlong Bang Ta and Koh Chan Bridge, for a period of ten years between March 2005 and January 2015. Physical, chemical, and biological parameters measured were water temperature, turbidity, salinity, conductivity, total solid, total dissolved solid, pH, dissolved oxygen, biological oxygen demand, ammonia, nitrite, nitrate, total phosphorus, and total coliform bacteria. Changes in the water quality parameters between two sampling locations and over the period of study were assessed. The relationships between water quality factors and land use around the canal basin were then examined. Results showed that the water quality at the Koh Chan Bridge was dramatically worse than that at the Ban Khlong Bang Ta and overall water quality was worse at the end of the study than at the beginning. Land use was a key factor determining water quality in the studied basin. There was no statistically significant difference in the water quality between seasons at both stations.

Keywords: Land use; Spatial-temporal variation; Thailand river; Water quality indicator.

1. INTRODUCTION

Wastewater is a major pollution problem in Thailand. The main sources of wastewater are from community and industrial sectors. Population growth, advancement in agriculture, urbanization and industrialization also have made surface water pollution a great problem and have decreased the availability of drinking water (Elmaci et al., 2008). The Bang Yai Canal ("Khlong Bang Yai") in Phuket connects Kathu Waterfall to the Andaman Sea, flowing through Kathu Municipality. The increasing sediment depositions in the canal from rainfall, from river bank slumping, and from construction in the area around the canal, are the major causes of soil erosion in the watershed area (Heednacram and Samitalampa, 2014). Sediment carried in the canal water increases the physical-chemical impurities in the water.

Bang Yai canal, which leads directly to the sea from Saphan Hin, is currently so clogged with garbage and sediment that local fishermen can not use the waterway to get out to sea during low tides. Figure 1 shows Bang Yai Canal. The sediment and garbage clogging the canal is also blamed for causing floods in Phuket Town. About 22,360 cubic metres of sediment is expected to be removed by the dredging, which will make the channel at least 25 metres wide and one meter deep.

The objectives of this study were to measure and examine the differences of the physical-chemical parameters concentration in two selected stations; Ban Klong Bang Ta and Koh Chan Bridge over 10 years from 2005-2015 in the monsoonal wet season and dry season.

2. MATERIALS AND METHODS

2.1. Study Area

The Bang Yai Canal is located in Kathu district and Muang Phuket of Phuket province of Thailand, with a watershed of approximately 63 km^2 and runs through the town of Kathu (Fig. 1). The upstream is from Khuan Wa, Kathu waterfall area in Kathu district, Phuket province flowing down to the southeast through foot hills which have quite a lot of gradient. In the upper parts of the canal flowing through urban areas in Phuket City Municipality with several short tributaries. The canal has a length of approximately 17 km (Wongmongkondate, 2011).

Figure 1. Study Area of Bang Yai Canal, Phuket, Thailand

This area of Phuket province, Thailand has a wet season from April to November and a dry season from December to March. This dry–wet seasonal pattern differs from the east coast of Thailand, facing the Gulf of Thailand, because of the influences of two distinct monsoons. During the southwest monsoon, the Gulf Coast experiences relatively low rainfall, compared with that of the Andaman Coast, with its heavy rain and storms during the northeast monsoon. During the southwest monsoon, the west coast experiences relatively high rainfall, compared with that of the east coast of Thailand, with its very dry conditions during the northeast monsoon (Lueangthuwapranit et al., 2011).

The Bang Yai Canal served as a water source in the Phuket Municipality for a long time. In the past, this canal was used as a shipping canal to export tin and was also used for irrigation, transport, agriculture, and recreation. Currently, Bang Yai canal has problems with sewage pollution on a regular basis. This has an impact on the lives of people in Phuket City (Russell, 2007).

2.2. Data

2.2.1. Water Quality Data

The water sampling data obtained from the Regional Environment Office (15^{th} edition) Phuket province were collected from 2 sampling stations on the Bang Yai Canal: Ban Khlong Bang Ta and Koh Chang Bridge, the samplings were collected from 2005 until 2015 twice or three times a year. Eighty-four samples were taken during the rainy and the dry seasons. The physic-chemical parameters measured were Water Temperature, pH, Turbidity, Conductivity, Salinity, Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Total Coliform Bacteria (TCB), Ammonia (NH_3), Nitrite (NO_2^-), Nitrate (NO_3^-), Total Phosphorus (TP), Total Solid (TS), Total Dissolved Solid (TDS). There are a lot of studies measuring physic-chemical parameter on waste water, lake, pond, river etc. For example, a study in a Lake Uluabat, in Marmara Region, Turkey were analysed according to standard methods. The parameters measured included DO, Total Nitrogen, NO_3, NH_3, TP, Main cations and anions (Ca, Mg, N, K, Cl, CO_3, HCO, SO_4), Hardness, Alkalinity, Chemical Oxygen Demand, and BOD 1.

2.2.2. Land Use Data

Data used in the study were Phuket land use in 2009 and land use in 2013 obtained from the Land Development Department, Ministry of Agriculture and Cooperatives of Thailand. These land use data were regularly updated once every 5 years. Therefore, information on water quality in 2006 - 2010 was incorprorated into the land use in 2009, and information on water quality in 2011 - 2015 into the Land use in 2013. The effect of environmental factors on land-use patterns can be evaluated by establishing a correlation between water quality index and the area of corresponding patterns (Hong et al., 2016).

Methods

GIS Processes

Each sampling location boundary was delineated and land use types were classified in buffer zones within two kilometres adjacent to each location using ArcGIS 10.2 software. The categories that have been examined were Abandoned Aquacultural Land; Abandoned Tin Mine, Pit; Beach; Cashew Plantation; Casuarina; Cemetery; City, Town, Commercial; Coconut Plantation; Dense Deciduous Forest; Dense Evergreen Forest; Dense Mangrove Forest; Disturbed Mangrove Forest; Factory; Farm pond; Golf course; Grass; Harbour; Institutional Land; Jack Fruit Plantation; Marsh and Swamp; Mixed orchard; Mixed perennial; Natural water resource; Oil Palm Plantation; Para Rubber Plantation; Poultry Farm Shed; Recreation Area; Reservoir; Resort, Hotel, Guesthouse; River, Canal; Road; Scrub; Active Shrimp Farm, and Village.

Statistical analysis

Temporal and spatial distributions of the water quality parameters studied and the two-way ANOVA to identify the changes of parameters that happened over the years at both stations were performed with the R programming (2015). The factors in this analysis were Date, Month, Year, Station, and Season. The parameters that used in the analysis were Water Temperature, pH, Conductivity, Salinity, DO, BOD, NO_3^-, NO_2^-, NH_3, TP, TS, TDS, and TCB. Studies in 2015 also used ANOVA tests for significant differences among different soil and water quality parameters in relation to the five different land use types along the Mara River, Kenya and in Tasmania (Matano et al., 2015).

3. RESULT AND DISCUSSION

Figure 2 shows the physical parameters in both of stations over the years from 2005-2015, the physical parameters shown on Figure 2 are water temperature, turbidity, conductivity, salinity, TS and TDS. Graph (A) shows that the highest temperatures of Bang Yai Canal were in 2006 and Koh Chan Bridge in 2008. The dry season in Thailand makes the water hotter than its normal temperature (F-statistic: 2.612 on 9 and 56 DF, p-value: 0.01357). Even though there are no factors that were significant, from Graph (B) it can be shown that the highest turbidity at both of stations, Bang Yai Canal and Koh Chan Bridge were in 2010. Turbidity in Bang Yai Canal was much higher than at the Koh Chan Bridge, F-statistic: 15.45 on 1 and 64 DF, p-value: 0.0002102. From Graph (C) it can be shown that the conductivity in Koh Chan Bridge was higher than in Bang Yai Canal, especially in 2009. The difference of percentage was up to 62%, F-statistic: 0.9715 on 10 and 55 DF, p-value: 0.4784. Salinity in Koh Chan Bridge in 2005 had the highest salinity, see Graph (E). The salinity decreased over time, but in 2010 the salinity was up to 7.5 ppt. Salinity in Bang Yai Canal was lower thanat Koh Chan Bridge, the highest salinity in Bang Yai Canal was in 2011, (F-statistic: 20.91 on 1 and 64 DF, p-value: 2.262e-05. For TS), Koh Chan Bridge was higher than Bang Yai Canal (Graph F). The highest total solid was in 2006 and it decreased in subsequent years. For the Bang Yai Canal, the highest TS was in 2010. Total solid in Koh Chan Bridge was high because there was a larger village area than at Bang Yai Canal station (F-statistic: 7.635 on 1 and 62 DF, p-value: 0.007525). The residential area near the station has lot of impact of the water quality around the area. Koh Chan Bridge has a larger residential area and has a higher concentration of TDS than Bang Yai Canal (Graph (G), F-statistic: 95.8 on 1 and 64 DF, p-value: 2.438e-14).

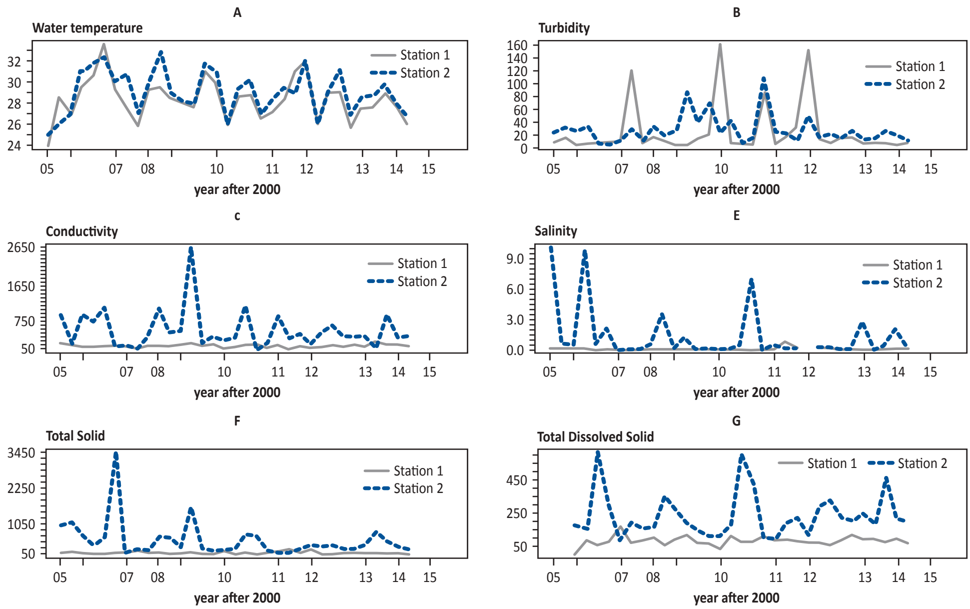


Figure 2. Physical Parameters in Both of Stations over the Years from 2005-2015

Figure 3 shows chemical parameters at both stations over 10 years, the chemical parameters shown in Figure 4 were pH, DO, BOD, NO_2, NO_3, NH_3, and TS. pH at the two stations were not significantly different, but it was obvious that Koh Chan was more alkaline than Bang Yai Canal (Graph-H), but in 2010 pH in Bang Yai Canal was highest compared to the rest of the years at both stations (F-statistic: 2.407 on 8 and 52 DF, p-value: 0.02726). Dissolved Oxygen (DO) in Bang Yai Canal was higher than at Koh Chan Bridge station G Graph (I) shows how the difference of the DO changed in in both stations over the year. The highest concentration of DO was in 2009 (F-statistic: 22.25 on 1 and 45 DF, p-value: 2.343e-05). For the BOD shown in Graph (J), Koh Chan Bridge had the higher BOD compared Bang Yai Canal station. In Koh Chan Bridge the highest concentration of BOD was in 2010, but at Bang Yai Canal the highest BOD was in 2007. The Koh Chan Bridge had higher BOD because there are lot of residential areas near the station. The wastewater and urban runoff from the residential area effected the water quality (F-statistic: 3.14 on 10 and 48 DF, p-value: 0.003678). The Nitrite results for both stations are shown in Graph (K). It shows that Bang Yai Canal in 2010 had more Nitrite than the Koh Chan Bridge. The highest concentration of Nitrite at Koh Chan Bridge was in 2010 (F-statistic: 18.51 on 1 and 60 DF, p-value: 6.336e-05). 2010 was also the year that had the highest concentration of Nitrate in both of stations among the other years: Graph (L) shows that Bang Yai Canal was the station that had more frequent high concentrations of Nitrate than the Koh Chan Bridge site (F-statistic: 14.17 on 1 and 64 DF, p-value: 0.0003657). Koh Chan Bridge had higher concentrations of NH_3 than Bang Yai Canal Station because Koh Chan Bridge station has more industry and residential areas than the Bang Yai Canal station (Graph (M), F-statistic: 38.28 on 1 and 60 DF, p-value: 5.967e-08). Beside other physic-chemical properties, Koh Chan Bridge also had the highest concentration of Total Phosphorus. Graph (N) shows how the TP increased by year, 2010 was the highest TP to appear at the Koh Chan Bridge; for Bang Yai Canal, the highest TP was in 2009 (F-statistic: 1.867 on 1 and 64 DF, p-value: 0.1766). Total Coliform Bacteria (TCB) over the years. Concentration of TCB in both of station Koh Chan Bridge and Bang Yai Canal were very high in 2012 compared to other analyzed years (Fig 7-O). Analysis of water temperature data shows that Year was the only factor that was significant (F-statistic: 28.9 on 1 and 64 DF, p-value: 1.145e-0).

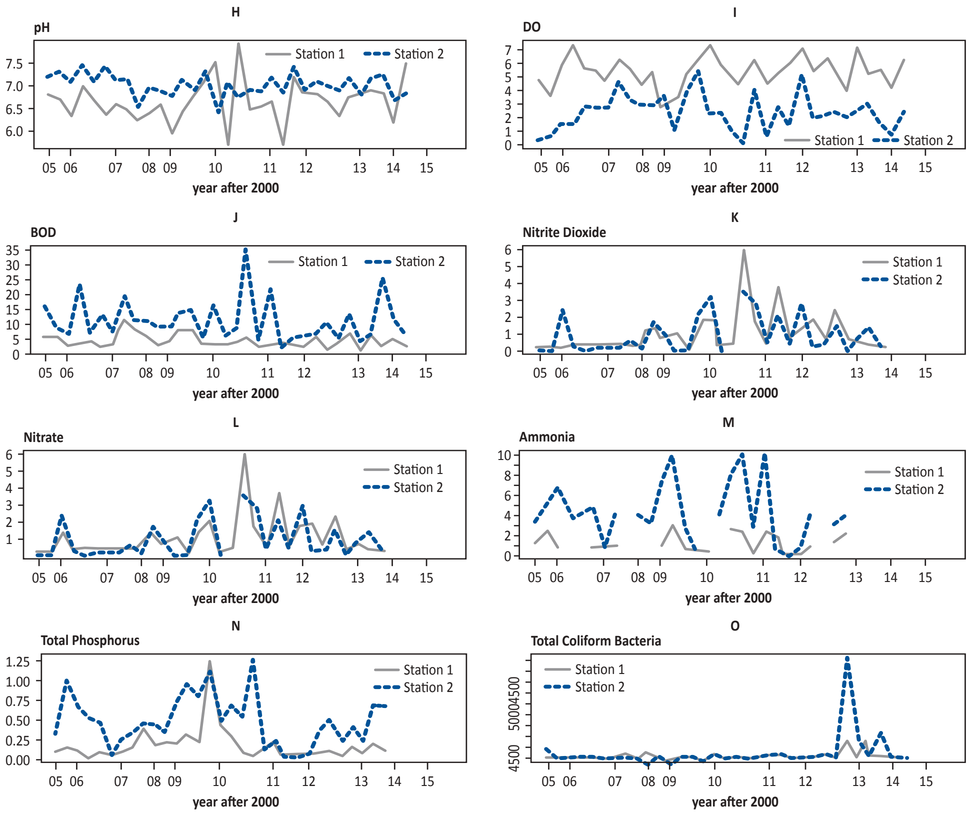


Figure 3. Chemical Parameters and Total Coliform Bacteria at Both Stations over 10 Years

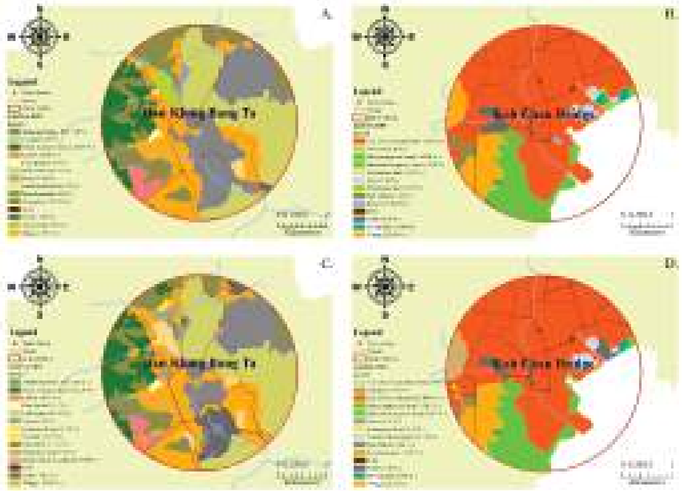


Figure 4. Land Use within 2-km Buffer Zone from Both of the Stations Examined; Ban Klong Bang Ta and Koh Chan Bridge

Figure 4 shows land use categories within a buffer of two kilometers from both of the stations examined; Ban Klong Bang Ta and Koh Chan Bridge, in years 2009 and 2013, and also shows its percentage of each category with a different colour to make it easier to understand the landuse figure and. The village area in Bang Klong Bang Ta in 2013 increased to 12.7% of the total area from 9.97% in 2009. The village area in Koh Chan Bridge in 2013 increased to 27.45% of the total area from 25.68% in 2009.

The increasing village area would be expected to effect the local water quality. Land-use change is one of the main drivers of environmental change being a major issue of global environmental change and an important component in understanding the sequence of changes in the catchment characteristics and the interactions of the human activities with the environment. This change influences the basic land resources and a variety of natural processes, including the soils, which are not static, and hence more susceptible to changes in their nutrient and moisture content (Nisar & Lone, 2013).

Land-use is a primary factor causing habitat degradation and poor water quality. Development activities such as agriculture, urbanization, forestry and industries often lead to more intensive land use which increases runoff and produces urban runoff, and consequent transport of pollutants directly into the rivers 6. Few studies on land use change and its impact on various resources of the region especially water (lakes, wetlands) and biodiversity have been carried out (Amin and Fazal, 2012; Iqbal et al., 2012; Thakor and Dharaiya, 2014). Physico-chemical characterization of certain soils in the region have also been done in some other studies [13]; [9].

4. CONCLUSION

Approximately 190 million people fall ill and 60,000 people die from a range of other diseases and injuries associated with water pollution each year (Wang and Yang, 2016). A good water quality assessment method should not only provide the water quality rank, but also accurately reflect the spatial and temporal variations of water quality condition. At the same time, a water quality assessment method which can be widely used in environmental management should be easy to calculate and master, in addition to its scientific value and accuracy (Li et al., 2016).

The chemical, biology and physical parameters from two stations: Ban Klong Bang Ta and Koh Chan Bridge were evaluated. Koh Chan Bridge Station has the higher concentration of dissolved ions, physical and biology parameters than Ban Klong Bang Ta station, because of the higher percentage of residential area around the Koh Chan Bridge and there was an increase in residential area at both sites over the study period: Koh Chan Bridge, 25% (2009) and 27% (2013) compared with Ban Klong Bang Ta which had 9.9% residential area in 2009 and 12% in 2013. Much of the pollution could be attributed to urban runoff. The presence of nitrite is a good indicator of large scale anaerobic activity in the canal sediment and water column, consistent with the high BOD levels measured. The environmental health of the canal got worse over the course of the study.

5. ACKNOWLEDGEMENTS

The authors wish to thank Prince of Songkla University for providing the computer and software facilities and funding for the project. This project was a joint project by the authors as part of a Masters in Science program in The Faculty of Technology and Environment at Prince of Songkla University-Phuket. The authors also thank Dr. Sangdao Wongsai for her help with the statistical analyses, and Associate Professor Raymond J. RITCHIE for his assistance in proof reading. The co-operation of the Regional Environment Office, Phuket Province, Thailand Ministry of Natural Resources and Environment is gratefully acknowledged.

REFERENCES

A. Elmaci, F. O. Topac, N. Ozengin, A. Teksoy, H. S. Baskaya, and L. Uluabat, "Evaluation of physical, chemical and microbiological properties of lake Uluabat, Turkey," *J. Environ. Biol.*, vol. 29, pp. 205–210, 2008.

A. Heednacram and T. Samitalampa, "Suspended Sediment Forecast of Khlong Bang Yai , Phuket," *Int. J. Eng. Technol.*, vol. 6, pp. 338–345, 2014.

Suthi Wongmongkondate, "Cultural Landscape Study of the Old Phuket Town, Graduate School, Silpakorn University. pp. 80 – 82," 2011.

C. Lueangthuwapranit, U. Sampantarak, and S. Wongsai, "Distribution and Abundance of Phytoplankton : Influence of Salinity and Turbidity Gradients in the Na Thap River , Songkhla Province , Thailand Distribution and Abundance of Phytoplankton : Influence of Salinity and Turbidity Gradients in the Na Thap Rive," *J. Coast. Res.*, vol. 27, pp. 585–594, 2011.

J. Russell, *Bangkok to Beijing via Laos and Vietnam February – April 2007*, Bangkok: Travel Diaries, 2007.

C. Hong, Z. Xiaode, and W. Wei, "Land use change and its effects on water quality in typical inland lake of arid area in China," *J. Environ. Biol.*, vol. 37, pp. 603–609, 2016.

I. Curebal and A. Soykan, "Land use and land cover change detection in Karinca river catchment (NW Turkey) using GIS and RS techniques," *J. Environ. Biol.*, vol. 33, pp. 439–447, 2012.

A. Matano *et al.*, "Effects of Land Use Change on Land Degradation Reflected by Soil Properties along Mara River , Kenya and Tanzania," *Open J. Soil Sci.*, vol. 5, pp. 20–38, 2015.

M. Nisar and F. A. Lone, "Effect Of Landuse / Landcover Change On Soils Of A Kashmir Himalayan Catchment-Sindh," *Int. J. Res. Earth Environ. Sci.*, vol. 1, pp. 13–27, 2013.

A. Amin and S. Fazal, "Land Transformation Analysis Using Remote Sensing and GIS Techniques (A Case Study)," vol. 2012, pp. 229–236, 2012.

M. Iqbal, S. M. Rashid, H. Sajjad, M. A. Siddiqui, and L. Siddiqui, "Anthropogenic impact on Landuse / Landcover in Dudhganga Watershed of Kashmir Valley, India," *Int. J. GEOMATICS* vol.2, No.3, 2012.

K. M. Thakor and N. Dharaiya, "Spatial variability mapping of micronutrients to improve productivity using Geo-informatics Technology in Gujarat state, India," *Int. J. Sci. Eng. Res.*, Vol.5, PP.1021-1027, 2014.

S. Ramzan and M. A. Bhat, "Fractionation of Zinc and their Association with Soil Properties in Soils of Kashmir Himalayas," *Int. Invent. J. Agric. Soil Sci.*, vol. 2, pp. 132–142, 2014.

Q. Wang and Z. Yang, "Industrial water pollution , water environment treatment , and health risks in China," *Environ. Pollut.*, vol. 218, pp. 358–365, 2016.

R. Li, Z. Zou, and Y. An, "ScienceDirect Water quality assessment in Qu River based on fuzzy water R Development Core Team. "R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing," Retrieved from http://www.r-project.org. 2015

Estimating Medical Treatment Costs for Violence-related Injury in Thailand

Wichayaporn Thongpeth,
PhD Candidate, Program in Research Methodology, Prince of Songkla University, Muang, Pattani, 94000 Thailand
wich14232008@gmail.com

Emml-Benjamin Atta Owusu Mintah,
Master degree, Program in Research Methodology, Prince of Songkla University Pattani, 94000 Thailand

Apiradee Lim
Assistant Professor, Program in Research Methodology, Prince of Songkla University,
Muang, Pattani, 94000 Thailand – api_45@hotmail.com
Department of Mathematics and Computer Science, Prince of Songkla University

Abstract

Hospitals in Thailand routinely record details of treatments for patients entering the emergency room. In the three southernmost provinces, many residents have suffered injuries from the terrorist insurgency that began in 2004. The data contained information from 7,404 subjects who made up to five visits (with 9,701 visits in total) during the five-year period from 2007 to 2011, We examined hospital data for the subjects who needed ER treatment on occasions. In 2007 the Ministry of Public Health in Thailand established a violence-related injury surveillance (VIS) database to record data on victims from the insurgency in the Deep South of Thailand (Narathiwat, Yala, and Pattani provinces and the four southernmost districts of Songkla province), but data on charges are incomplete. In this study, we illustrate a method of imputing unknown or unrecorded charges for visits using length of stay and number of diagnoses, based on a sub-sample of the patients who made visits. Factors considered are the principal diagnosis of the type of injury (ICD-10 code group) grouped by severity of injury, the number of diagnoses, and the length of stay. Results show that higher charges were incurred by victims suffering abdominal and pelvic content injuries.

The aim of the study was to analyze and compare some corresponding costs in Thailand. To do this we investigated costs for injuries caused by gunshot and bombs to persons entering emergency rooms at hospitals in southern Thailand.

The conclusions suggest that the linear regression model provides a good fit for the estimate of treatment costs based on the coefficients.

Keywords: Medical Treatment Costs; Violence-related Injury Surveillance; Linear regression model
JEL classification: G00

1. INTRODUCTION

The Violence-related Injury Surveillance (VIS) System for the Southern Border Provinces Area was established in January 2007 to develop the data system that would facilitate the development of emergency medical services, determination of strategies and plans, resource allocation, control and prevention of injuries, healing, and recovery for those who were affected by violence in the Southern Border Provinces area. The target group (population under surveillance) are all individuals who were injured or deceased from intentional injury who received treatment or whose autopsy was performed at the 48 governmental hospitals in Songkla, Satul, Pattani, Yala, and Narathiwat Provinces. Estimation of the average total cost for treating trauma patients is often complicated by the fact that the survival times are censored on some study subjects and their subsequent costs are unknown. This study presents the results of analysis of the data from the surveillance system, which only includes incidents resulting from the situation of unrest in the Southern Border Provinces. Injury deaths compared to other leading causes of mortality. The deaths caused by injuries, have an immeasurable impact on the families and communities affected, whose lives are often changed irrevocably by these tragedies. Injuries and violence have been neglected from the global health agenda for many years, despite being predictable and largely preventable. Evidence from many countries shows that dramatic successes in preventing injuries and violence can be achieved through concerted efforts that involve, but is not limited to, the health sector. The international community needs to work with governments and civil society around the world to implement these proven measures and reduce the unnecessary loss of life that occurs each day as a result of injuries and violence. Injuries are a global public health problem about 5 million people die each year as a result of injuries. Other main causes of death from injuries are falls, drowning, burns, poisoning, and war (World Health Organization, 2014).

Violence related injury surveillance also has diverse effects on the economy of many developing countries. The total cost of injuries and violence in the United States was $671 billion in 2013, according to two Morbidity and Mortality Weekly Reports (MMWR) released today by the Centers for Disease Control and Prevention (CDC). The cost associated with fatal injuries was $214 billion; nonfatal injuries were $457 billion (CDC, 2015). This lost cost 1-3% of gross national product of the government for low and middle-income countries annually. There doesn't report on the total cost of emergency medical service in Thailand. This amount is more than the total aid provided to low and middle income countries for developing health systems to prevent accidents (World Health Organization, 2008). Moreover, there is a high cost estimated benefit arising from preventing such accidents. Given the extent of this burden being confronted by low and middle-income countries, there is the need to prevent at the forefront of public health initiatives (World Health Organization, 2008). This study aimed to analyze and compare some corresponding costs in Thailand. And then to do this we investigated the cost of injuries caused by gunshot and bombs to persons entering emergency rooms at a hospital in the Depth south of Thailand.

2. MATERIALS AND METHODS

2.1. Data and variables

A retrospective analysis of Violence injury surveillance in Thailand. We obtained relevant data from the Deep South Coordination Centre (DSCC) database, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand. Data management for systematic analysis of the violence injury surveillance data were checked for errors and missing records. All of the missing values were cleaned before performing data analyses. Since cost of medical treatment was not complete recorded, we estimated the cost based on the coefficients estimated from linear regression model of VIS data. The model has three categorical determinants comprising length of stay, number of diagnosis and diagnosis group. The estimated violence related injury treatment costs were obtained and filled in the database.

2.2. Statistical methods

The relationship between diagnosis group, length of stay, and number of diagnosis of violence related injuries for each determinants will be analyzed using linear regression model. The medical treatment cost of violence related injuries will be estimated using linear regression model.

Linear regression

The linear regression can be used to explore the relationships between a dependent variable and a set of independent variables. The general method of estimation that leads to the least squares function under the linear regression model (when the error is normally distributed) is called maximum likelihood. Under the assumptions of linear regression, the method of least squares yields estimators with a number of desirable statistical properties. The specific form of the linear regression model is as equation 1.

$$y_i = a + \sum_{i=1}^{n} b_i x_i + \varepsilon_i \tag{1}$$

Where, is outcome (diagnosis group, length of stay, number of diagnosis), *a* is an intercept of y, the value of y when equals zero, $_1, _2, _3, \ldots, _9$ are, mechanism of injuries and respectively are determinants, b_1, b_2, $b_3, \ldots, b_9$ *are* coefficient of each variable, ε_i is the error.

In addition, the estimate medical treatment costs from linear regression model

We call *y* the independent or response variable (estimated cost of injuries); x_1, x_2, and x_3 are the independent variables (diagnosis group, length of stay, number of diagnosis). We call *a* the "true" intercept term; b_1, b_2, and b_3 are called the slope or gradient of the line. It is the increase in y corresponding to an increase of one unit in x (Bland, 2000; McNeil, 2006).

Linear regression is a statistical method widely used to model the association between a continuous outcome and a set of fixed determinants. The model expresses the outcome variable as an additive function of the determinants. We used R software (R Development Core Team, 2011) to produce all statistical results and graphs.

2.3. Analysis strategy

To remove skewness in the linear model we transformed the incidence rates by taking their logarithms, after replacing zero counts by 0.5 to ensure finiteness. We fitted models with two additive factors as determinants for statistical reasons aimed at reducing the standard errors of the estimated parameters. This model was chosen because it is arguably more appropriate for studying.

3. RESULTS

A total of 9,701 visits and 7,404 patients were available for the economic analysis. After adjusting for patient demographic and characteristics included the number of conditions present on admission, and length of stay. As the length of stay is also often right-skewed, we transformed it to log (length of stay+1)

From this regression, we then calculated a given hospital adjusted charge for the average statewide patient for that DRG, where the adjusted charges represented standardized log charge/(day+1). This gave us a singer adjusted charge per day for each event, representing the predicted charge for a patient with the same average clinical and demographic and characteristics, which we then used as the dependent variable in our second stage regression.

4. DISCUSSION

We considered 9,701 visits from 7,404 patients entering hospitals in four provinces during year 2007-2011.We used commands in R program and structured these data as a database table indexed by each patient'13 –digit Personal Identification Number and injury and hospital entry date/time with 20 fields as follows. Main outcomes are the cost (in Baht) and final outcomes are status arrive in hospital 0 or 1: died. We separated patients into groups according to the number of occasions they visited a hospital. We gained some detailed understanding of the process, we now focus on the three subjects who made five visits using the following R commands. We viewed data for the three subjects with the most visits. We saw that the first of the three subjects. When doing a study like this it is useful to get some understanding of the process generating the data before embarking on further analysis, and carefully investigate the information available from them in detail. Next step, we analyzed the costs associated with visits from patients entering the emergency room in southern Thai hospitals for treatment of terrorism- related injuries. Since many of these charges do not appear in the database, we will develop a method for imputing these missing these missing outcomes, and illustrate this method using a relatively small sample of patients with three or four visits. We fitted model to charge using number of diagnosis and length of stay for predicting charges Charge unknown Charges have skewed distributions so need to be transformed to satisfy the normality assumption in linear models, then transformed back to get fund total. There subjects had 881 visits, their lengths of stay (LS) ranged from 0 to 215 days. But there are 171 missing values for length of stay. We created the grouped length of stay varies LS1 to have 8 levels, after imputing missing values to be the median (0). and then we created grouped variable number dignosis1 to have 5 levels by combining last three levels. The charge unknown 253 visits and their charge unreported code as -1. This data have charge having a highly skewed distribution. And then we created un-skewed variable charge1 after replacing missing values by NA and transforming using logarithms. We separated data according to charges reported or omitted. And then we fit a linear model to charges reported charges. This model accounts for 50 percentage of squared variation. The normality assumption for the errors is plausible. Use model regression coefficients to impute values of charge1 for cases with unreported charge. We merged the data tables with the reporters and imputed charges. Transform to baht. We could compare the total charges for the reported and the reported plus imputed amounts.

The estimated total charge for the 881 visits by the 284 patients who had 3 or 4 visit was 13.7 million bath. Or the average 48,289 baht per visits. Our estimate of the medical cost of each violence fatality in Thailand in 2001-2011 was 26,126,325 Baht but the government paid true cost being 50,967,541baht. We calculated the number and cost of injuries by violence related injuries using the method in Thailand, between 2007 and 2011.

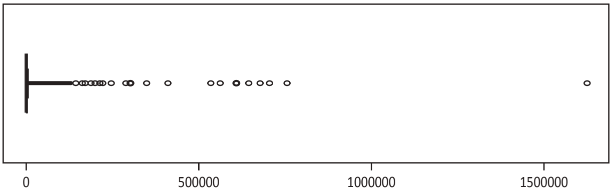


Figure 1. Box Plot Graph Show Medical Cost Among 9701 Visits (P =.005). Reported Hospital Charges from Each Visit and 75th Percentiles, with the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

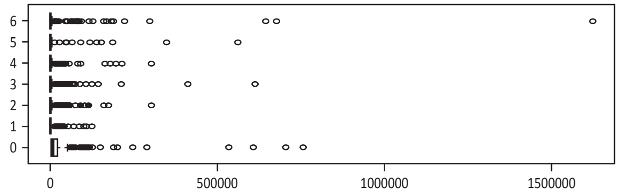


Figure 2. A Box Plot Graph Comparing Medical Cost Among 9701 Visits (P =.005). Reported Hospital Charges from Each the Number of Diagnosis. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

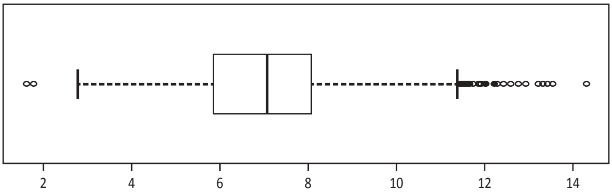


Figure 3. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each the Length of Stays Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median.

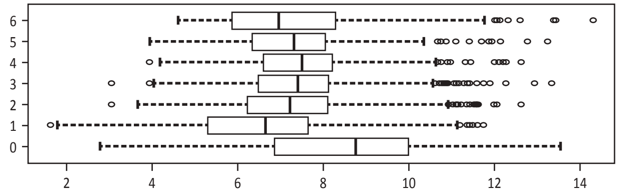


Figure 4. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each the Number of Diagnosis were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

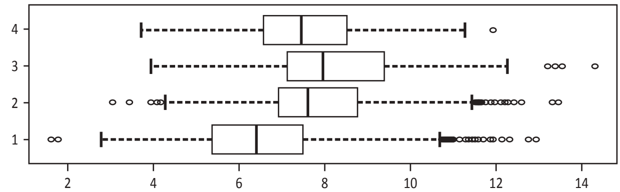


Figure 5. Box Plot Graph Comparing Mean the Length of Stay 14 Days (P = .005). Reported Hospital Charges From Each Diagnosis Group Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th and 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

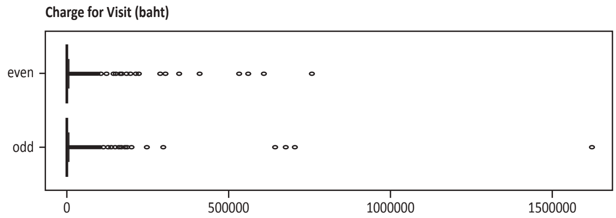


Figure 6. Box Plot Graph Comparing Mean Costs of Medical Cost Among 9701 Visits (P = .005). Reported Hospital Charges From Each Institution Were Converted to Costs by Applying Medicare Cost-to-charge Ratios.

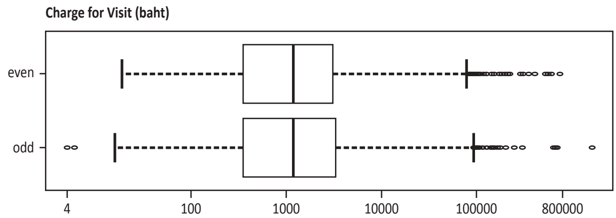


Figure 7. Box Plot Graph Comparing Mean Costs of Trauma Care Among 7404 Patients (P = .005). Reported Hospital Charges From Each Institution Were Converted to Costs by Applying Medicare Cost-to-charge Ratios. The Box is Defined by the 25th And 75th Percentiles, With the Line Through the Box Representing the Median. The Whiskers on the Box are Equal to the 150% of the Interquartile Range Centered at the Mean

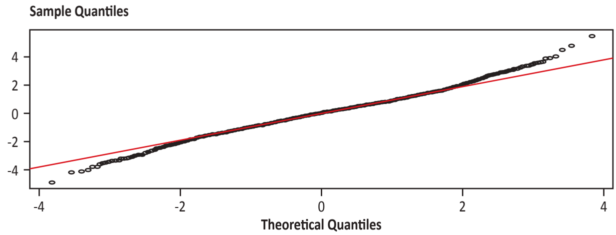


Figure 8. A Scatter Plot Showing the Data of Hospital Charge in Violence Injury Surveillance Compare the Theoretical Data.

5. CONCLUSIONS

Much of appeliez statistics may be viewed as an elaboration of the linear regression model and associated estimation methods of least-squares. In beginning to describe these techniques Mosteller and Tukey (1977) in their inuential text remark: What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily, this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributionsQuantile regression methods useful to describe associations between grouped length of stay variable and hospital charge.And then it can describe associations between grouped number of diagnoses variable and hospital charge in Thailand. To do this we investigated costs for injuries caused by gunshot and bombs to persons entering emergency rooms at hospitals in southern Thailand. We suggested that the linear regression model provides a good fit for the estimate of treatment costs based on the coefficients.

REFERENCES

Ash, A.A., Ellis, R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H. Lezzoni, L.I., Mackay, E., and Yu, W. (2000). Using diagnosis to describe poputations and predict costs. Health Care Financing Review. 21(3), 7-28.

Brilleman, S.L., Gravelle, H., Hollinghurst, S., Purdy, S., Salisbury C., Windmeijer, F. (2014). Journal of Health Economics. 35, 109-122.

McNeil, D.(2007). Modern Statistics, Pearson SprintPrint Ltd, Australia, pp. 229-233.

Nilsen, P., Hudson, D., and Lindqvist, K.(2006). Economic analysis of injury prevention applying results and methodologies from cost-of injury studies. International Journal of Injury Control and Safety Promotion.13(1), 7–13.

World Health Organization. (2014). Special Tabulation Lists for Mortality and Morbidity; Mortality tabulation list1. In International Statistical Classification of Diseases and Health Related Problems Tenth Revision Volume 1. World Health Organization,Geneva, Switzerland, pp.2-5.

World Health Organization. (2014). Preventing Injuries and Violence: a Guide for Ministries of Health. World Health Organization, Geneva, Switzerland, pp.6-9.

World Health Organization. (2014). The Global Burden of Disease: 2004 Update. World Health Organization, Geneva, Switzerland, pp.120-121.

Yang, L., Lam, L.T., Liu, Y., Geng, W.K., and Liu, D.C. (2005). Epidemiological profile of mortality due to injuries in three cities in the Guangxi Province, in China. Accident Analysis and Prevention. 37(1), 137-141.

Statistical Modeling for Classification Cause of HIV Death based on the 2005 Verbal Autopsy Data

Amornrat Chutinantakul*
Office of Disease Prevention and Control Region 11, Nakhon Si Thammarat, Thailand - yaiamorn@yahoo.com

Don McNeil
Prince of Songkla University, Pattani Campus, Pattani, Thailand - don.mcneil@mq.edu.au

Phattrawan Tongkumchum
Prince of Songkla University, Pattani Campus, Pattani, Thailand - phattrawan@gmail.com

Kanitta Bundhamcharoen
International Health Policy Program (IHPP), Bureau of Policy and Strategy, Ministry of Public Health, Nonthaburi, Thailand - kanitta@ihpp.thaigov.net

Abstract

Thai death registration data is known to be of poor quality with respect to cause of death, mainly because certificates for deaths outside hospital are not necessarily certified by doctors. However, a verbal autopsy (VA) data were used to estimate individuals' true causes of death. This sample comprised 9,495 cases of reported deaths aged 5 years and older. Using HIV cause as outcome, and province, gender-age group and reported cause in and outside hospital as determinant factors, we compared three data analytic approaches, (1) a simple cross-referencing method used by the medical scientists who first analysed the data and published results in 2010, (2) a complex logistic regression model developed by bio-statisticians and published in 2014, and (3) a very simple recursive partitioning method proposed by data scientists. Using area under the ROC curve above the diagonal as a measure of goodness-of-fit, we found that (1) gave 76%, (2) gave 94%, and (3) gave 89%. Our conclusion is that for such data the tree-based method compares favourably with the purely statistical approach, mainly because it automatically handles interactions between study factors and does not require data transformations.

Keywords: Simple cross-referencing, Logistic regression, Recursive partitioning, Verbal autopsy data

1. INTRODUCTION

Thai mortality data are of poor quality (Prasartkul et al. (2007); Rao et al. (2010); Vapattanawong & Prasartkul (2011)). Since 1996 to 2015, they presented ill-defined varied from 27.2% - 41.7% (Bureau of Policy and Strategy, 2010). High percentages of ill-defined lead to inaccuracy of cause of death. Due to high percentages of ill-defined in death registry in Thailand, verbal autopsy method was used to determine cause of death. The latest verbal autopsy (VA) was conducted in 2005. The results presented in four studies (Rao et al., 2010; Pattaraarchachai et al., 2010; Polprasert et al., 2010; Porapakkham et al., 2010). However, classification cause of deaths is not used statistical modeling. Byass (2010) mentioned that the estimation of mortality from the 2005 VA study had some uncertainties and suggested the probabilistic modeling.

Aim of this study is to propose an appropriate method to classify cause of HIV death. The 2005 VA data were used as gold standard to correct misclassification of cause of HIV deaths. Simple cross-referencing, a tree-based method known as recursive partitioning and logistic regression methods were compared.

2. METHODOLOGY

2.1. Statistical approaches

We compared three data analytic approaches, such as 1) a simple cross-referencing method, 2) a complex logistic regression model and 3) a very simple recursive partitioning method.

2.1.1. *A simple cross-referencing method* used by the medical scientists who first analysed the mortality data (Porapakkham et al., 2010). It is similar to simple logistic regression.

Logistic regression (McNeil, 1996; Woodward, 1999) formulates the logit of the probability that a person died from HIV as a linear function of the determinant factors. A simple cross-referencing is formulated as

$$logit(P_i) = ln\left(\frac{P_i}{1-P_i}\right) = \mu + \alpha_i$$

where P_i is the probability of death due to HIV, μ is a constant and α_i is the only parameter of DR cause-location i (Chutinantakul et al., 2014).

2.1.2. *A complex logistic regression model* developed by bio-statisticians. It includes an additive linear function of the three determinant factors, which could be expressed as

$$logit(P_{ijk}) = ln\left(\frac{P_{ijk}}{1-P_{ijk}}\right) = \mu + \alpha_i + \beta_j + \gamma_k$$

where P_{ijk} is the probability of death due to HIV and α_i, β_j and γ_k are individual parameters specifying DR cause-location group i, sex-age group j and province k, respectively (Chutinantakul et al., 2014).

2.1.3. *A very simple recursive partitioning or tree-based method*

A simple recursive partitioning method works for both categorical and continuous input and output variables. It is mostly used in classification problems. This method split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter/differentiator in input variables.

The endpoint for a tree is a partition of the space X. We compare trees by how well that partition corresponds to the correct decision rule for the problem. In logical problems the easiest way to compare partitions is to count the number of errors, or, if we have a prior over the space X, to compute the probability of error. The recursive partitioning method is described by Venables and Ripley (2002). The main benefit of using this algorithm is that it automatically (1) unearths informative interactions between variables and (2) incorporates transformations using multiple binary splits on predictors.

2.2. Data for the comparison

The 2005 VA data were used as a tool to correct misclassification cause of HIV deaths. This data comprised a sample of 9,644 deaths (3,316 in-hospital and 6,328 outside-hospital) from 28 selected districts in nine provinces of four regions. These were Bangkok, Nakhon Nayok, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Suphan Buri, Chumphon, and Songkhla. The VA data comprised variables of interest. These are sex, age, region, location of death, ICD-10 reported cause groups and ICD-10 VA-assessed cause groups.

This study was confined to deaths of persons aged 5 years and older, for which HIV death is common and often misclassified (age under 5 years separate for special attention in other study and few cases of HIV deaths). Thus, this study comprised the sample of 9,495 deaths (3,212 in-hospital and 6,283 outside-hospital). Accordingly, the chapter-block classifications of ICD-10 codes that categorized mainly by human organs, they were used to create 21 major cause groups based on the distribution of VA-assessed deaths that obtain sufficient sample size. For HIV cause group, it comprised of 512 cases at age 5 years and older.

2.3. Model's assessment

The Receiver Operating Characteristic (ROC) curve reflects the relationship between sensitivity and the false positive rate (1- specificity). Areas under the ROC curve (AUC) are used to evaluate models ability (Fan, et al., 2006; Sakar, & Midi, 2010). The AUC measures the performance of a model and represents model accuracy. An area of 1 represents a perfect model; an area of 0.5 represents a worthless model. This study used the AUC for comparing models ability for the recursive partitioning method, simple and complex logistic regression model. The greater area indicated the appropriate model.

3. RESULTS

We compared three methods for predicting HIV death from the 2005 VA data comprised of 9,495 deaths by fitting the simple logistic regression model, complex logistic regression model and recursive partitioning method. We assess models ability using the ROC curve. The AUC above the diagonal line is used as a measure of goodness-of-fit.

Figure 1 shows the ROC curves representing three methods plotted on the same graph. The accuracy of the model depends on how well the model separates the group being tested into those with and without HIV deaths. The simple cross-referencing model comprised of only factor of DR causes and location of death (drGh only in Figure 1) as shown in pink line gave areas 76% whereas the ROC curves from complex logistic regression model and recursive partitioning method are shown in black and brown lines that gave areas 94%, and 89%, respectively. Complex logistic regression model gave a bigger area than recursive partitioning method 5% and 18% for a simple cross-referencing model. In general, ROC curves with an AUC equal or less than 75% are not clinically useful.

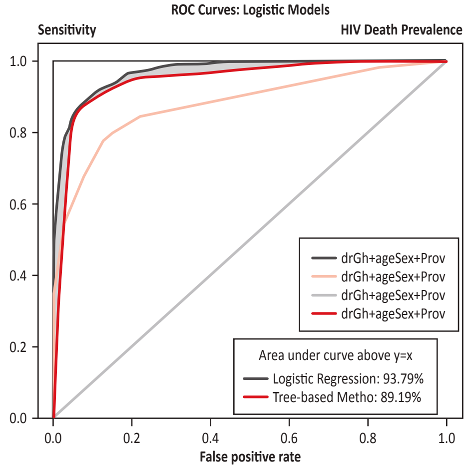


Figure 1. The ROC Curve of Logistic Regression Model and Tree-based Method

4. DISCUSSION

Many methods are used to build a prediction. Our study aimed to compare the effectiveness for classification cause of HIV deaths from three methods comprising simple cross-referencing model, complex or multiple logistic regression model and tree-based or recursive partitioning method. We found complex logistic regression model outperform the others two. Our complex logistic regression model was used to estimate deaths from HIV (Chutinantakul et al., 2014).

This study illustrated that complex logistic regression model performed better than the tree-based method. This result accorded to a study of comparison of decision tree and logistic regression model in transportation problem (Chen, 2009) including the study of classifying acute cardiac ischemia (Long, 1993). On the other hand, some studies showed no significant different in the two methods, such as the study of diagnosis of carpal tunnel syndrome (Rudolfer, Paliouras, Peers, 1999).

Although, the complex logistic regression model is slightly outperform the recursive partitioning but the advantages of recursive partitioning or tree-based method are easy to understand, useful in data exploration, less data cleaning required and data type is not a constraint. Tree-based method is considered to be a non-parametric method that have no assumptions about the space distribution and the classifier structure.

5. CONCLUSIONS

This study aimed to propose an appropriate method to classify cause of HIV death. For such data the recursive partitioning method compares favourably with the purely statistical approach, mainly because it automatically handles interactions between study factors, does not require data transformations.

REFERENCES

Bureau of Policy and Strategy. (2010). Death registration database. Ministry of Public Health.

Byass, P. (2010). Integrated multisource estimates of mortality for Thailand in 2005. Population Health Matrics.

Chen, X. (2009). A comparison of decision tree and logistic regression model. Available http://www.mwsug.org/proceedings/2009/stats/MWSUG-2009-D02.pdf

Chutinantakul, A., Tongkumchum, P., Bundhamcharoen, K. & Chongsuvivatwong, V. (2014). Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009. Population Health Matrics.

Fan, J., Upadhye, S. & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine.

Long, W.J., Griffith, J.L., Selker, H.P. & Agostino, R.B.D. (1993). A comparison of logistic regression to decision-tree induction in a medical domain.

McNeil, D. (1996). Epidemiological Research Methods. John Wiley & Sons.

Pattaraarchachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Pao-in, W., Singwerathum, N. & Lopez, A.D. (2010). Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. Population Health Matrics.

Pigott, T.D. (2001). A review of methods for missing data. Educational Research and Evaluation.

Polprasert, W., Rao, C., Adair, T., Pattaraarchachai, J., Porapakkham, Y. & Lopez, A.D. (2010). Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. Population Health Matrics.

Porapakkham, Y., Rao, C., Pattaraarchachai, J., Polprasert, W., Vos, T., Adair, T. & Lopez, A.D. (2010). Estimated causes of death in Thailand, 2005: implications for health policy. Population Health Matrics.

Prasartkul, P., Porapakham, Y., Vapattanawong, P. & Rittirong, J. (2007). Development of a verbal autopsy tool for investigating cause of death: the Kanchanaburi project. Journal of Population and Social Studies.

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swarnpunyalert, N. & Lopez, A.D (2010). Verifying causes of death in Thailand: rationale and methods for empirical investigation. Population Health Matrics.

Rudolfer, S.M., Paliouras, G. & Peers, I.S. (1999). A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. Computer and Biomedical Research.

Sakar, S. & Midi, H. (2010). Importance of assessing the model adequacy of binary logistic regression. Journal of Applied Sciences.

Vapattanawong, P. & Prasartkul, P. (2011). Under-registration of deaths in Thailand in 2005-2006: results of cross-matching data from two sources. Bulletin of the World Health Organization.

Venables, W.N. & Ripley, BD. (2002). Modern Applied Statistics with S. Springer.

Woodward, M. (1999). Epidemiology: Study Design and Data Analysis. Chapman & Hall.

This page intentionally left blank

IPS18:
Statistics and Public Health

Statistical Methods for Public Health and Medicine

Haikady N. Nagaraja *
The Ohio State University – nagaraja.1@osu.edu

Abstract

Probabilistic modeling, statistical design, and inferential methods form the backbone of the remarkable advances in medicine and public health. General goals of inference are hypothesis testing (as in clinical trials), estimation (of risk for a disease), and prediction (of a future condition). We illustrate them by introducing examples, data types, statistical models, and methods. With summary statistics on commonly used statistical concepts in major public health and medical journals, we discuss popular statistical methods that drive current research in public health and medical science. We examine trends in biostatistical research and observe the evolving field of data science and bioinformatics.

Keywords: Statistical Methods, Public Health, Statistical Design

1 For examples of this public debate refer to Bindseil et al. (2015)

Statistical Assessment of Blinding

Anil Gore, Sharayu Paranjpe, Meenakshi Mahanta, Ashwini Joshi, Prashant Kulkarni
CYTEL Statistical Software and Services Pvt. Ltd. Pune (India) - anil.gore@cytel.com

Abstract

Blinding is an important aspect of the design of a clinical trial. Traditional approach to measuring efficacy of blinding is 'the blinding index' based on conjecture of subjects [about the treatment they received] at the end of the trial. This paper examines a situation with subject responses recorded at multiple time points. The key issue is how to test if there is progressive unblinding. Two related questions are (a) how to measure the extent of unblinding and (b) how to apportion it between primary cause (trial design) and secondary cause (AE or efficacy effect). It is indeed possible to answer all these questions. Sections 1-3 prepare the backdrop and Section 4 states the precise question of interest. Section 5 addresses this question using various alternative methods. One method is generalized McNemar test for marginal homogeneity. Second is a weighted least squares approach described by Stokes et al (2000). Third is application of polytomous logistic regression. Fourth and last is a simulation approach. This is the main part of the paper. Other questions are answered in section 6. Numerical illustrations are given based on artificial data. Relevant program codes are available if needed.

1. INTRODUCTION

Clinical trials are experiments in which efficacy of a drug/treatment is tested on human subjects. Blinding (keeping a subject/investigator ignorant of which treatment the subject is getting) is a vital part of the design of a clinical trial. It is now practiced widely. This is because belief about or knowledge of the treatment given (unblinding) can affect the response. A subject can experience improvement in health even when she is on placebo provided she <u>believes</u> that she is getting a drug. Clearly this placebo effect has to be factored in while judging effect of a drug. Even treating and evaluating doctors have to be blinded to ensure that their actions or assessments are not influenced by the knowledge of who is getting what treatment. How effective is the provision for blinding in a trial? In other words, does the subject in fact remain ignorant of treatment received or does she guess it right? If blinding is not effective, the provision for blinding may be no more than a ritual. No wonder regulators are keen to find out if in fact the attempt at blinding was successful (fully or partially).["*We recommend that you administer a questionnaire at study completion to investigate the effectiveness of blinding the subjects and treating and evaluating physicians" (FDA 2003). "DRUDP requests that subjects and investigators state at the end of the subject's participation as to what treatment assignment they think was made, in order to assess the adequacy of blinding*"] FDA recommends reporting "<u>how the success of blinding was evaluated</u>." Present document reviews methods proposed in literature for evaluation of blinding when subjects guess [at the end of the study] the treatment received. It then takes up the issue of assessing progressive unblinding when the subject gets multiple opportunities to guess the treatment received. Need for this extension appears to have escaped attention of clinical trial statisticians. Various alternatives are suggested. The intent is to adapt inference procedures available in statistical literature to assess progression in blinding.

2. ASSESSMENT OF BLINDING

Conventional method of checking effectiveness of blinding is using a blinding questionnaire administered to subjects at the end of the trial. Let us assume that we have a trial with two treatment arms namely T (test) and P (placebo). Each subject is asked to guess which group (s)he belongs to. This gives rise to a 2X2 table of counts shown below.

Table 1. Results of Blinding Questionnaire (Counts)

	Subject's Guess		Total
True grouping	T	P	
T	N_{11}	N_{12}	N_{1*}
P	N_{21}	N_{22}	N_{2*}
Total	N_{*1}	N_{*2}	N

So as shown in the above table, there were altogether N subjects out of which N_{1*} subjects received the test and the remaining (N- N_{1*}) i.e. N_{2*} subjects received the placebo. In response to the question about their guesses, N_{11} subjects correctly guessed that they received T while N_{12} subjects wrongly thought that they got P etc. Our numerical measure of blinding has to be based on the four cell counts.

What does it mean if all subjects know which is the treatment arm for them (Test or Placebo)? It means blinding attempts failed miserably. So it is clear that $N_{1*} = N_{11}$ (and hence N_{12} is zero) shows poor blinding. The same is true if N_{2*} equals N_{22} (and N_{21} is zero). These are extreme cases. In more realistic cases, there will be non-zero counts in all four cells. Ideally, if the blinding procedure works, the true grouping and subject's guess should be independent. It can be checked by applying the chi-square test of independence of two attributes. This method was proposed by (Hughes & Krahn , 1985).

There is one hurdle in this method. It assumes that each subject can make a guess. In practice, some subjects plead ignorance and simply say that they do not know. In that case the investigator has two options. One is to force the subjects to make a choice (which is not feasible or advisable) or alternatively to record these cases as DK (do not know). This latter alternative leads to the so called 2X3 format for the table as shown below.

Table 2. Results of Blinding Questionnaire allowing 'Don't Know' (Counts)

	Subject's Guess			Total
True grouping	T	P	DK	
T	N_{11}	N_{12}	N_{1DK}	N_{1*}
P	N_{21}	N_{22}	N_{2DK}	N_{2*}
Total	N_{*1}	N_{*2}	N_{*DK}	N

We will generally work with this format in our paper. To check effectiveness of blinding we can still use the chi-square test of independence if we ignore the third column DK. This seems like failure to use key information. If everyone says DK, then it suggests very effective blinding. So we can apply the chi-square test to the entire 2X3 table. A more popular approach is that of James et al (1996) who proposed an index for blinding efficacy. This index BI takes values between zero and 1. If all guesses are correct, there is no blinding and value of BI is zero. If everyone responds DK, BI equals 1 and there is complete blinding. If there is no case of DK, and half the guesses are right and the other half are wrong, then BI =0.5 (random guessing). Authors of this index have given a formula for SE of the estimated index (based on observed counts). Assuming normality, a confidence interval can be calculated for the true population value of BI. If this interval falls entirely below 0.5, it is regarded as evidence of poor blinding.

Among other things, this index is for the entire trial and does not give separate assessments for the two arms. Bang et al (2005) have proposed another index competing with James index. The Bang index is supposed to rectify some of the deficiencies of the James index. In this paper we will restrict ourselves to James BI.

3. PROBLEM OF EXTENSION OF JAMES INDEX

We now introduce another degree of realism in the problem. Here the basic 2X3 table is obtained at multiple time points during the trial. The concern is that after experiencing adverse events or efficacy benefits, subjects may be able to improve their success rate in guessing whether they are in the test group or placebo. In other words, there may be progressive unblinding. Three questions are of interest here. (i) Is there a statistically significant progressive unblinding? (ii) If yes, what is the extent of unblinding? (iii) How can we apportion this effect between primary (design of the trial) or secondary (AE/efficacy) causes?

4. TEST FOR PROGRESSIVE UNBLINDING

The main thrust of literature on assessment of blinding is on estimation and there seems to be a lack of emphasis on testing hypotheses. In fact a formal test would be very useful. Let us begin by recognizing that we are now handling a three way contingency table. Rows are treatment arms, columns are guesses at time point 1 and sub-columns are guesses at time point 2 as shown in Table 3.

Table 3. Counts of Responses to Blinding Question (at two time points)

	Guess at time point 1								
	T			P			DK		
	Guess at time 2			Guess at time 2			Guess at time 2		
Treatment	T	P	DK	T	P	DK	T	P	DK
T	n_{TTT}	n	n_{TTDK}	n_{TPT}	n_{TPP}	n_{TPDK}	n_{TPT}	n_{TPP}	n_{TPDK}
P	n_{PTT}	n	n_{PTDK}	n_{PPT}	n_{PPP}	n_{PPDK}	n_{PDKT}	n_{PDKP}	n_{PDKDK}

At each time point we have a 2X3 table with true treatment arms as the rows and subject's guess as the columns. We can compute James BI or any other measure of blinding at each time point. What is not clear is how to test

H0 : unblinding remains the same at both time points against the alternative

H1 : unblinding worsens over time.

This question appears to have been ignored in literature thus far. It can be approached in different ways. Note that the alternative hypothesis is essentially one sided. However some tests using chi-square statistics may be inherently two sided. In that case, we can examine the direction in which counts are changing to judge which way the wind blows.

We begin with a simple intuitive idea. If there is progressive unblinding over time, then the proportion of DK should decline. (The extent of decline may be different in T and P groups. Perhaps it will be greater in T than P). So we will cast the data in the form of 2Xk table of counts. Rows will be DK and [Tor P]. (we have the option of keeping T and P separate). Columns will be time points. We can test the null hypothesis that probability of DK is the same at all time points. This is a simple homogeneity test. We can improve it a bit by adopting an ordered alternative that the proportion decreases over time. If the null hypothesis cannot be rejected, we will infer that secondary unblinding has not occurred. If the null hypothesis is indeed rejected (as we expect) we will conclude that there is statistically significant progressive unblinding and then the issue of how to measure it etc will have to be taken up. There are more sophisticated test procedures in literature and we will review them here.

(a) Generalized McNemar test- When there are two time points and each subject chooses T or P or DK as the guess at each time point, there are 9 possible combinations. This leads to a 2X9 contingency table, rows being the trial arm and columns being the choices made at the two time points. Now we can arrange the counts in 9 cells in a row into a 3X3 table as shown below.

Table 4. Data for Test of Marginal Homogeneity

	Guesses at time point 2				
Guesses at Time point 1		T	P	DK	Marginal totals
	T	n_{TT}	n_{TP}	n_{TDK}	$n_{T.}$
	P	n_{PT}	n_{PP}	n_{PDK}	$n_{P.}$
	DK	n_{DKT}	n_{DKP}	n_{DKDK}	$n_{DK.}$
Marginal totals		$n_{.t}$	$n_{.P}$	$n_{.DK}$	Total count= n..

If the situation is similar at two time points, then the marginal totals should also be similar. This can be tested using a test for marginal homogeneity. The procedure is described on page 253 of Agresti (2007). If the null hypothesis is rejected it implies that degree of unblinding has changed over time. In general the test statistic is a quadratic form. When the square table has only 3 rows and 3 columns there are explicit formulas given by Xuezheng Sun and Zhao Yang (2008) (in http://www2.sas.com/proceedings/forum2008/382-2008.pdf).

Here is an illustration of this test.

Following is the assumed data set with 100 subjects.

Table 5. Assumed Data Set with 100 Subjects for Test of Marginal Homogeneity

Treatment arm	Pair of guesses at two time points								
	TT	TP	TDK	PT	PP	PDK	DKT	DKP	DKDK
T	5	5	2	5	5	2	8	9	9
P	5	4	3	7	4	2	5	5	15
Total	10	9	5	12	9	4	13	14	24

Note that the first column with title TT gives counts of subjects who guessed TT at the two opportunities. Among the 50 subjects in the group T, 5 subjects guessed that they were on test product (on query at both occasions).

Hypothesis to be tested is

Ho: 'Degree of unblinding does not change over time' against the alternative

H1: Degree of unblinding changes over time.

Test Statistic is

Z0 =[$\underline{n23}$(d1)^2+ $\underline{n13}$(d2)^2+ $\underline{n12}$(d3)^2]/ 2 ($\underline{n12}$ $\underline{n23}$ + $\underline{n12}$ $\underline{n13}$ + $\underline{n13}$ $\underline{n23}$)]

which follows a chi-square distribution with 2 d.f when the null hypothesis is true.

We will to do the analysis separately for T and P groups. First the T group

Table 6. Test of Marginal Homogeneity for T Group

		Time 1			
		T	P	DK	Total
Time 2	T	5	5	8	18
	p	5	5	9	19
	DK	2	2	9	13
	Total	12	12	26	50

d1= 12-18= -6 d2= 12- 19 = -7 d3= 26-13=13 ; n12bar= 5, n13bar=5 (8+2/2), n23bar= 14 so the statistic is Z0= [14*36+5*49+5*169]/2(5*5+5*14+5*14)= 4.83 and p value is 0.09 > so for T group there is no change over time. Results for P group are similar.

(b) Comparison of average scores (weighted least squares approach): Here we visualize two arms (T and P) of the clinical trial as two populations. Guesses T, P and DK are categories of response (second dimension) and time point is the third dimension. So the data are seen as a table below. Here cell entries are counts of subjects in a given arm with a given guess at a given time point.

Table 7. The Comparison of Average Scores Between T Arm and P Arm

Treatment arm	Guess	Time 1	Table 7	Time 3	Time 4
T	T	n_{TT1}	n_{TT2}	n_{TT3}	n_{TT4}
	P	n_{TP1}	n_{TP2}	n_{TP3}	n_{TP4}
	DK	n_{TDK1}	n_{TDK2}	n_{TDK3}	n_{TDK4}
P	T	n_{PT1}	n_{PT2}	n_{PT3}	n_{PT4}
	P	n_{PP1}	n_{PP2}	n_{PP3}	n_{PP4}
	DK	n_{PDK1}	n_{PDK2}	n_{PDK3}	n_{PDK4}

Analysis of such data is described in 'Categorical Data Analysis Using The SAS® System' by Maura E. Stokes, Charles S. Davis and Gary G. Koch, 2nd edition (2000) Wiley page 437. Here we use scores (suggested by James et al. (1996)) for the responses. The score is 0 if the guess is correct, 0.5 if the guess is wrong and 1 if response is DK. So, note that the scores are slightly different for the two arms since the correct answer is different. The question of interest is whether the mean scores for two time points differ significantly. We refer to the data in Table 5 once again. Table 8 below gives the results of the analysis. Notice that only one effect is significant. It is the time effect. In other words, the average scores for the two time points differ significantly for T.

Table 8A. Mean Scores by Time and Treatment Arm

Response Functions and Design Matrix						
Sample	**Function Number**	**Response Function**	**Design Matrix**			
			1	**2**	**3**	**4**
P	1	0.62000	1	1	0	1
	2	0.57000	1	1	0	-1
T	1	0.64000	1	-1	1	0
	2	0.45000	1	-1	-1	0

Table 8B. Results of Tests

Analysis of Weighted Least Squares Estimates					
Effect	**Parameter**	**Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**
Intercept	1	0.5700	0.0319	318.61	<.0001
arm	2	0.0250	0.0319	0.61	0.4337
time(arm=T)	3	0.0950	0.0367	6.71	0.0096
time(arm=P)	4	0.0250	0.0341	0.54	0.4634

Note that the change in average score is significant for T but not for P. In other words, there is significant progressive unblinding for the treatment arm but not for the placebo arm.

(c)Polytomous logistic regression: The data consist of response in three categories (T, P and DK) and hence is amenable to a logistic regression model. Various possibilities exist. First consider a separate model for each treatment arm. Here we use the correct guess as the reference category of response and the model for each time point is $ln\left(\frac{\pi_{1i}}{\pi_{11}}\right) = \alpha_i + \beta * time,$ for i = 2, 3 (these being the categories other than the reference category).

Notice the following points: (i) We fit different models to different treatment arms but slope for time is assumed to be the same. (ii)If chance of DK declines over time, the slope β should be negative and significant. (iii) The situation can be different for the two rows in the 2X3 table. Let us consider one illustration.

Table 9. Case of Three Time Points

True group	Guess at time 1			Guess at time 2			Guess at time 3		
	T	P	DK	T	P	DK	T	P	DK
T	41	55	130	51	75	100	61	75	90
P	14	99	106	24	109	86	29	114	76

Treatment T:

Table 10. Result of Treatment T

Response		Point Estimate		Confidence Interval and P-Value for Beta			
					95 %CI		2*1-sided
Model Term/ Category	Type	Parameter	SE	Type	Lower	Upper	P-Value
Intercept for Guess=P	MLE	0.8145	0.261	Asymptotic	0.3029	1.326	1.81E-03
Intercept for Guess=DK	MLE	1.26	0.2576	Asymptotic	0.7548	1.765	1.05E-06
Time	MLE	0.255	0.1137	Asymptotic	-0.4779	-0.03213	0.02493

We see that the regression coefficient for time is -0.255 and p value is 0.025. In other words, according to this model, probabilities of P and DK (relative to probability of T) decline over time. This is evidence of progressive unblinding. Now the same thing can be repeated for arm P. We can introduce separate intercept and slope parameters as well. Of course there is the issue of how good the model fit is. It can be checked using a chi-square test. We see below that the p value is large and there is no concern about fit of the model. (This is done for T. Results for P are similar).

Table 11. The Result of Chi-Square Test

STATISTICS	VALUE	DF	P-VALUE
-2 loglikelihood	3.724	3	0.2929
Likelihood Ratio	288.3	3	1.37E-09

Above results are obtained from Cytel's Software for logistic regression **LogXact**.

(d) Direct comparison of James' BI:

It is possible that instead of comparing other features of distributions one wishes to compare the blinding indices at two time points directly. This can be done using simulation. T and P groups have to be analyzed separately. The basic input will be the 3X3 array of observed counts. These will be converted into proportions. Samples will be generated from this multinomial population with 9 classes. For each sample, the data for the two time points will be generated. Then the BI will be calculated at each time point. This process will be repeated N times. We will then have N pairs of BI values and hence N differences. They will be ranked and then the proportion of cases in which the actual difference observed is exceeded is calculated. It is the p-value for the test of the hypothesis that the BI value has remained the same for two time points. Alternatively, CI based on normality assumption is also possible.

Table 12. Result of Simulated Value of James BI

Treatment arm T		BI at time point 1	BI at time point 2	Difference: BI 2 - BI 1
Mean	True value	0.6203	0.76	0.139
	Simulated value	0.6289	0.7586	0.1383
Standard Deviation	Simulation	0.02133	0.026311	0.0303
Confidence Interval	Simulation	(0.5862,0.6715)	(0.7059, 0.8112)	(0.069, 0.1902)

In the above table, we see that at each time point mean of simulated values of BI is close to the true value. The CI for the difference excludes the value zero. This means that the two indices differ significantly. Index at the later time point is larger. In other words, there is significant improvement in blinding in the arm T. On the other hand the story is different for arm P. Here the CI for difference in BI contains the value zero. So we conclude that the difference between blinding indices is not significant. In other words, in the arm P, there is lack of evidence to support progressive unblinding. Notice that conclusions seem to change with choice of analysis method. This is not uncommon when comparing alternative methods of analysis. This method can be extended to more than two time points without difficulty.

5. EXTENT OF UNBLINDING AND ITS BREAK UP

First we need a measure of the extent of unblinding. A simple measure of current level of unblinding may be the proportion of DK. We can calculate it separately for T and P if desired. Conventionally, unblinding question is asked at the end of a trial. We propose that it should be asked after randomizing subjects to different arms, as soon as the first dose is given. At this point, there is no efficacy and hence unblinding if any is attributable to the trial design. Any reduction from this value, in the proportion of DK may be attributed to the experience of the subject.

REFERENCES

Hughes J R & Krahn D (1985) Blindness and validity of double blind drug trials. J. Clin. Psychopharmacol. Vol.5 pp.138-142.

Stokes, M E , Davis C S and Koch G G (2000) Categorical Data Analysis using SAS system 2nd edition John Wiley.

James KE, Bloch DA, Lee KK, Kraemer HC, Fuller RK (1996). An index for assessing blindness in a multicentre clinical trial: Disulfiram for alcohol cessation – a VA cooperative study. Stat Med. 15:1421–34.

Bang H, Ni L, Davis CE (2004). Assessment of blinding in clinical trials. Control Clin Trials.25: 143–56.

Alan Agresti (2007) An Introduction to Categorical Data Analysis. John Wiley

Xuezheng Sun, Zhao Yang (2008) Generalized McNemar's Test for Homogeneity of the Marginal Distributions Paper 382 SAS Global Forum , Statistics and Data Analysis.

Non-Inferiority and Equivalence Tests in a Sequential Multiple-Assignment Randomized Trial

Palash Ghosh*

Centre for Quantitative Medicine, Duke-National University of Singapore Medical School, Singapore – palash.ghosh@duke-nus.edu.sg

Abstract

In a sequential multiple-assignment randomized trial (SMART), a patient is given a sequence of treatments over the trial period. SMART designs are applicable in various branches of medical science like depression, behavioral science, cancer etc. The treatments/drugs used in a SMART could be from multiple pharmaceutical companies. Generally, pharmaceutical companies are reluctant to do a comparison of different marketed drugs from different companies by a traditional one/two-sided test, because a negative result may damage their business severely. In this context, a general perception is that a SMART design may not be a suitable choice for pharmaceutical companies. In this article, contrary to that general perception, we argue that a SMART design could be profitable for pharmaceutical companies. It may be a win-win situation for all the companies involved in a SMART. We develop the non-inferiority and the equivalence test procedures to compare treatment-sequences in a SMART, to facilitate the uptake of SMART designs in the pharma-sponsored trials. We show that the trade-off between the efficacy and the quality of life (QOL) from a treatment-sequence can be addressed by a non-inferiority/equivalence test. Sample size and power issues have been discussed with supporting simulations. We show the applicability of the developed test procedures in the SMART context using the STAR*D data. This is joint work with Bibhas Chakraborty.

Keywords: SMART, SMART Design, Non-inferiority

Hybrid Filtering Purchase Prediction Modeling: A Case Study of an Online Healthcare Store

Hermawan Adi Budyanto*
University College London, London, United Kingdom -
hermawan.budyanto.15@ucl.ac.uk / hermawan.budyanto@id.streamintelligence.com
Shorful Islam
Stream Intelligence, London, United Kingdom - shorful@streamintelligence.com
Delmiro Fernandez-Reyes
University College London, London, United Kingdom -
delmiro.fernandes-reyes@crick.ac.uk
Agus Nur Hidayat
University College London, London, United Kingdom -
agus.hidayat.15@ucl.ac.uk / agus.hidayat@id.streamintelligence.com

Abstract

Investing in customer experience is paramount to the success of an online marketplace. Many modern web services thus make an intensive use of recommender systems to provide personalised product recommendation to satisfy unique customers' preference. Nevertheless, there are several difficulties in building recommender systems with past purchase data: (1) the data generally do not observe negative samples of items resulting in class imbalance problem (2) the data are voluminous and highly prone to erroneous information. We approach this data mining problem using hybrid filtering models, by combining item-based collaborative and content-based filtering. This study aims at precisely predicting which customer-specific product is most likely to be purchased. The machine learning algorithms used in this study include Extreme Gradient Boosting, Neural Network, Support Vector Machine, and Penalised Logistics Regression. We test these models on two-year historical transaction of an online healthcare retailer, to predict complete sales for repeating purchases. We assess these models' accuracy against two baseline methods; namely popular items and random search. The result shows that Extreme Gradient Boosting algorithm outperforms other models, consistently achieving higher accuracy metric and fastest running time. The only caveat is that we would need to balance between the number of trees created to optimize computation times. In addition, we design R programs which allow for parallel computations. The resulting codes efficiently scale up to work in a large data set processing. In a commercial environment, the results of this study can help to control customer attrition rate by offering them relevant product recommendations, while at the same time increase economic value-add from loyal customers.

Keywords: Data mining, Hybrid filtering, Recommender systems, Machine learning, Transaction history data, Item-based CF, Statistical modelling

1. INTRODUCTION

The existence of a rating system is pertinent to online marketplaces. Either implicitly or explicitly, dozens of customers' past activity is utilized to predict which item is likely to be purchased or so called the recommender systems. Recommender systems studies so far are mainly concentrated between collaborative filtering (CF) using different forms of techniques (Jannach et al., 2010; Sarwar et al., 2000) and content-based model. CF approach is built upon the principle that if several customers share a common interest towards a similar item, this may indicate their preference overlaps strongly. On the other hand, the content-based model matches the items to customers based on as much as of the existing knowledge that is available allowing the extension of such method to apply machine learning (ML) systems.

In this research, we exclusively address the problem of implicit feedback of an UK-based online contact lenses retailer. Implicit feedback is inferred from customer interaction with items, e.g. purchase actions. Therefore, primary characters emerged from such data is that it is very common to have a highly imbalanced class and sparse (Pan, et al., 2008). Two strategies are commonly implemented to deal with implicit feedback; namely: all missing all negatives (AMAN) and all missing all unknowns (AMAU).

For predicting specific items of interest to a customer, we construct a content-based recommendation whose features consisting of both item description, past purchase history, and item-based CF similarity. We use two stages to infuse CF to content based model. The first stage is the construction of CF item-based similarity matrix. If such item is found in a customer, we calculate the prediction score based on weighted average between the top-N most similar items and the customer's past product history. In the second stage, we engineer customer's profile, item information, and customer interaction with items (e.g. past purchase history).

2. EXPERIMENT PROCEDURES AND EVALUATION METHODOLOGY

Since we intend to predict future items that will be likely purchased by customers, we decompose our dataset into training set (2014 data), holdout (January 2015 to April 2015 data), and testing data (the rest of 2015 data). The construction of these data is described in subsection 2.2. The purchases matrix summarises the entire dataset as described in subsection 2.3. In our setup, we treat missing observations as the negative values and took samples at random – as detailed in Subsection 2.4. Lastly, the evaluation metrics are defined in Subsection 2.5.

2.1. Data

We are provided with two years of customers-transactions data over two years, from 2014 to 2015. The complete data includes transactions history, product, customer data – all of which has total size of 8 GB. The original dataset contains two million transactions, 165,000 customers, and 238 products over two years of observation period from 2014 to 2015.

In addition to that, we scrape the client's website to get more information about customer interactions with the company's online sales website. The data is strictly confidential and we have agreed not to release any sensitive information related to customer data.

2.2. Data splitting technique

We are interested to see the model effectiveness in predicting products that will be purchased by repeat purchasers. To reflect this future event from the same type of customers, we build the machine learning models on the 2014 data. We then tune the parameters of the model using a holdout set from dataset taken between January 2015 and April 2015. After obtaining the optimal parameters, we run the expreriment using both 2014 data and hold out data. Finally, we evaluate the model using May 2015 until December 2015 data. The details of training and testing dataset is given in table 1.

Table 1A. Schematic Design for Data Splitting Strategy

Data	Model Tuning		Final Model	Testing
	Training	Holdout		
2014 training	x		x	
2015 holdout		x	x	
2015 validation				x

2.3. Purchase matrix

Our algorithm is designed to work at transaction levels data, as opposed to the standard CF approach for which it works at customer-levels data. Hence, it is possible to see multiple observations that come from a customer if she/he has multiple transactions. We denote a purchase matrix as P, which contain Boolean values of either 0 and 1. The entry of 1 in cell $P(i,j)$ means that an item has been bought in that respective order and 0 simply means the opposite.

2.4. Negative sampling scheme

To introduce the negative samples from missing products, we take randomly products that have not yet been bought by a customer — the missing observations. The implementation of this approach is inexpensive and flexible. Following is the detailed sampling procedure:

Input: a set of past purchase history at transaction-levels

1. Group data of items that were bought by a unique key — consisting of a customer id and order no. This creates a positive sample dataset in the binary classification model.
2. Draw negative samples of n items from an item list that was sold in 2014 as well as the new products sold in 2015
3. Drop observations which contain any products that have been bought from the dataset
4. Append the negative dataset produced in the previous step to the positive dataset.

2.5. Evaluation methodology

2.5.1. Evaluation metrics

To avoid confusion in interpreting the results, in this study, the actual positive target data (the dataset which contains products that were purchased) is recognised as the negative class by R. Thus, in this context specifity is defined as the number of true negative (number of observations where the actual positive data is correctly predicted) divided by the sum of true negative (number of observation where actual negative data is correctly predicted) and false positive (number of actual negative data that is mistakenly predicted). We use specificity and area under receiver operating characteristics (AUROC) to evaluate the quality of a recommendation list generated by our models.

2.5.2. Model evaluation

We compare the results of several machine learning algorithms against baseline models such as popular items and random search. Popular item model will recommend items that were purchased by a majority number of customers in the training dataset. This is particularly fascinating because we observe that the detailed transaction data shown a long tail – meaning that top 5 items responsible for 50% sales volume. Furthermore, random search model selects items at random for the recommendation and thus 50% accuracy should be expected.

3. ALGORITHMS EVALUATED

3.1. Item-based CF

We calculate the pairwise similarity of two items using the cosine distance formula as below:

$$sim(i,j) = \frac{(P_i^T P_j)}{\sqrt{P_i^T P_i}\sqrt{P_j^T P_j}}$$

Where P_i denotes the vector column of item i^{th} and P^T is simply the transpose form of such vector. If the total number of items is n, the similarity matrix would have a dimension of $n \times n$.

3.2. Item-based CF scores prediction

The prediction score of an item *i* for a customer is computed using following approach:

1. For each product *i*, in *P*, select the top-N most similar items from the similarity matrix **S.** Next, we obtain $||P||_i$ or a vector that contains top-N items' names of those top-N similar items to product
2. The top-N item list (*topN.sim*) is sorted by item-similarity values decreasingly. In addition, the respective transaction history of a consumer's pertaining to the top-N items was also extracted (*topN.records*) in the form of Boolean values– either 0 or 1.
3. Calculate the prediction score for product *I* using this formula:

$$\frac{crossproduct(topN.records, topN.sim)}{N * sum(topN.sim)}$$

It is important to note that the selection of N could affect the machine learning accuracy performance. There is not such an exact rule to rely on, we in fact experimenting with several N values. In the end, considering the computational time and model efficiency, we decided to choose $N = 5$ items.

The nested loop process for computing cosine similarity distance is considerably slow. One reason is that the computation is done in a sequential process from the outer loop through the inner loop. We derive an optimal computation algorithm so that it enables backend core systems to run simultaneously; for example: 1 to 1000 customers on core 1; 1000 to 2000 customers on core 2, etc. By parallelising available loops, we can significantly speed up the entire process and get a seriously improved performance. The detailed implementation of parallelised item-based similarity and score prediction in R can be found in[1].

3.3. Machine Learning models

In this research, we experiment several models including penalised logistics regression (penalised logistics) (Friedman et al., 2000), support vector machine – using kernels from both radial and polynomial degree (SVM Radial and Poly) (Karatzoglou et al., 2004), neural networks (NN) (Bergmeir and Benitez Sanchez, 2012), and extreme gradient boosting (Xgboost) (Chen and Guestrin, 2016).

3.4. The development environment

We use R version 3.3.0 in Amazon EC2 R3x2 which has following specification: 61GB Memory,8 vCPU, and the physical processor from Intel Xeon E5-26712. Table 2 shows R open source packages that we use to run the machine learning (ML) models.

4. EXPERIMENTAL RESULTS

Table 2. The Corresponding R Packages Used to Run The Machine Learning Model

ML Models	The corresponding R package
Xgboost	XGBoost
NN	nnet
SVM	kernlab
Penalised Logistics	glmnet or caret

4.1. Parallelization results in R

We calculate the item-similarity matrix using different scenarios of number of customers. The result in table 3 confirms the parallelised loop can be 3.5 more optimal than the traditional loop. The reason for

1 https://github.com/hbudyanto/masterthesis/blob/master/singlescr/func.R

such inefficiency observed for smaller sample dataset is that it takes some time for R to send the request to enable backend cores to work in parallel. Thus, when the number goes small, running in parallel will likely to overkill the entire performance.

Table 3. The Completion Time of Running Cosine Similarity (in seconds)

Samples taken (customers)	10	100	1000	10,000
Traditional	0.033	0.384	9.880	208.11
Parallel	0.171	0.968	8.346	63.84

4.2. Evaluation of predicted repeat purchasers model

Due to massive dataset size, we decide to take a random sample of repeat purchasers (N). Here, we define the repeat purchasers as the customers who at least have one transaction in each observation year - 2014 and 2015. In doing so, we were able to relate the past history with the future transaction.

Table 4. Summary of Accuracy Metrics on Returning Customers Using Different Sampling Numbers

Algorithms	N = 10,000		N= 25,000		N=50,000	
	AUC	Spec	AUC	Spec	AUC	Spec
XgBoost	89.0%	71.9%	91.0%	73.4%	89.9%	72.9%
NN	89.2%	68.8%	89.7%	69.9%	89.1%	68.0%
SVM Radial	88.9%	68.6%	88.4%	69.3%	87.1%	67.8%
SVM Poly	89.5%	68.7%	89.6%	68.6%	88.9%	67.4%
Penalised Logistics	88.9%	63.4%	89.7%	64.1%	88.7%	63.3%
Popular Items	80.3%	67.5%	84%	68.3%	89.0%	69.0%
Random Search	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%

Table 4 shows the performance of the models in the testing set using the setting that we configured from Section 2.2. Across all sampling scheme, the extreme gradient boosting outperforms other models - achieving in average 72% in specificity metric. Meanwhile, the other ML algorithms except penalised logistics regression model reported about the same accuracy - 69% or 1pt higher than popular items baseline model. One possible explanation for a small gap between the four ML algorithm and popular item is that the historical item sales data was mainly fragmented - top 5 items account for 50% of total sales. Thus, it is not very surprising to see that popular items can be quite powerful.

In contrast, the penalised logistics regression performs poorly against the popular item baseline model. We noticed the there is an indication of overfitting in the model performance. One solution to counter this is to create more options in grid search to find the most optimal penalisation constant in the model which may lead to a longer computation time.

Table 5 presents the comparison of running time amongst machine learning algorithms used in the experiment. Again, the Xgboost comes top – recording a mere 7 minutes long to run hyperparameters optimization search while the SVM model take the longest time. Figure 1 shows the ROC curve of all ML models when they are tested on the testing dataset. Again, the Xgboost achieves the highest AUC score - with c.91% or 1pt higher than the neural nets in the second place.

Table 5. Running Time (in minutes) Comparison Between ML Algorithms Used in The Experiment (N = 25,000)

Algorithms	Number of Grid Search	Average running time per search
XgBoost	30	7.25
NN	24	9.56
SVM Radial	5	15.78
SVM Poly	20	14.71
Penalised Logistics	3	6.9

Considering all aspect of computational time and accuracy performance, these results suggest Xgboost is the best method on all measures. However, note that a careful approach should be taken when growing trees in this model. Too many trees could lead to overfitting, whereas too little trees may only result in a small increment in accuracy metric.

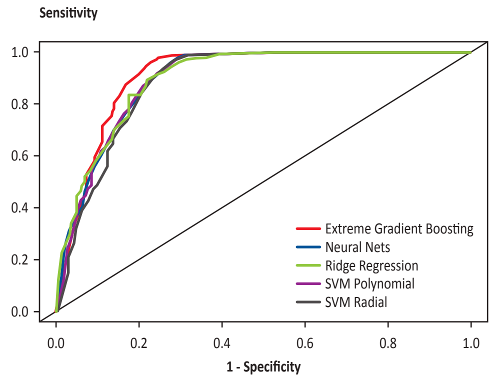


Figure 1. ROC of All Machine Learning Models

4.3. Feature Importance

We experiment item-based CF score prediction to be applied in a range of item profiles. This includes: product, manufacturer, lens type, brand, category, and subcategory. As we notice from Table 6, three out of five most important predictors are reported come from item-based CF score prediction.

Table 6. Feature Selection using Random Forest Algorithm - to display top 5

Features	Score
Item-based CF score prediction on product	100
Customers' past purchase history of lens type and brand type	35
Customers' past records of subcategory and brand type	25
Item-based CF score prediction on manufacturer	6
Item-based CF score prediction on category	5

5. CONCLUSIONS

This study exemplifies that purchase history can be used to develop recommender systems. In such cases, we explore solution using the combination of item-based CF and content-based model. In particular, we develop a binary classification models where we obtain the negative samples by treating missing variables as negative target dataset.

Our result shows that the construction of hybrid filtering content model using features from item-based CF was proved to perform better than baseline model (extreme gradient boosting vs popular items model) in predicting items that will be purchased by repeat customers. Using random forest algorithm, we see that three item-based CF features sit on top-5 ranking of most important features.

One evident application of this model is to deliver a holistic customer relationship management (CRM) program. Thus, once the company has implemented customer segmentation and churn model, it could continue further by offering a tailored deal to retain high churn and high value customers (e.g. send discount promotion) as well as to improve values of low churn but high value customers (e.g. offer up sales or cross sales). Another benefit of having recommendation systems is that it can be used by purchasing division to estimate number of item orders. Moreover, using the same recommender systems, the marketing division will have the capacity to estimate new items that has high likelihood to perform well and later can efficiently utilize marketing budget to advertise these products.

REFERENCES

Amatriain, X., Pujol, J. M., & Oliver, N. (2009). I like it... i like it not: Evaluating user ratings noise in recommender systems. *International Conference on User Modeling, Adaptation, and Personalization.*

Bergmeir, C. N., & Benitez Sanchez, J. M. (2012). Neural networks in R using the Stuttgart neural network simulator: RSNNS.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754.*

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics, 28*(2), 337-407.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction.* Cambridge: Cambridge University Press.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). *Kernlab-an S4 package for kernel methods in R.* Vienna: Institut fur Statistik und Mathematik, WU Vienna University of Economics and Business.

Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., & Yang, Q. (2008). Digitizing the consumer decision journey. *2008 Eighth IEEE International Conference on Data Mining.*

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *Proceedings of the 2nd ACM conference on Electronic commerce.*

This page intentionally left blank

IPS21:
Goodness of Fit and Change Point Problems

Bootstrap Parameter Change Test for Location-Scale Time Series Models with Heteroscedasticity

Sangyeol Lee*
Seoul national University, Seoul, Korea– sylee@stats.snu.ac.kr

Abstract

This study considers the bootstrap CUSUM test for a parameter change in location-scale time series models with heteroscedasticity. The CUSUM test has been popular for detecting an abrupt change in time series models because it performs well in many applications. However, it has severe size distortions in many situations. As a remedy, we consider the bootstrap CUSUM test, particularly focusing on the score vector based-CUSUM test, and demonstrate the weak consistency of the bootstrap test for its justification.

Keywords: Location-scale time series models with heteroscedasticity; Parameter change test; CUSUM test; Bootstrap method

An Adaptive-to-Model Test for Parametric Single-Index Errors-in-Variables Models

Hira Koul
Michigan State University, East Lansing,USA – koul@stt.msu.edu
Chuanlong Xie
Hong Kong Baptist University, Hong Kong, China – 14485435@life.hkbu.edu.hk
Lixing Zhu*
Hong Kong Baptist University, Hong Kong, China – lzhu@hkbu.edu.hk

Abstract

This paper provides some useful tests for fitting a parametric single-index regression model when covariates are measured with error and validation data is available. We propose two tests whose consistency rates do not depend on the dimension of the covariate vector when an adaptive-to-model strategy is applied. One of these tests has a bias term that becomes arbitrarily large with increasing sample size but its asymptotic variance is smaller, and the other is asymptotically unbiased with larger asymptotic variance. Compared with the existing local smoothing tests, the new tests behave like a classical local smoothing test with only one covariate, and still are omnibus against general alternatives. This avoids the difficulty associated with the curse of dimensionality. Further, a systematic study is conducted to give an insight on the effect of the values of the ratio between the sample size and the size of validation data on the asymptotic behavior of these tests. Simulations are conducted to examine the performance in several finite sample scenarios.

Keywords: Adaptive-to-model test; Dimension reduction; Errors-in-variables model.

A Data-Dependent Choice of the Tuning Parameter for Goodness-of- Fit Tests Employing Bootstrapped Critical Values

L. Santana
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa
W.D. Schutte
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa
& J. S. Allison
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

Abstract

Allison & Santana (2015) discussed a data-dependent choice of a tuning parameter that appears in many goodness-of-fit tests. However, this method is only applicable to a class of tests for which the null distribution is independent of unknown parameters, but can be approximated through simple Monte Carlo methods. This data-dependent choice of the tuning parameter is obtained by maximising the bootstrap power of a test when calculated from the Monte Carlo approximation of the appropriate critical value. We now consider a slightly broader class of tests, that is, where the null distribution is unknown and cannot be approximated via Monte Carlo methods. Unlike the approach followed in Allison & Santana (2015) the approach followed for these tests is to approximate the null distribution (and resulting critical values) using the bootstrap. Typical tests that fall in this class include testing for symmetry, and testing goodness-of-fit for the gamma, generalised exponential, skewed normal, and normal mixture distributions, to name but a few. The new method to obtain the data-dependent choice of the tuning parameter in these tests does not rely on maximising bootstrap power, but rather relies on attempting to find the value of the tuning parameter that allows the test to come as close as possible to the specified nominal significance level. An iterative bootstrap algorithm which employs the Monte Carlo *warp speed* method of Giacomini, Politis and White (2013) is provided and the results of the performance of our new method is investigated in two testing scenarios: testing symmetry of the error distribution of a regression model, and testing goodness-of-fit for the gamma distribution. Various test statistics containing a tuning parameter are employed for these scenarios.

Keywords. Bootstrap; Data-dependent choice of tuning parameter; Empirical characteristic function; Goodness-of-fit; Monte Carlo; Symmetric distribution.

1. INTRODUCTION

Many of the tests that appear in the goodness-of-fit literature are based on, among others, the empirical characteristic function, the empirical distribution function, or the Laplace transform. (see, for example, Epps and Pulley, 1983; Henze and Meintanis, 2002; Baringhaus and Henze, 1991; and Meintanis, et al., 2014). However, the majority of these tests suffer from having to select the appropriate value of a so-called "tuning parameter" that often appears in a weight function in these tests statistics. These parameters hamper the practical application of these tests, since these values are typically dependent on the properties of the unknown distribution of the underlying data. For example, one can see from the power tables generated in Henze & Meintanis (2002, 2005) for testing for exponentiality, the behaviour of the approximated powers across the sequence of values of the tuning parameter fluctuates wildly depending on which alternative is used in the simulation. One therefore requires a method to obtain the value of the tuning parameter that allows the test to perform well, irrespective of one's knowledge of the distribution of the underlying data (which is unknown in practice).

An immediate solution proposed in many papers is to select the parameter value by making use of a "compromise choice" of the parameter that produces the highest power for the majority of alternatives when considering the tables of simulated powers presented in a particular paper (see, for example, the tables presented in Henze and Meintanis (2002, 2005) and Meintanis, et al. (2014), and the corresponding compromise choices proposed therein). Naturally, this procedure is unique to each application and would require the time-consuming simulation and tabulation of approximated powers for a sequence of parameters values and a wide variety of alternative distributions for any new test. In addition, while these compromise choices fare well for many of the alternatives considered in a simulation study, it is clear that they can also fare extremely poorly for others (Allison & Santana, 2015). The compromise choice is therefore undesirable since the knowledge concerning which alternative originally generated the observed data is therefore required to effectively judge whether any particular test employed in practice will perform well or not. With this in mind, one would prefer to make use of a data-dependent choice of the parameter.

One approach to obtain a data-dependent choice of the tuning parameter is discussed in Allison & Santana (2015). However, this data-dependent choice can only be used in the class of problems where the null distribution of the test statistic is independent of unknown parameters, and can be approximated through simple Monte Carlo methods. Many goodness-of-fit tests in the literature satisfy this condition (see, for example, Epps and Pulley, 1983; Henze and Meintanis, 2002; Henze and Meintanis, 2005; Jammalamadaka and Taufer , 2006, and Meintanis, et al., 2014). The data-dependent procedure presented in Allison & Santana (2015) involves selecting the tuning parameter value that maximises the bootstrap power of the test across a sequence of values for the tuning parameter.

In this paper we propose a data-dependent choice of a tuning parameter for a class of tests where the null distribution is dependent on unknown parameters and therefore cannot be obtained through Monte Carlo or exact methods. In these cases, the null distribution must rather be estimated using a bootstrap procedure. Typical tests that fall in this class include testing for symmetry, and testing goodness-of-fit for the gamma, generalised exponential, skewed normal, and normal mixture distributions, to name but a few. The added complication of using the bootstrap estimated critical value means that the approach of maximising bootstrap power followed in Allison and Santana (2015) is not applicable. The method followed now will be to find a value of the tuning parameter that produces a bootstrapped achieved significance level that most closely approximates the specified significance level. The remainder of the paper discusses the formal details of the general approach to obtain the new data-dependent choice of the tuning parameter. The paper concludes with bootstrap algorithms for the specific implementation of the method when applied to testing goodness-of-fit for the Gamma distribution as well as for testing symmetry of the error distribution of a linear model.

2. A DATA-DEPENDENT CHOICE OF THE TUNING PARAMETER

Let $X_n=(X_1,\ldots,X_n)$ be identical and independently distributed random variables from some unknown distribution function, F, with unknown parameter θ and let the empirical distribution function (EDF) constructed from X_n be denoted F_n. For the purposes of this paper we will consider the following general hypothesis:

H_0: $F\in\mathfrak{F}_\Theta$ for some $\theta\in\Theta$,

where $\mathfrak{F}_\Theta=\{F_\theta:\theta\in\Theta\}$ denotes a general family of distributions encompassing fully parametric (such as that of the family of exponential distributions) and semi-parametric (such as that of the family of all symmetric distributions) instances, and is indexed by some parameter $\theta\in\Theta$, and where Θ denotes the parameter space assumed to be an open subset of $\mathbb{R}^q$ for some $q \geq 1$, Let the test statistic to be used in these hypotheses be denoted by $T_{n,a}(X_n)$, and let a denote the tuning parameter. In the discussions and algorithms that follow, we will assume that $a\in\{a_1,a_2,\ldots,a_S\}$.

If we denote the specified significance level by and if we choose "correctly", then it stands to reason that

$$P(T_{n,a}(\boldsymbol{X}_n)\geq C_{n,a}(\alpha,\boldsymbol{X}_n) \mid H_0)\approx\alpha, \tag{1}$$

where $C_{n,a}(\alpha,\boldsymbol{X}_n)$ is a bootstrap estimate of the critical value. Note that the quantity in (1) is typically the measure one would use to evaluate the performance of a bootstrap critical value by means of Monte Carlo simulation. To define $C_{n,a}(\alpha,\boldsymbol{X}_n)$, first consider the transformed values $\boldsymbol{W}_n^0=(W_1^0,W_2^0,\ldots W_n^0)$, with $\boldsymbol{W}_i^0\equiv \boldsymbol{W}_i^0(\boldsymbol{X}_n)$, i=1,2,…,n, and where the transformation is defined such that G_n, the EDF of $\boldsymbol{W}_n^0$, reflects the null hypothesis. Next, obtain the *bootstrap (re)sample* $\boldsymbol{W}_n^{0*}=(W_1^{0*},W_2^{0*},\ldots W_n^{0*})$ by sampling independently from G_n, and calculate the statistic $T_{n,a}(\boldsymbol{W}_n^{0*})$. The value $C_{n,a}(\alpha,\boldsymbol{X}_n)$ is then obtained such that it satisfies the expression

$$P^* \left(T_{n,a} \left(\boldsymbol{W}_n^{0*}\right) \geq C_{n,a} \left(\alpha, \boldsymbol{X}_n\right) = \alpha, \right. \tag{2}$$

where P^* denotes the conditional probability operator given $\boldsymbol{X}n$. For practical computer simulations, this value can be obtained by independently generating B bootstrap samples $\boldsymbol{W}_n^{0*}$, calculating the statistic $T_{n,a}\,(\boldsymbol{W}_n^{0*})$ for each one, sorting the results, and then determining the value that appears at the B×(1-α) position.

Naturally, since the value of a affects the test statistic value and the resulting value of $C_{n,a}\,(\alpha,\boldsymbol{X}_n)$, one would like to evaluate each choice of a to determine whether the expression in (1) is true or not for the given choice of a. In order to perform this evaluation using Monte Carlo methods, one would need full knowledge of the distribution stated in the null hypothesis. In the scenarios presented in this paper we only address the class of hypotheses where the distribution is not fully specified, and so this form of Monte Carlo evaluation is not possible. However, since one cannot perform this evaluation process by drawing samples from F, the evaluation can be shifted to make use of G_n instead, thereby evaluating the performance of a not using "real" data, but rather using "bootstrap" data and an iterated bootstrap algorithm. The resulting optimal value of a that is found using the bootstrap evaluation, denoted $\hat{a}$, is then our data-dependent estimator for the tuning parameter.

The general procedure to obtain $\hat{a}$ is described next. Given $\boldsymbol{X}_n$, transform the values to reflect the null hypothesis, i.e., $W_i^0 = W_i^0\,(\boldsymbol{X}_n)$, i=1,2,…,$n$, (once again, denote the EDF of these data by G_n) and sample with replacement from these values to obtain $W_1^{0*},W_2^{0*},\ldots,W_n^{0*}$. Treating the EDF G_n as if it was the distribution specified in the null hypothesis (but is now fully specified), the procedure to obtain evaluate the values will involve applying the bootstrap estimation of the critical value to the data $W_1^{0*},W_2^{0*},\ldots,W_n^{0*}$, as opposed to the original data $X_1,X_2,\ldots,X_n$, and thus requires a second level of bootstrap sampling. Following the same approach as before, but now applied to $W_1^{0*},W_2^{0*},\ldots,W_n^{0*}$, we start by transforming the data to reflect the null hypothesis, i.e., $V_i^{0*} \equiv V_i^{0*}\,(\boldsymbol{W}_n^{0*})$, i=1,2,…,n. Next, we sample with replacement from these transformed values to get $\boldsymbol{V}_n^{0**} = (V_1^{0**},V_2^{0**},\ldots,V_n^{0**})$. Finally, the evaluation of each of the values of in $\{a_1,a_2,\ldots,a_S\}$ involves first calculating the quantity

$$\hat{\alpha}_a^* := P^* \left(T_a \left(\boldsymbol{W}_n^{0*}\right) \geq C_a \left(\alpha; \boldsymbol{W}_n^{0*}\right)\right), \tag{3}$$

where $C_a\,(\alpha;\boldsymbol{W}^{0*})$ is that value such that

$$P^{**} \left(T_a \left(V_n^{0**}\right) \geq C_a \left(\alpha; W_n^{0*}\right)\right) \approx \alpha,$$

and where P^{**} denotes the conditional probability operator given $\boldsymbol{W}_n^{0*}$. Having performed the calculation in (3) for each a in $\{a_1,a_2,\ldots,a_S\}$, the sequence $\{\hat{\alpha}_{a_1}^*,\hat{\alpha}_{a_2}^*,\ldots,\alpha_{a_S}^*\}$ is obtained, and the problem now simply reduces to finding the value of a that has a corresponding $\hat{\alpha}_a^*$ value closest to the specified significance level a. That is, the data-dependent choice of the tuning parameter is calculated as:

$$\hat{a} := \underset{a \in \{a_1,a_2,\ldots,a_S\}}{\arg\min} |\hat{\alpha}_a^* - \alpha|. \tag{4}$$

Note that the calculation of (3) will require an iterated bootstrap. To improve computational time, the *warp speed* method of Giacomini, et al. (2013) will be is incorporated to perform these calculations.

3. BOOTSTRAP ALGORITHMS

The data-dependent choice of the tuning parameter described in the previous section will now be applied to two testing scenarios: The first scenario involves a goodness-of-fit test for the Gamma distribution and the second concerns a testing goodness-of-fit for the symmetry of the error distribution in a linear regression model.

3.1. Testing goodness-of-fit for the Gamma distribution

The bootstrap algorithm to obtain the data dependent choice of the tuning parameter a appearing in a test statistic of the form $T_{n,a}$ is now presented when one tests the hypothesis

H_0: the law of X is $\Gamma(\vartheta,c)$ for some $\vartheta,c>0$,

i.e., X follows a Gamma distribution with parameters ϑ and $c>0$. To illustrate how we go about choosing the tuning parameter data dependently, we make use of the test statistic in Henze, et al., (2012). For their test they made use of the *parametric bootstrap* to calculate the critical value. We thus also employ the *parametric bootstrap* in our algorithm, given below:

1. Given the sample $X_1,X_2,\ldots,X_n$, estimate the Gamma distribution parameters $\theta=(\vartheta,c)$ using some suitable estimator, $\hat{\theta}=(\hat{\vartheta},\hat{c})$, where $\hat{c}$ is a scale invariant estimator for c. Next, fix a set of tuning parameter values $\{a_1,a_2,\ldots,a_S\}$.
2. For a chosen value, $s\in\{1,2,\ldots,S\}$ proceed as follows.
   a. Obtain the parametric bootstrap sample $\boldsymbol{W}^{0*}=(W_1^{0*},W_2^{0*},\ldots,W_n^{0*})$ by sampling n observations independently from $\Gamma(\hat{\vartheta},1)$.
   b. Estimate the Gamma parameters ϑ and c again from the data $W_1^{0*},W_2^{0*},\ldots,W_n^{0*}$ and name them $\hat{\vartheta}^*$ and $\hat{c}^*$. Next, scale the W_i^{0*} values as follows $Y_i^{0*}=W_i^{0*}/\hat{c}^*$, $i=1,2,\ldots,n$.
   c. Calculate the test statistic $T_{n,a_s}^* = T_{a_s}(Y_1^{0*},Y_2^{0*},\ldots,Y_n^{0*})$.
      i. Obtain a double bootstrap sample $\boldsymbol{V}^{0**}=(V_1^{0**},V_2^{0**},\ldots,V_n^{0**})$ by sampling n observations independently from $\Gamma(\hat{\vartheta}^*,1)$.
      ii. Once again, obtain the estimated parameters from $V_1^{0**},V_2^{0**},\ldots,V_n^{0**}$, now denoted $\hat{\vartheta}^{**}$ and $\hat{c}^{**}$ and scale the values using $Y_i^{0**}=V_i^{0**}/\hat{c}^{**}$, $i=1,2,\ldots,n$.
      iii. Calculate the statistic $T_{n,a_s}^{**} = T_{a_s}(Y_1^{0**},Y_2^{0**},\ldots,Y_n^{0**})$.
   d. Repeat step 2(a), (b), and (c) B times so that we now have $T_{n,a_s,1}^*,T_{n,a_s,2}^*,\ldots,T_{n,a_s,B}^*$ and $T_{n,a_s,1}^{**},T_{n,a_s,2}^{**},\ldots,T_{n,a_s,B}^{**}$. Ordering the "double bootstrap" values we have $T_{n,a_s,(1)}^{**}\leq T_{n,a_s,(2)}^{**}\leq\cdots\leq T_{n,a_s,(B)}^{**}$.
   e. Approximate the double bootstrap critical value: $C_{a_s,B}(\alpha;\boldsymbol{W}^{0*}) = T_{n,a_s,(\lfloor(1-\alpha)B\rfloor)}^{**}$.
   f. Calculate $\Delta(a_s) := \left|\frac{1}{B}\sum_{b=1}^{B} I\left(T_{n,a_s,b}^* \geq C_{a_s,B}(\alpha;\boldsymbol{W}^{0*})\right) - \alpha\right|$
3. Repeat step (2) for each $s\in\{1,2,\ldots,S\}$ to obtain $\Delta(a_1),\Delta(a_2),\ldots,\Delta(a_S)$.
4. Calculate $\hat{a}_B := \underset{a\in\{a_1,a_2,\ldots,a_S\}}{\arg\min}\,\Delta(a)$.

3.2. Testing symmetry of the error distribution in regression models

The algorithm presented here details the implementation of the data-dependent choice of the tuning parameter for tests for the symmetry of the error terms in a linear regression model. We are thus interested in the hypothesis

$H_0: F(x) = 1 - F(-x)$, for all $x\geq 0$,

where $F(x) = P(\varepsilon_i<x)$, $i=1,2,\ldots,n$, and ε_i is the error term of the regression model

$Y_i = m(X_i) + \varepsilon_i$,

and is some known or unknown function of the observed regressors X_i. For simplicity we will assume in the remainder of the discussion that this function will be of the form of a simple linear regression, i.e., $m(X_i)=\beta_0+\beta_1 X_i$. The *wild bootstrap* procedure followed by Husková and Meintanis (2012) is adapted here to illustrate the data-dependent choice of the tuning parameter appearing in test statistics of the form $T_{n,a}$. In our study we will consider three tests that contain a tuning parameter to investigate the performance of our procedure for selecting the tuning parameter data-dependently, namely

- a test based on the probability weighted empirical characteristic function (Meintanis et al., 2016),
- a test based on the characteristic function with weight function $e^{-a|t|}$ and one with weight function e^{-at^2} (see Hušková and Meintanis, 2012).
- A test based on a characterisation of symmetry by Ahsanullah (1992) and developed in Allison & Pretorius (2017).

1. Given the sample of pairs $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, estimate the vector $\boldsymbol{\beta}=(\beta_0,\beta_1)$ by $\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1)$ using ordinary least squares. Then obtain the sample residuals $e_i=Y_i-(\hat{\beta}_0-(\hat{\beta}_1 X_i$, $i=1,2,\ldots,n$ and fix a set of tuning parameter values $\{a_1,a_2,\ldots,a_S\}$.
2. For a chosen value, $s\in\{1,2,\ldots,S\}$ proceed as follows.
   a. Following a wild bootstrap approach, generate independent random variables $(u_1,u_2,\ldots,u_n)$, with $P(u_i=1)=P(u_i=-1)=0.5$.
   b. Generate bootstrap response values that reflect the null hypothesis of symmetry by calculating $W_i^{0*}=\hat{\beta}_0+\hat{\beta}_1 X_i+u_i\, e_i$
   c. Obtain the ordinary least squares estimator of $\boldsymbol{\beta}$ based on the pairs $(X_1,W_1^{0*}),(X_2,W_2^{0*}),\ldots,(X_n,W_n^{0*})$, and call it $\hat{\boldsymbol{\beta}}^*=(\hat{\beta}_0^*,\hat{\beta}_1^*)$.
   d. Calculate the test statistic $T_{n,a_s}^* = T_{a_s}(e_1^*,e_2^*,\ldots,e_n^*)$, where $e_i^*=W_i^{0*}-\hat{\beta}_0^*-(\hat{\beta}_1^* X_i$, $i=1,2,\ldots,n$.
      i. Once again generate independent random variables $(u_1,u_2,\ldots,u_n)$, with $P(u_i=1)=P(u_i=-1)=0.5$ and calculate $V_i^{0**}=\hat{\beta}_0^*+\hat{\beta}_1^* X_i+u_i\, e_i^*$, $i=1,2,\ldots,n$.
      ii. Denote the ordinary least squares estimator of $\boldsymbol{\beta}$ calculated from the pairs $(X_1,V_1^{0**}),(X_2,V_2^{0**}),\ldots,(X_n,V_n^{0**})$ by $\hat{\boldsymbol{\beta}}^{**}=(\hat{\beta}_0^{**},\hat{\beta}_1^{**})$.
      iii. Calculate the test statistic $T_{n,a_s}^{**} = T_{a_s}(e_1^{**},e_2^{**},\ldots,e_n^{**})$, where $e_i^{**}=V_i^{0**}-\hat{\beta}_0^{**}-\hat{\beta}_1^{**}X_i$, $i=1,2,\ldots,n$.

The algorithm then continues in the same way as in the previous algorithm from step (2)(d).

4. FINITE SAMPLE STUDY

We will investigate the finite sample performance of the newly proposed method for choosing the tuning parameter data dependently for each of the two testing scenarios described above.

REFERENCES

Ahsanullah, M. (1992). On some characteristic property of symmetric distributions. Pakistan Journal of Statistics, Series B, 8, 19–22.

Allison, J.S. and Pretorius, C. (2017). A Monte Carlo evaluation of the performance of two new tests for symmetry. Computational Statistics, doi:10.1007/s00180-016-0680-4.

Baringhaus L. and Henze N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. Annals of the Institute of Statistical Mathematics, 43, 197–192.

Epps T.W. and Pulley L.B. (1983). A test for normality based on the empirical characteristic function. Biometrika. 70, 723–726.

Giacomini R., Politis, D. N. and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. Econometric Theory, 29, 567–589.

Henze N. and Meintanis S.G. (2002) Test of fit for exponentiality based on the empirical Laplace transform. Statistics, 36, 147–161.

Henze, N. and Meintanis, S.G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. Metrika. 61, 29–45.

Henze, N, Meintanis, S.G., and Ebner, B. (2012) Goodness-of-Fit Tests for the Gamma distribution based on the empirical Laplace transform. Communications in Statistics—Theory and Methods, 41, 1543–1556.

Hušková, M. and Meintanis S.G. (2012). Tests for symmetric error distribution in linear and nonparametric regression models. Communications in Statistics - Simulation and Computation, 41, 833–851.

Jammalamadaka, S.R. and Taufer, E. (2006). Use of mean residual life in testing departures from exponentiality. Journal of Nonparametric Statistics, 18, 277–292.

Meintanis, S.G, Swanepoel J. and Allison J. (2014). The probability weighted characteristic function and goodness-of-fit testing. Journal of Statistical Planning and Inference, 146, 122–132.

This page intentionally left blank

IPS22:
Financial Account and Balance Sheet

The Challenges in Compilation of National and Regional Balance Sheet

Bagus Dwi Karyawan
Central Bank of Indonesia

Abstract

Preventing financial crisis need monetary and financial data, a set of tools as well as a method that be able to describe economic condition, identify financial risk and imbalances. Likewise, a country that consists of lots of autonomous region, in one hand, could create and encourage economic growth. However, on the other hand, decentralized region in a country with liberated sectors institutions could bring additional complexity and contagion risk in financial system. Financial imbalance in one region could create risk to other regions. Therefore, a formulation needs to be formed in order to identify financial imbalances and risk in every region within a country.

Some indicators such as GDP, Balance of Payments, Fiscal Budget, Flows of Fund have been used to describe financial conditions but mostly at the national level and flows. It could not sufficient to identify the risk. The tools that combine the flows andfa stock would be more comprehensive and defined in capturing the potential risk in the financial system. The Integrated Sector Balance Sheet is a comprehensive tool that can describe the relationship among sector, financial imbalances between institutions, domestic and external sector at the national and regional level.

The challenges arise in compiling balance sheet, particularly when dealing with nonfinancial corporation and household sectors that should be devoted in particular region. Furthermore, there is no reference and theoretical concept of Regional Balance Sheets. Hence, availability of high-quality data in certain sector and region is crucial to formulate the National and Regional Balance Sheets.

Keywords : Financial Account, Balance Sheet, Flows of Fund, Integrated Sector Balance Sheet

The Use of Financial Account & Balance Sheet in Measuring Financial System Vulnerabilities

Arlyana Abubakar[1]

Abstract

The main purpose of this paper is to identify financial imbalances, as reflected through financial inter-linkages among sectors and potential risk contagion in financial system, using National Financial Account and Balance Sheet (NFABS). The NFABS consists of balance sheets of each sector in financial system, namely the central government, the regional government, the central bank, banks, other financial corporations, the non-financial corporations and households, and the external sector (rest of the world). There are 3 main tools used in measuring financial system vulnerabilities, namely sectoral risk profile analysis, network analysis and sensitivity analysis. Sectoral risk profile analysis shows that the financial system in Indonesia remain stable over 3nd quarter of 2016. Further, the network analysis reveals the importance of external funding in Indonesia particularly through non-financial corporation channel. The sensitivity analysis discloses that the effect of exchange rate depreciation shock is relatively small on non-financial corporation due to the high of foreign currency asset, whereas the combined shock lead to the decrease in value of net external position of non-financial corporation. However, the exposure from the banking sector rise significantly as the non-financial corporation has to replace their foreign funding with domestic funding.

Keywords : Financial Account, Balance Sheet, Financial System Vulnerabilities, Financial Imbalances

1 Senior researcher at Macroprudential Policy Department – Bank Indonesia (arlyana@bi.go.id)

National Balance Sheet of Thailand: How To Make It Work?

Yuwawan Rattakul Boonyaleephan*
Bank of Thailand, Bangkok, Thailand – yuwawanr@bot.or.th
Somsachee Siksamat
Bank of Thailand, Bangkok, Thailand – somsachp@bot.or.th

Abstract

The National Balance Sheet (NBS) is a very useful tool for economic and financial stability analysis. It contains comprehensive information which helps identify risks and vulnerabilities of the economy. Its importance is demonstrated by the endorsement of the G-20 Finance Ministers and Central Bank Governors, the Financial Stability Board and the International Monetary Fund (IMF) in dedicating one data gap initiative to sectoral accounts, wherein the NBS forms part. In addition, they also introduce changes in the Data Dissemination Standard requirement. Challenges in compiling the NBS lie in the data source availability, quality, and consistency. The compilation procedure involves five Cs of compilation approach: namely, **Collection, Counterparts, Confrontation, Control-total** and **Collaboration**. The **collection** of data source from various balance sheets of major economic sectors, such as Financial, General Government and Rest of the World sectors, is conducted and supplied by country member to the IMF under Article VIII. From **counterpart** information on the balance sheet of the financial sectors, balance sheets of the non-financial corporate and household sectors can be derived, which is very helpful when the direct information is not available at reasonable costs. The **confrontation** of data from various sources involves data validation and quality check of the input data. Most often the administrative data source is considered a superior data source. The **control-total** helps reduce the inconsistency and put in place the holistic view of each financial instrument. Last but not least, the **collaboration** between compiling agencies is crucial, as the majority of financial balance sheet data source are available on a periodic basis at the central bank as the financial regulator and statistics compiler, while the rest of the national accounts are compiled by the National Economic and Social Development Board office. The clear role and responsibility should be defined and agreed, information should be shared and easily accessible, and the compilation and reconciliation process should be pre-designed between the two agencies. The NBS plays vital role in the Balance Sheet Approach of the International Monetary Fund. The approach utilizes detailed information in the NBS to identify four types of balance-sheet mismatches; maturity, currency, capital structure mismatches and solvency problems. The NBS is also employed in stress-testing and shock propagation to other sectors. In many countries, the NBS serve as a basis for key financial indicators of household and non-financial corporate sectors. Although perfecting the NBS is almost impossible, the 5C steps help tackle major challenges faced by the compilers. Even the partial coverage proves to be useful for policy makers to gauge the economic vulnerabilities.

Keywords: System of National Accounts; Balance Sheet Approach; Data Gap Initiatives; Macroeconomic vulnerability.

JEL Classification: E01; E21; E63

1. INTRODUCTION

The National Balance Sheet (NBS) is a very essential tool for macroeconomic and financial stability analysis. It contains comprehensive information that supports holistic analysis of risk and vulnerability of an economy. The NBS is a macroeconomic statistics based on the System of National Accounts 2008 framework that comprises overall assets and liabilities of a nation categorized by exhaustive lists of assets and other interesting information, such as the owner of the assets and the associated debtor, maturity, monetary and currency dimensions.

The G-20 Finance Ministers and Central Bank Governors (FMCBG) has endorsed the NBS importance by dedicating one data gap initiative (DGI) to sectoral accounts and balance sheet to assess vulnerability of domestic economies to shocks since 2009[1]. Since then, the progress towards implementing full set of accounts of each country has differed markedly. FSB and IMF (2015) concluded that there was a need for improvement in this area from the first phase of DGI implementation (DGI-1) in 2015 and carried it to the 2nd phase (DGI-2). In parallel with the DGI, the IMF has leveled up the Special Data Dissemination Standard (SDDS) to the improved SDDS Plus by incorporating many DGI-I recommendations to the data requirements, including the sectoral financial balance sheet.

Although Thailand is not a G-20 member, the DGI-1 serves well as guidelines and suggestions of new tools for policy analysis. The Bank of Thailand (BOT) finds the guidelines useful and adopts them as the BOT has been striving to apply macro prudential policy and thus in much need for systemic risks statistics. The existing data of the financial accounts (flows only) dated back to 1993 and is based on the annual basis with fourteen month lag. It has been compiled by the National Economic and Social Development Board Office (NESDB), an organisation which is widely recognised as the official compiler of the National Accounts in Thailand. Given the need for policy planning, the BOT started working on compiling quarterly financial balance sheet in 2010 with four month lag in close co-operation with the NESDB, and working towards having the national financial accounts.

2. COMPILATION CHALLENGES

The compilation of the NBS is a daunting task, since it involves a wide range of information of all economic sectors. The NESDB has many studies to support compilation of the NBS since 2007, including Khlaisuan (2011) which had focused on estimating land ownerships and values and found that lands had the largest share of 90% in total assets. However most of the works were focused on compilation of annual data, one-time-off during the period 2005-2009, and involved hiring extra-resources to carry out the project. From that experience as well as ours at the BOT, I have concluded major challenges in compiling the NBS as the following: data source availability, data quality, and data consistency.

The first challenge is **data availability**, which covers existence, coverage and frequency. The most difficult task was how to acquire 20 million household balance sheets, which are not publicly available; would require expensive collection cost; and would expose household privacy thus their unwillingness to share information. Consequently, compiling household balance sheet by census on a quarterly basis was just out of the question. Setting the willingness aside, sampling household survey, did have the problem of representation. The data coverage was also a problem where we collected data based on stratified random sampling. Though we were fortunate to have the balance sheet of 300,000 corporations, they were on an annual basis.

Next challenge is **data quality**. An ideal scenario is when the data is of more granular detail, the easier compilation process will be. Most of the publicly available data had few information, while the detail-probing survey made it difficult to get response. Additionally, the balance sheets of 300,000 corporations have generally come in different account names of sub-categories. Though they were based on the International Accounting Standard, the sub-categories were not compatible with the System of National Accounts (SNA). To make it more difficult, most of the vital information for compilation were in hard copies, not digital information, making us incapable to verify data using summation of account names. Thus corporation balance sheets was almost useless.

The third major challenge is **data consistency**, which is almost akin to the quality. In contrast to unavailable data, occasionally we came across happy problems of having more than one source of data. Most of data from financial institutions did have a counterparty information, varying from low to high details. Not all the dataset were cross-checked with the other. The Monetary and Financial Statistics and the Balance of Payment Statistics at the BoT were compiled separately and selected based on the source which served their respective purposes. Thus we found many of the transactions between the financial

1 The IMF staff and Financial Stability Board secretariat have written a report comprises of 20 recommendations presented to the G-20 FMCBG meeting in 2009, which can be found on the following link: http://www.imf.org/external/np/g20/pdf/102909.pdf

sector and the rest of the world from the MFS or IIP were largely different. The inconsistencies were also found within the balance sheet of the same sector. For example, in the other deposit-taking corporations (ODCs) account the claims on the ODCs were not equal to the ODCs Liabilities from themselves. The most common cause for inconsistency problem lied in the use of different sources, valuation, time of recording, and different classification of data standard.

These are common challenges faced by compilers all over the globe. To overcome these, we have applied a widely accepted methodology to minimise their impact on compilation. There are also other challenges like how to make use of the available data source and the quality improvement of data source that might be specific and need local or international experts to recommend the adjustment.

3. FIVE Cs OF COMPILATION APPROACH

Many countries has overcome common challenges mentioned above and described them in many statistics manuals[2], such as the SNA2008, the Financial Production, Flows and Stocks in the System of National Accounts, and on the website of the compiling agencies. As for Thailand, we have adopted the compilation procedure, so-called **Five Cs of compilation approach**: namely, **Collection, Counterparts, Confrontation, Control-total** and **Collaboration**.

The collections of data source from various balance sheets of major economic sectors, such as Financial, General Government and the Rest of the World sectors, were supplied to the IMF under Article VIII.[3] The Required **collections** of data include the balance sheet based on the MFS for Central Bank, Other Depository Corporations, the Other Financial Corporations; the balance sheet based on the GFS for Central Government, Local government and Social Securities Fund; the IIP statistics for the Rest of the World. With the above information, we can fill in more than half of the NBS. The credit registration and securities database are also favourable additions. The more details of balance sheet and sub-sectors, the easier the compilation will be. In our particular case example, we still lack balance sheet of government sectors except the SSF. Thus we resorted to the next C step of the compilation: Counterpart.

From **counterpart** information on the balance sheet of the financial sector, balance sheets of the non-financial corporate and household sectors can be derived. This is very helpful as the direct information is not available at reasonable collection costs. As mentioned in the first challenge, having NFC and HHs data is suffice for our NBS compilation purposes. The counterpart information, such as the deposits from HH and NFCs and loans from the financial corporations to HH and NFC and so forth, fill in our need while also providing more reliable, timely, and cost-effective sets of data. To address the lack of the general government sector and subsectors balance sheet, we deduced them from the counterpart information, which yield information of government assets and liabilities with the financial sector, so data collection from other government agencies took less effort.

With all data compiled to establish NBS, the next step is Data **Confrontation**. The confrontation of data from various sources involves validating data by checking their consistency and verifying quality of the input. The finding from this exercise, in our case, was the large gap of inconsistency between inter-sectors and intra-sectors. The result obliged us to start working together across compilers within the BOT, establish common ground procedures and adopt the best practice. The compilation process and data sources were adjusted to be in sync. In the meantime, the decision on which data source is reliable and to be used had to be made and was made. As a general rule of thumb, the administrative data source is a superior data source, with recording basis close to the SNA as an extra criterion.

The **control-total** helps reduce the inconsistency and put in place the holistic view of financial instruments and economic sectors. This is the measure to assess if the adjustments improve the data quality and serve as a quality control. In Thailand those measures applied for horizontal controls are data on deposits from banking sector, debt securities, equities in the stock exchange, and loans from a financial sector. As for vertical controls, financial sector and subsectors are used, and help complete the remaining information needed.

---

2 See European Central Bank (2008), International Monetary Fund (2016), United Nations, et.al. (2009) and United Nations, et.al. (2015).

3 Under Article VIII section 5, member countries are required to submit the international investment positions (mandatory) along with other information as arranged with the IMF. See International Monetary Fund (2016a).

Last but not least, the collaboration between compiling agencies is crucial. The BOT could help compiling quarterly financial balance sheets from each account view on request to support the macroprudential and monetary policy planning. The current goal of the NESDB is to have the sectoral breakdown of the current accounts and annual non-financial balance sheet, within five-year timeframe. There are a lot of action items needed to be accomplished by each agency. The NESDB and the BoT could discuss and agree to collaborate to complete the National Accounts. The BoT would supply data from the financial balance sheet and financial accounts. Prior to the full National Accounts being officially released, agreements on clear role and responsibility, arrangement of information sharing, and the compilation and reconciliation process should be reached between the two.

4. USES OF NBS

A good case study to demonstrate how NBS could provide early warning prior to the total collapse of the economy was the 1998 Asian Financial Crisis. Thailand was the trigger of the Asian Financial turmoil in 1998 with her ongoing deficit of current account and soaring real estate prices. Contrastingly, the economy's balance of payments was performing well with a surplus due to the fixed exchange rate regime. This led to the attack on the currency by the global investors to profit from the country fixed exchange rate to the boiling point when the financial regulator gave up the defense of fixed exchange and surrendered to the market force. This brought about the real estate bubble burst, the capital flow reversal, and the exchange rate depreciation of more than 80%. All sectors suffered the dramatic decline; household and non-financial corporations alike. At the point of defending the fixed exchange rate regime, the economy was already on the last line of defense. This could have been prevented with the NBS implementation akin to setting observation posts to strengthen the army intelligence long before the enemy approaches the firing line.

The periodic compilation of NBS data would have indicated the maturity profiling of financial corporations and Non-Financial corporations (NFCs) and the high risks of aggregate exposure to **maturity** and **capital structure mismatches**. Those corporations were highly leveraged with their short-term debts to finance long-term loans. In addition, the NBS data could have given the financial regulator early warning signals of sector exposure to **currency mismatches**, when Non-Financial corporations (NFCs) had financed their real estate projects with their low-cost foreign currency borrowing due to the fixed exchange rate. The real estates were non-tradeable goods with long-term yield that would not have generated adequate foreign currency revenue to finance debts. When the fixed exchange rate regime was replaced with the float regime, this consequently led to **solvency problem** of the sectors**.** The high mismatches in balance sheet of those sectors led to the vulnerabilities and solvency risks which were triggered by the spike depreciation in domestic currency and capital flow reversal.

There was a silver lining in every crisis and this financial crisis brought just that to Thailand. It brought to the attention of the BoT the realisation, hindsight and lessons learned of the dire need to make available better early warning indicators to monitor foreign currency debt. We have learned that economic flow data would have helped alerting the red flag of the state of the economy; though it would be unable to measure the depth of the impact affecting the health of the economy, nor the duration the economy could sustain before collapse. Another lesson learned out of this crisis was the balance sheet mismatches, and all these inspired work on the BSA of Allen, et al. (2002).

The NBS is also employed in stress-testing and shock propagation to other sectors. By collecting and compiling sectoral balance sheet from each account view, the presentation could be changed to match an analytical need. The simplest form is the intersectoral financial asset and liability matrix. This matrix show financing interactions in the economy, while classification can go as far as compilers can collect[4]. Furthermore, it can answer what financial instruments the creditors could claim on the debtors. The basic summary of shock propagation of Thailand that can be derived from NBS is shown in the table below.

4 Ridgeway (2011) has mentioned that Statistics Canada could go as much as 30 sectors. While in Thailand although we could compile 13 sector balance sheet, but due to limitation of the counterpart information we can create only an 8-by-8 intersectoral matrix.

Table 1. Intersectoral Asset and Liability Matrix as end June 2016

(Billion bath)

Creditor / Debitor	Central Bank	General Government	Other Depository Corporations	Other Financial Corporations	Public Nonfinancial Corporations	Other Nonfinancial Corporations	Hous eholds and NPIS Hs	Rest of the World	Total
Central Bank		919	3,611	690	131	319	996	215	7,081
General Govenment	275		1,448	2,010	80	158	1,087	720	5,779
Other Depository Corpoations	187	1,834		1,837	743	3,845	12,760	3,075	24,282
Other Financial Corporations	0	897	1,195		15	968	7,226	1,836	12,135
Public Nonfinancial Corporations	0	2,111	299	484			291	687	3,873
Other Nonfinancial Corporations	0	543	6,378	2,659			11,919	8,410	29,909
Hous e Holds and NPIS Hs	4	416	10,011	1,796	0	1,951		0	14,179
Rest of the World	6,543	39	1,495	1,620	798	2,556	6		13,058
Total	7,009	6,760	24,438	11,296	1,767	9,796	34,284	14,945	110,296

From this table, the household and NPISH (HHs & NPISHs) sector is the most important creditor of the economy in Thailand with the highest financial claims on other sectors of 34 Tn Baht, while the major debtor is the other non-financial corporations (ONFCs) with liabilities (including equities) of 29 Tn Baht. If the ONFCs suffered from the economic slowdown, HHs & NPISHs and the Rest of the World (ROW) (with asset exposures of 11.9 and 8.4 Tn Baht) would be affected through losses in equity holdings. If the situation get worsened, the creditor such as ODCs and/or ROW would have bad debts, and this would worsen the HHs and NPISHs financial position through equities and deposits holding. (The data of connection from which instruments are not shown in the table above.)

In many countries, like Canada, Australia, and EU countries, the NBS serves as a basis for key financial indicators of household and non-financial corporate sectors, since financial data on these areas are hard to come by from other published macroeconomic statistics. The IMF and the Bank for International Settlement (BIS) used the counterpart technique to render the Balance Sheet and indicators respectively, based on the ODC Balance sheet and the IIP (IMF), or the International Banking Statistics (BIS) to monitor member countries economic risk and performance (See Shrestha et.al (2012) and BIS (2016)). In Thailand, at present the BOT has not published the financial balance sheet, but information of (ODCs) credits to private sector or (Financial Institution) Loans to Households are among key indicators published on a monthly or quarterly basis. The coverage of those indicators are partial, not the entire economy with credit provision. However, it includes the major sectors providing credits and shows trends of demands for credits.

Although perfecting the NBS is almost impossible due to the large household sector and non-financial assets, the 5C steps help tackle major challenges faced by the compilers. The 5C steps help building framework for compiling the NBS, and it could help to overcome many challenges mentioned above, though some remains. The partial coverage proves to be quite useful for policy makers to gauge the economic vulnerabilities; however, there are always rooms for improvement due to the availability of new data sources with the compilation heavily based on assumptions and estimates where data is scarce.

5. CONCLUSIONS

The Bank of Thailand is relatively a new player in the NBS field, but it has high expertise in basic central banking statistics. Adopting NBS approach, the BOT still finds the compilation challenging particularly at these early days, due to the challenges of data availability, quality and consistency. Five Cs compilation approach, namely Collections, Counterpart data, Confrontations, Controlled Total, and Collaboration helped us setup and running. Collections help in the preparation process with the check-list of required data source. Counter parts help where the direct financing data cannot be obtained, thus reduce cost, support timely data and control data consistency. Confrontation is not to be taken in its literal sense; it helps in checking the quality and consistency between different sets of central banking statistics, while the administrative data is always a criterion for choosing the better data source. Controlled total helps in setting up the framework to fill in the puzzle pieces. Collaboration among statistics agencies ensure successful compilation for sharing data, dividing roles and responsibilities and setting the compilation and reconciliation best practice procedures. Once the NBS is systemically up and running, regular monitoring and continuous improvement would be next to further improve compilation and data outcome. The use of the NBS is wide-ranging from serving as indicators for each data dimension to the holistic view of the intersectoral linkages in the economy. Though partial data compiled by the IMF and BIS is still useful for policy maker, countries are encouraged to compile first-hand data so that they have a full control of data quality and source.

REFERENCES

Allen, M., Rosenberg, C., Keller, C., Setser, B., and Roubini, N. (2002) A Balance Sheet Approach to Financial Crisis, IMF Working Paper 02/210. Washington, D.C. (https://www.imf.org/external/pubs/ft/wp/2002/wp02210.pdf)

Bank for International Settlement (2016). Long series on total credit and domestic bank credit to the private non-financial sector. Documentation on Data. BIS. Basel. (http://www.bis.org/statistics/totcredit/credpriv_doc.pdf)

Financial Stability Board and International Monetary Fund (2009). The financial crisis and information gaps: Report to the G-20 Finance Ministers and Central Bank Governors (October). Washington, D.C. (http://www.imf.org/external/np/g20/pdf/102909.pdf)

____________ (2015). The financial crisis and information gap: Sixth progress report on the implementation of the G-20 Data Gap Initiatives (September). Washington, D.C. (http://www.imf.org/external/np/g20/pdf/2015/6thprogressrep.pdf)

European Central Bank (2008). Handbook on quarterly financial accounts for the euro area: sources and methods. Frankfurt am Main. (http://www.ecb.europa.eu/stats/pdf/eaa/Handbook_on_quarterly_financial_accounts.pdf)

International Monetary Fund (2016a) Articles of Agreement of the International Monetary Fund. Washington, D.C. (https://www.imf.org/external/pubs/ft/aa/)

____________ (2016b) Monetary and financial Statistics Manual and Compilation Guide. Washington, D.C. (http://www.imf.org/~/media/Files/Data/Guides/mfsmcg_merged-web-pdf.ashx)

Ridgeway, A. (2011) Development and Use of Flow of Funds Data: Canada. Paper presented at the IMF-OECD conference on strengthening sectoral position and flow data in the macroeconomic accounts, 28 February-2March, Washington, D.C. (http://www.imf.org/external/np/seminars/eng/2011/sta/pdf/canada.pdf)

Shrestha, M., S. Fassler, and R. Mink (2011). An integrated framework for financial flows and positions on a from-whom-to-whom basis. Paper presented at the IMF-OECD conference on strengthening sectoral position and flow data in the macroeconomic accounts, 28 February-2March, Washington, D.C. (http://www.imf.org/external/np/seminars/eng/2011/sta/pdf/whom.pdf)

United Nations, European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development and World Bank (2010). System of National Accounts 2008. Sales No.: 08.XVII.29. New York.

United Nations and European Central Bank (2015). Handbook on National Accounting “Financial Production, Flows and Stocks in the System of National Accounts”. Sales No.: 14.XVII.7. New York. (http://unstats.un.org/unsd/publication/Seriesf/seriesF_113e.pdf)

Khlaisuan, Wannapa (2011) “Wealth of Nation from National Accounts’ Perspective”, NESDB Annual Economic Conference. NESDB, Bangkok. (in Thai)

Rich Debt, Poor Debt: Assessing Household Indebtedness and Debt Repayment Capacity[1,2,3]

Muhamad Shukri Abdul Rani
Bank Negara Malaysia, Kuala Lumpur, Malaysia - shukri@bnm.gov.my
Siti Hanifah Borhan Nordin*
Bank Negara Malaysia, Kuala Lumpur, Malaysia - hanee@bnm.gov.my
Chin Ching Lau
Bank Negara Malaysia, Kuala Lumpur, Malaysia - lau@bnm.gov.my
Sheng Ling Lim*
Bank Negara Malaysia, Kuala Lumpur, Malaysia - sheng@bnm.gov.my
Zhen Shing Siow
Bank Negara Malaysia, Kuala Lumpur, Malaysia - siowzs@bnm.gov.my

Abstract

In this study, we explore the relationship between the debt service ratio (DSR) and ability to withstand shocks for individual borrowers in Malaysia, using a micro-level dataset that matches their debt and income. We achieve this by quantifying the financial resilience of individual borrowers and subsequently simulating a model on loan default and credit losses in response to generated financial and economic shocks (debt repayment, cost of living and variable income). The simulation allows us to differentiate the sensitivity of borrowers in the different income segments to each shock and estimate the banking system's exposures to borrowers that are more likely to default. The observations and findings can contribute towards the formulation of more targeted policies to manage household indebtedness in Malaysia. The results show that, in the pre-shock scenario, borrowers across all income groups are more likely to have negative financial margin if their DSR is above 60%. However, for borrowers in the bottom 40th percentile income group, some borrowers with DSR of less than 60% also recorded negative financial margin. In the post-shock scenario, borrowers across all income groups are more likely to have negative financial margin if their DSR is above 40%. On aggregate, borrowers are most sensitive to an income shock, particularly those in the middle income group.

Keywords: Household debt, Household balance sheet, Stress Test, Debt service ratio

JEL classification: E20, E21, E58, G28

1. INTRODUCTION

The household debt-to-GDP ratio has been commonly used to measure the level of household indebtedness. In Malaysia, household debt has expanded annually by 10.7% since 2009 to account for 88.7% of GDP as of 2016Q3. However, a high household debt-to-GDP ratio may not necessarily imply that households are over-indebted or in financial distress (Bilston et al, 2015). The composition of household debt is an important consideration as some debt are acquired for the purpose of accumulating wealth, which over time, can add on to the financial buffer when required (Bank Negara Malaysia, 2015b). While the number of delinquent and impaired loans in Malaysia has remained low relative to total credit extended to the household sector, the attendant credit risk and its implication to financial stability should not be underestimated. Taking this into account, this paper explores the nexus between debt servicing capacity across the income spectrum and how individual borrowers respond to potential economic and financial shocks. To achieve this, this paper leverages on borrowers' income and debt data obtained from a Bank Negara Malaysia's internal database to calculate their financial margin. We then apply various simulated economic and financial shocks to the financial margin of these individuals to

---

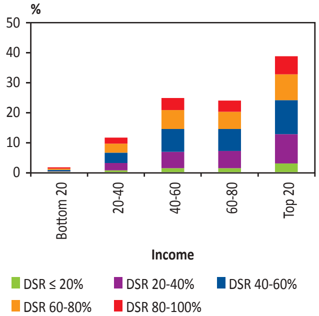
1 The views expressed in this paper should be attributed to the authors and do not necessarily represent those of Bank Negara Malaysia or its policies. The authors would like to thank Madelena Mohamed and Rafidah Mohd Zahari for their valuable contribution to the paper.
2 All authors are from Bank Negara Malaysia, Kuala Lumpur, Malaysia.
3 Refer to http://www.bis.org/publ/bppdf/bispap91k.pdf for full version of the paper

identify the potential deterioration in their debt repayment capacity. We apply the sensitivity analysis model adopted by Bilston, et al. (2015) and Albacete and Fessler (2010). This paper is organised as follow. Section 2 describes the distribution of household debt in Malaysia. Section 3 and Section 4 introduce the model and results respectively. Section 5 discusses the limitations of the model. The final section concludes the paper with a discussion on policy design implications.

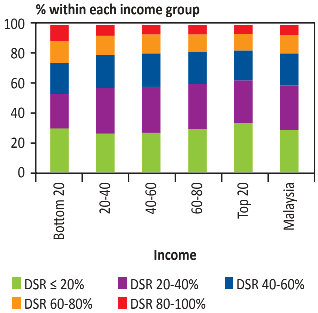
2. STYLISED FACTS

In Malaysia, about 38% of household debt is owed by those in the top 20% of the income distribution (Graphs 1 and 2). The median DSR for this income segment is relatively low at 30% (Graph 3). The debt held by borrowers in the bottom 40% of the income distribution accounted for only 13.2% of total household debt. The average size of debt peaked at RM283,000 for those in the top 20% of the distribution (Graph 4), largely supported by higher income and debt servicing capacity.

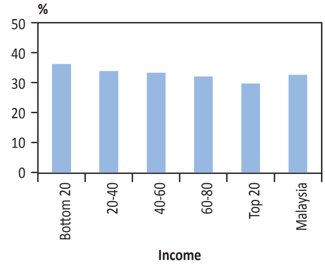
The higher income segment has a larger share of debt for wealth accumulation purposes such as the purchase of residential properties and securities (Graph 5). In contrast, lower income borrowers have most of their loans for consumption purposes (eg purchase of motor vehicles and personal use). About 52% of debt in value terms owed by those in the bottom 40% of the income distribution is based on variable rate, compared to 72% of debt owed by those in the top 20% (Graph 6).



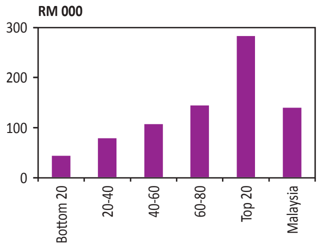
Graph 1. Debt by Income and DSR



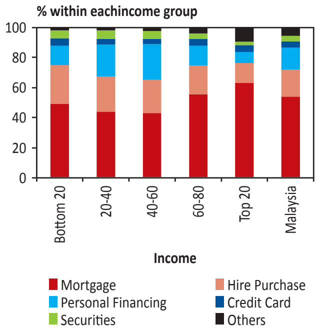
Graph 2. DSR by Income Group



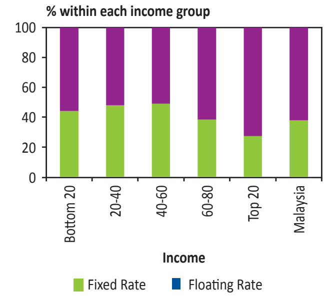
Graph 3. Median DSR



Graph 4. Average Debt per Borrower



Graph 5. Debt by Loan Purpose by Income Group



Graph 6. Debt by Type of Interest Rate by Income Group

3. METHODOLOGY

The main dataset employed in this paper is the Integrated Income Indebtedness Database (IIID) established by Bank Negara Malaysia using the latest available cross-sectional data on taxable income and credit in 2014. The sample used in this article has close to two million individual borrowers and captures approximately RM200 billion or 20% of total household debt in Malaysia. This represents about 5% and 10% of Malaysian population and labour workforce, respectively. We also leveraged on the Household Expenditure Survey (HES) 2014 and Household Income and Basic Amenities Survey (HIBA) 2014 (Department of Statistics Malaysia, 2015a and 2015b) to estimate individuals' expenditure on basic necessities[4]. To ensure consistency with other official publications, the clustering of income groups in this paper is based on the structure of HIBA 2014.

This paper adopts the financial margin approach, similar to the methodology applied by Bilston, et al. (2015) and Albacete and Fessler (2010) to determine the level of financial distress among individuals.

Table 1. Model Assessment Framework

Step	Parameters of interest	Interpretation		
1	Financial Margin	Residual income, sufficient to make ends meet		
2	Probability of Default	Default if financial margin falls below zero		
3	Debt-at-risk (DAR)	Share of net debt (defaulted) to total household debt that lenders cannot recover		
4	Scenarios	Compute Step 1-3 for each scenario		
		Pre-Shock	Individual Shocks	Combine Shocks

Financial margin is the residual income (net of statutory obligations, debt repayment and expenditure on basic necessities) that is sufficient for an individual to make ends meet. An individual is assumed to be in default with certainty if financial margin or residual income is less than zero. We assume that an individual do not use his/her assets, neither savings nor any form of liquid assets to avoid default (Albacete and Fessler, 2010). Total income is derived from the IIID. Each individual's personal disposable income is computed by deducting the estimated income tax and mandatory contribution to the Employee Provident Fund (EPF). Using the granular credit information in the IIID enables us to estimate the monthly debt obligation for every debt facility for each individual. Meanwhile, expenditure patterns for each income group were obtained from the HES 2014.

The debt-at-risk (DAR), which is essentially the expected loan losses to lenders, is the share of net debt from those who defaulted to total household debt. In our analysis, we assume that lenders are only able to recoup 50% of the collateral value for housing assets. Other loans are assumed to have a loss given default of 100%. Individuals' financial resilience is further assessed under different stress scenarios. This paper considers three different financial and macroeconomic shocks – higher cost of borrowing and cost of living[5], and lower income for borrowers. The impact on the financial resilience is derived from the individual shocks as well as combined shocks.

4. RESULTS

Pre-Shock Scenario

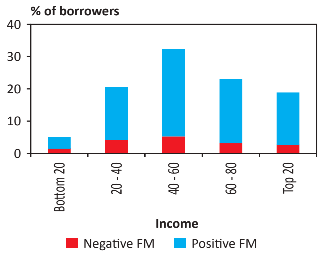
In Malaysia, individuals with negative financial margin (FM) are estimated at 16.6%[6] of total borrowers in the dataset, comprising mainly those from below the 60th percentile income groups (Graph 7) or those with a high DSR of above 60% (Graph 8). The associated debt held by those with negative FM accounted for 32.1% of total household debt, comprising mainly loans for the purchase of residential property. These were acquired mainly by the mid- to high-income earners (Graphs 9 and 10). Generally, borrowers with a DSR less than 60% do not exhibit negative FM (Graph 11). However, for those in the bottom 40th percentile income group, some borrowers with DSR of less than 60% also recorded negative FM. These only accounted for 6.9% of total number of individuals in the bottom 40th percentile income group (Graph 12).

4 Items in the basic necessities include (i) food and non-alcoholic beverages; (ii) housing and water, electricity and gas and other fuels; (iii) transportation; (iv) education; and (v) healthcare (Bank Negara Malaysia, 2015b).

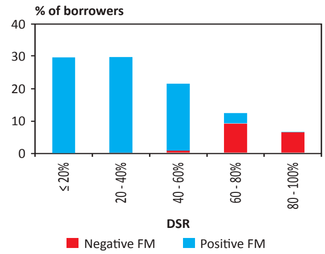
5 This paper assumes that the proportion of expenditure rise at equal measure for all income levels and substitution effects do not take place.

6 Similar with the findings by Albacete and Fessler (2010) that the percentage of households in financial distress in Austria is up to 15.6%.

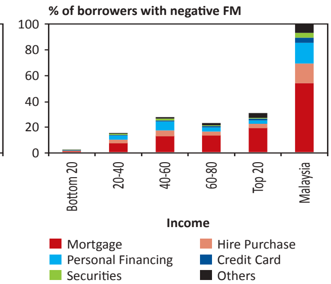
Delinquent[7] and impaired loans[8] in this sample accounted for 0.8% and 0.5% of outstanding loans respectively. About 60% of the delinquent and impaired loans are contributed by those from the bottom 60th income percentile, half of which are from the bottom 40th income percentile. Across all income groups, for individual borrowers with negative FM, delinquent and impaired loan ratios are higher for those with DSR 60% and above (Graphs 13 and 14).



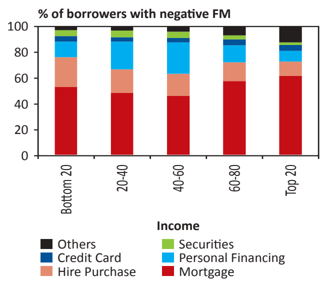
Graph 7. Financial Margin by Income Group



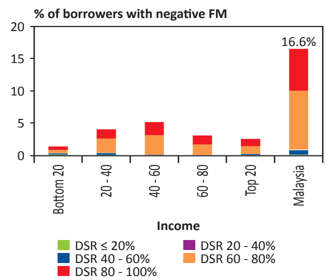
Graph 8. Financial Margin by DSR Group



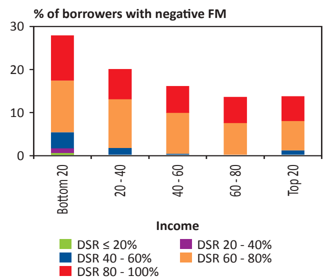
Graph 9. Debt by Loan Purpose by Income Group



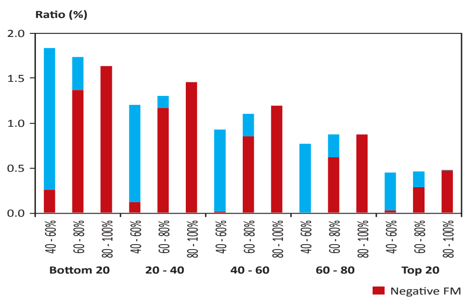
Graph 10. Debt by Loan Purpose within Each Income Group



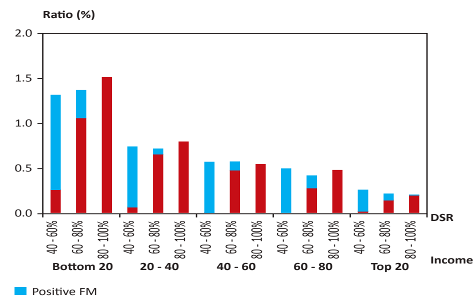
Graph 11. DSR for Borrowers by Income Group



Graph 12. DSR for Borrowers within Each Income Group



Graph 13. Gross Delinquent Loans

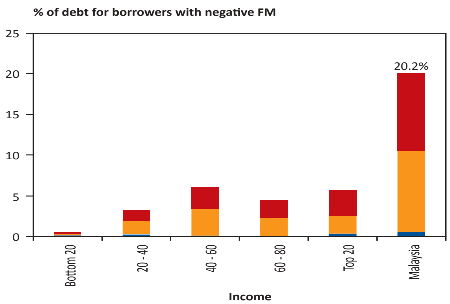


Graph 14. Gross Impaired Loans

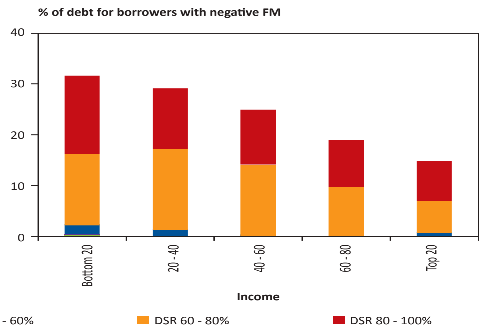
The overall DAR of individuals with negative FM in the pre-shock scenario is estimated at 20.2% of the total debt in the dataset, with a risky net debt of RM43.2 billion. Individuals with the DSR higher than 60% mainly contribute to this. When we break down by income percentile, the DAR is highest for the 40th~60th percentile income category, contributing about 30% to the overall DAR. This is attributed mainly to the higher proportion of non-collateralised debt, which is mainly loans for personal use and car purchase (Graph 10).

7 Loans-in-arrears of between one and three months
8 Loans-in-arrears of three months or more

About half of the DAR is from the 60th income percentile and above. This simply reflects the higher size of debt owed by this group (Graph 1) and does not imply that the credit risk exposures to this group are riskier. Further analysis within each income group shows that borrowers will likely have a lower DAR as earnings increase. The DAR accounts for 18.9% and 14.8% of the DAR in the 60th~80th and top 20th income percentiles, respectively, within the same income categories. In contrast, this is higher at 31.6% for borrowers in the lowest income percentile. All DARs are computed without taking into account other financial assets or savings that can be drawn upon, which could potentially result in a lower DAR.



Graph 15. DAR by Income Group



Graph 16. DAR within Each Income Group

Post-Shock Scenario

Our analysis suggests that, even after combining all shocks based on the worst-case scenarios, the banking system in Malaysia is able to withstand the associated potential losses (Table 2). The amount of risky net debt to the banks, after taking into account the collateral value for housing loans, is estimated at RM77.5 billion, well within the total capital of the banking system in 2014.

Table 2. Scenario Simulation

Scenario Simulation	Financially Distressed Borrowers (PD=1; FM<0)					
	% of Total Number of Borrowers	% of Amount of Debt	Debt-at-Risk (%)	Risky Net Debt (RM billion)		
				Banks	Non-banks	Total
Pre-shock	16.6	32.1	20.2	35.2	8.0	43.2
Post shock: Single						
Cost of Borrowing Shock (150 bps)	19.1	38.1	23.2	41.3	8.6	49.9
Cost of Living (20%)	22.7	41.2	25.9	45.1	10.5	55.6
Income Shock (20%)	34.7	57.7	36.4	63.2	14.8	78.0
Post shock: Combined						
(all of the above)	48.1	74.2	44.3	77.5	17.5	95.0

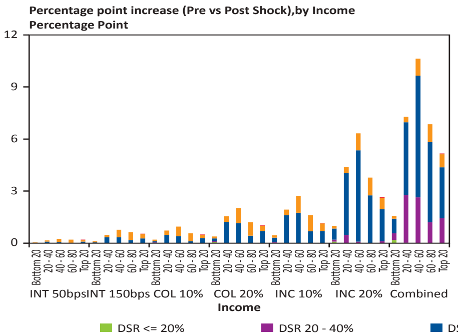
Source: Bank Negara Malaysia

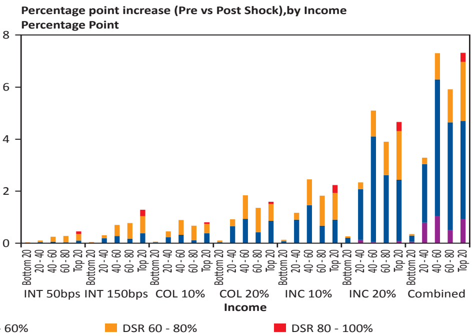
Cost of Borrowing Shock

The debt repayment capacity of borrowers is largely unaffected by the simulated 50 basis point hike in the lending rate (Graphs 17 and 18). Some signs of distress, however, are observed when the lending rate increases by 150 basis points. The impact is more evident amongst the mid- to high-income individuals. This is due mainly to the higher proportion of loans for the purchase of residential property based on variable interest rate acquired by these individuals (Graph 9). Of significance, a 150 basis point hike in lending rate raises the proportion of borrowers with DSR levels 40%~60% and DSR above 60% by 1.1 percentage points and 1.4 percentage points, respectively. On the other hand, DAR increases marginally by 1 percentage point (+RM1.2 billion) and 2.1 percentage points (+RM3.3 billion), respectively, as exposures are mainly for the purchase of properties.

Cost of Living Shock

The impact of a higher cost of living on individual borrowers is slightly higher than the cost-of-borrowing shock. The overall share of borrowers with negative FM and the DAR increases by 2.9 percentage points and 2.8 percentage points, respectively, when the cost of living goes up by 10% (Graphs 17 and 18). About two-thirds of the increase in the number of borrowers with negative FM is mainly from those below the 60th income group percentile. The higher income earners only become more financially stretched when the cost of living increases by 20%. The increase in the number of individuals who now exhibit negative FM is larger amongst those with DSR levels of 40%~80%, but the corresponding DAR is limited, even at the higher cost of living of 20%. Borrowers aged between 30 and 40 years old living in the city are found to be the most affected by the higher cost-of-living.





INT50bps, INT150bps: increase in interest rates by 50 bps and 150 bps respectively; COL10%, COL20%: increase in cost of living by 10% and 20% respectively; and INC10%, INC20%: decline in income by 10% and 20% respectively.

Graph 17. Negative FM – Individual and Combined Shocks

Graph 18: Debt-at-Risk – Individual and Combined Shocks

Income Shock

The impact from a decline in income is the most significant compared to other simulated shocks, particularly for those in the middle income group (Graph 17 and 18). On aggregate, a decline in income by 10% will increase the share of borrowers with negative FM by 7.9 percentage points, while DAR rises by 7.8 percentage points to 28%. Even borrowers in the 20%-40% DSR bucket begin to exhibit negative FM, particularly for those in the bottom 40th income percentile category. Of significance, every borrower who has a DSR greater than 60% also exhibits negative FM.

Combined Shock

The combination of all three simulated shocks is obtained by taking the extreme magnitude of each shock. The share of borrowers with negative FM increases by 31.5 percentage points across all income groups under the combined shock scenario (Graphs 17 and 18). To recap, under the pre-shock scenario, we have established that individuals with DSR of more than 60% are more susceptible to default. This is extended to those with DSR of 40%~60% under the individual macroeconomic and financial shocks. Under the combined shock scenario, individuals who exhibit negative FM extend to those with lower DSR of 20%~40%.

5. MODEL LIMITATIONS

We acknowledge that the methodology applied in this paper has some limitations that require further calibration in order to support policy analysis and formulation. We intentionally used a relatively simple model and assumptions given that this is the first attempt in using the IIID.

The probability of default computed in this model is likely to be overestimated as savings and/or liquid financial assets are not recognised as buffers against shocks. DAR could also be overestimated given the high haircut of 50% for residential properties in this model. Also, the model does not take into account any effects of possible policy intervention to manage the potential impact on the economy and financial system. The predictive ability to capture household financial stress using micro-level datasets is also relatively untested. Nevertheless, the model does provide a useful starting point for developing a holistic stress-testing framework to assess financial system exposures to individuals in distress.

Basing macroprudential analysis and decision on the data could be a challenge given the time lag. On one hand, while the structure and composition of debt and income is stable over the past years (BNM, 2015a; and BNM, 2015b), the significant change in the cost of living (BNM, 2015a) might have altered the overall results. On the other hand, the higher house prices from 2014 to 2016 may reflect higher wealth that acts as buffers against potential shocks.

6. POLICY IMPLICATION

The area of study covered in this paper can provide support for policy design and implementation to manage and mitigate risks arising from financial institutions' exposures to the household sector. For example, the relationship between the DSR and impaired household loans may provide guidance to both banks and regulators on indicative levels of the prudent threshold of the DSR for different income groups. This could contribute towards banks' credit underwriting, risk management and loan loss provisioning practices. It could also contribute towards more granular application of such a technique to different borrowers with different DSRs across different age groups, geographical location and type of financing facility.

For regulators, the area of study can better support the design and implementation of targeted macroprudential and microprudential policies to address specific sources of vulnerabilities. Through the application of proportionality of regulations, this can reduce potential unintended consequences of broad policies, such as reduced access to financing by eligible borrowers. In addition, the area of study could better inform the design and calibration of stress test scenarios and parameters in assessing the shock absorption capacity across households, banks and the financial system.

7. CONCLUSIONS

In this paper, we attempted to leverage on a new granular income and debt dataset (IIID) to study individual borrowers' debt repayment capacity. To do so, we calculated the financial margin of these borrowers and subjected them to various macroeconomic and financial shocks, including a higher cost of borrowing, an increase in the cost of living and a decline in income. We also estimated financial institutions' exposures to individuals who are more likely to be financially distressed and have a higher chance of default. The results from our empirical analysis suggest the following:

- Financial margins and the DSR are reasonable indicators to provide greater insights on the assessment on financial distress and debt repayment capacity of households across different income groups.
- The results show that, in the pre-shock scenario, borrowers across all income groups are more likely to have negative financial margin – hence, have greater likelihood of default – if their DSR is above 60%.
  - For borrowers in the bottom 40th percentile income group, a similar trend is observed if their DSR is above 40%.
  - In the post-shock scenario, borrowers across all income groups are more likely to have negative financial margin if their DSR is above 40%.
  - On aggregate, borrowers are most sensitive to an income shock, particularly those in the middle income group.
- Our analysis suggests that, even after combining all shocks based on the worst-case scenarios, the banking system in Malaysia is able to withstand the associated potential losses.

REFERENCES

Albacete, N. and Fessler, P. (2010). Stress testing Austrian households. *Oesterreichische Nationalbank Financial Stability Report* 19, pp 72–91, June.

Bank Negara Malaysia (2015a). *Annual Report 2015*, March.

Bank Negara Malaysia (2015b). *Financial Stability and Payments System Report 2015*, March.

Bilston, T., Johnson, R. and Read, M. (2015). Stress testing the Australian household sector using the HILDA survey. *Reserve Bank of Australia Research Discussion Paper* 2015-01, March.

Department of Statistics Malaysia (2015a). *Household Expenditure Survey 2014.*

Department of Statistics Malaysia (2015b). *Household Income and Basic Amenities Survey 2014*

CPS23: Statistics of Industry

A Control Chart for Monitoring Process Mean

Ah Hin Pooi
Sunway University, Petaling Jaya, Malaysia – ahhinp@sunway.edu.my

Abstract

This paper suggests a control chart which is based on the current and up to at most a fixed number N of the more recent samples. Given that the process mean is still in control, an (n+1)-dimensional multivariate power-normal distribution is fitted to the sample means of the current and n (n ≤ N) more recent samples. An (n+1)-dimensional sphere is used to form a criterion for deciding whether there is a shift in the process mean in the light of the current sample. The proposed chart is compared with the Shewhart, Cumulative Sum (cusum) and Exponentially Weighted Moving Average(EWMA) charts which have about the same in-control Average Run Lengths (ARL). It is found that for N = 5, the proposed chart always has a smaller out-of-control ARL than the Shewhart chart, and for the shift of sample mean larger than about two standard deviations, the proposed chart has a smaller out-of-control ARL than the three alternative charts. Furthermore, the performance of the proposed chart improves as we increase the value of N.

Keywords: Multivariate power-normal distribution; Multi-dimensional sphere; Average run-length.
JEL Classification: C15

1. INTRODUCTION

A process that operates only in the presence of random variation is said to be statistically in-control. If there is additional variability due to other assignable sources of variation such as materials, personnel and operating system, the process may operate at an unacceptable level of performance. In such a situation the process is said to be statistically out-of-control. An important aim of control chart is to detect assignable causes of variation from random variation. Shewhart chart was introduced by Walter A. Shewhart in 1924. It uses only the current sample to determine whether the process is in-control.

There are charts which use both the current sample and previous samples to monitor the process. These charts include the cumulative sum (cusum) chart (Ewan,1963; Page,1954; Page,1961; Gan,1991; Lucas,1976; Hawkins,1981,1993; Tan and Pooi,2005; Woodall and Adams596,1993) and exponentially weighted moving average (EWMA) chart (Roberts 1959; Crowder 1987; Lucas and Saccucci 1990).

The cusum and EWMA charts perform better than the Shewhart chart when the shifts in the process mean are small. However, for detecting large shifts in the process mean, the Shewhart chart performs better than the other two charts.

This paper suggests yet another chart which uses both the current and previous samples to monitor the process. A short description of the proposed chart is as follows.

When the first sample is available, the proposed chart declares that the process is out-of-control if the sample mean $\overline{x}$ deviates from the process mean μ by an amount beyond a chosen multiple L of the process standard deviation σ, that is

$$\left|\overline{x}-\mu\right| > L\sigma$$

Let $\overline{x}_i$ be the mean of the i-th sample. The value of the first sample mean $\overline{x}_1$ which does not provide enough evidence for us to declare that the process is out-of-control will be denoted by $\tilde{\overline{x}}_1$. A two-dimensional multivariate power-normal distribution (Pooi,2012) is next fitted to $(\tilde{\overline{x}}_1, \overline{x}_2)$. A two-dimensional sphere S_2 is used to form a criterion for deciding whether there is a shift in the process mean in the light of the observed value of $\overline{x}_2$. If the process is declared to be in-control, then the value of $\overline{x}_2$ will be denoted by $\tilde{\overline{x}}_2$. Next for $j \geq 3$ but less than or equal to some chosen integer N, a j-dimensional multivariate power-normal distribution is fitted to $(\tilde{\overline{x}}_1, \tilde{\overline{x}}_2, ..., \tilde{\overline{x}}_{j-1}, \overline{x}_j)$ and a j-dimensional sphere S_j is used to decide whether the process is still in-control in the light of the observed value of $\overline{x}_j$. For $j > N$, the value of $\left(\tilde{\overline{x}}_{j-(N-1)}, \tilde{\overline{x}}_{j-(N-2)}, \ldots, \tilde{\overline{x}}_{j-1}, \overline{x}_j\right)$ and the sphere S_N will be used to decide whether the process is still in-control in the light of the observed value of $\overline{x}_j$.

The performance of a chart may be measured by its average run length (ARL) which is the average number of samples taken before an out-of-control signal is obtained. For any two charts with the same ARL when there is no shift in the process mean, the chart with a smaller ARL when the process mean is shifted to $d\sigma$ is said to be more effective in detecting a shift of $d\sigma$.

The simulated results of ARL for the proposed chart show that when the in-control ARL is about 371, the proposed chart has an out-of-control ARL which is uniformly smaller than that of the Shewhart chart. When $d \geq 2$, the out-of-control ARL of the proposed chart for $N = 5$ tends to be the smaller than those of the cusum and EWMA charts. When N increases, the ARL of the proposed chart for $d < 2$ tends to decrease. Thus the performance of the proposed chart for $d < 2$ may be improved further by increasing the value of N.

The layout of the paper is as follows. In Section 2, we give a short introduction to the multivariate power-normal distribution. Section 3 describes the proposed chart. The values of the ARL of the proposed chart, Shewhart chart, cusum chart and EWMA chart are compared in Section 4. Section 5 concludes the paper.

2. MULTIVARIATE POWER-NORMAL DISTRIBUTION

To introduce the multivariate power-normal distribution, we may begin with the following power transformation introduced in Yeo and Johnson (2000):

$$\tilde{\varepsilon} = \psi(\lambda^+, \lambda^-, z) = \begin{cases} [(z+1)^{\lambda^+} - 1]/\lambda^+, & (z \geq 0, \lambda^+ \neq 0) \\ \log(z+1), & (z \geq 0, \lambda^+ = 0) \\ -[(-z+1)^{\lambda^-} - 1]/\lambda^-, & (z < 0, \lambda^- \neq 0) \\ -\log(-z+1), & (z < 0, \lambda^- = 0) \end{cases} \tag{2.1}$$

If z in Equation (2.1) has the standard normal distribution, then $\tilde{\varepsilon}$ is said to have a power-normal distribution.

Let y be a column vector of k' correlated random variables. The vector $\boldsymbol{y}$ is said to have a k-dimensional power-normal distribution with parameters $\boldsymbol{\mu}, \mathbf{H}, \lambda_i^+, \lambda_i^-, \sigma_i, 1 \leq i \leq k'$ if

$$y = \mu + \mathrm{H}\varepsilon \tag{2.2}$$

where $\boldsymbol{\mu} = E(\boldsymbol{y})$, $\mathbf{H}$ is an orthogonal matrix, and $\boldsymbol{\varepsilon}$ is a column vector comprising the uncorrelated variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon k'$ of which the i^{th} variable ε_i is given by $\varepsilon_i = \sigma_i[\tilde{\varepsilon}_i - E(\tilde{\varepsilon}_i)]/\{\text{var}(\tilde{\varepsilon}_i)\}^{\frac{1}{2}}$ with $\sigma_i > 0$ representing a constant, and $\tilde{\varepsilon}_i$ denoting a random variable having the power-normal distribution with parameters λ_i^+ and λ_i^-.

3. A CONTROL CHART BASED ON SPHERICAL REGIONS

Suppose that a quality characteristic is normally distributed with mean μ and standard deviation σ where μ and σ are both known. Given that the mean of the current sample of size n is $\overline{x}$, the Shewhart control chart for the process mean μ will declare that the process is out-of-control if the sample mean $\overline{x}$ deviate from the process mean by more than $z_{\alpha/2}\,\sigma/\sqrt{n}$ where/is the $100(1-\alpha/2)\%$ point of the standard normal distribution and α is a chosen small probability. A common choice of α is 0.0027 which corresponds to the value of 3 for $Z_{\alpha/2}$ and an in-control ARL of about 371.

A new control chart for the process mean may be constructed using the following steps:

Generate a large number M of the first sample mean $\overline{x}$ and denote the value of $\overline{x}_1$ by $\tilde{\overline{x}}_1$ if

$$|\overline{x}_1 - \mu| < z_{\alpha/2}\,\sigma/\sqrt{n}\,.$$

Suppose there are M_1 such value of $\tilde{\overline{x}}_1$.

(I) For each value of $\tilde{\overline{x}}_1$, generate a value of the second sample mean $\overline{x}_2$, and fit a 2-dimensional power-normal distribution to the M_1 values of $\tilde{\overline{\mathbf{x}}}^{(2)} = \left(\tilde{\overline{x}}_1, \overline{x}_2\right)^{\mathrm{T}}$.

Let the parameters of the fitted distribution be $\boldsymbol{\mu}^{(2)}, \mathbf{H}^{(2)}, \lambda_i^{(2)+}, \lambda_i^{(2)-}, \sigma_i^{(2)}, i = 1, 2$.

(II) Transform each value of $\tilde{\overline{\mathbf{x}}}^{(2)}$ in (B)(I) to $\mathbf{z}^{(2)}$ where $z_i^{(2)}$ satisfies Equation (2.1) with

$\tilde{\varepsilon}, z, \lambda^+, \lambda^-$ changed to $\tilde{\varepsilon}_i^{(2)}, z_i^{(2)}, \lambda_i^{(2)+}, \lambda_i^{(2)-}$ respectively, $\tilde{\varepsilon}_i^{(2)} = \mathrm{E}(\tilde{\varepsilon}_i^{(2)}) + \frac{s_i^{(2)}\{Var(\tilde{\varepsilon}_i^{(2)})\}^{1/2}}{\sigma_i^{(2)}}$, and $\mathbf{s}^{(2)} = \mathbf{H}^{(2)\mathrm{T}}(\tilde{\bar{\mathbf{x}}}^{(2)} - \boldsymbol{\mu}^{(2)})$ and denote the value of $(\tilde{\bar{x}}_1, \bar{x}_2)$ by $(\tilde{\bar{x}}_1, \tilde{\bar{x}}_2)$ if $Z^{(2)}$ lies inside the 2-dimensional sphere with mean 0 and initial radius $\{1-0.01 f_2 \chi^2_{2,\alpha}\}^{1/2}$ where $\chi^2_{2.\alpha}$ is the 100((1-α)% point of a chi square distribution with 2 degrees of freedom and f_2 is a chosen integer such that the total number M_2 of such $(\tilde{\bar{x}}_1, \tilde{\bar{x}}_2)$ is approximately equal to $(1-\alpha)M_1$.

(I) For $3 \leq j \leq N$, generate M_{j-1} values of the j-th sample mean $\bar{x}_j$ and fit a j-dimensional power-normal distribution to the M_{j-1} values of $\tilde{\bar{\mathbf{x}}}^{(j)} = (\tilde{\bar{x}}_1, \tilde{\bar{x}}_2, \ldots, \tilde{\bar{x}}_{j-1}, \bar{x}_j)^{\mathrm{T}}$.

Let the parameters of the fitted distribution be $\boldsymbol{\mu}^{(2)}, \mathbf{H}^{(2)}, \lambda_i^{(2)+}, \lambda_i^{(2)-}, \sigma_i^{(2)}, i = 1, 2.$

(II) Transform each value of $\tilde{\bar{\mathbf{x}}}^{(j)}$ in (C)(I) to $Z^{(j)}$ where $Z_i^{(j)}$ satisfies Equation (2.1) with $\tilde{\varepsilon}, z, \lambda^+, \lambda^-$ changed to $\tilde{\varepsilon}_i^{(j)}, z_i^{(j)}, \lambda_i^{(j)+}, \lambda_i^{(j)-}$ respectively, $\tilde{\varepsilon}_i^{(2)} = \mathrm{E}(\tilde{\varepsilon}_i^{(2)}) + \frac{s_i^{(2)}\{Var(\tilde{\varepsilon}_i^{(2)})\}^{1/2}}{\sigma_i^{(2)}}$, and $\mathbf{s}^{(2)} = \mathbf{H}^{(2)\mathrm{T}}(\tilde{\bar{\mathbf{x}}}^{(2)} - \boldsymbol{\mu}^{(2)})$ and denote the value of $(\tilde{\bar{x}}_1, \tilde{\bar{x}}_2, \ldots, \tilde{\bar{x}}_{j-1}, \bar{x}_j)$ by $(\tilde{\bar{x}}_1, \tilde{\bar{x}}_2, \ldots, \tilde{\bar{x}}_{j-1}, \tilde{\bar{x}}_j)$ if $Z^{(j)}$ lies inside the j-dimensional sphere with centre 0 and initial radius $\{1-0.01 f_j \chi^2_{j,\alpha}\}^{1/2}$ where $\chi^2_{j,\alpha}$ is the 100(1 - α)% point of a chi square distribution with j degrees of freedom and f_j is a chosen integer such that the total number M_j of such $(\tilde{\bar{x}}_1, \tilde{\bar{x}}_2, \ldots, \tilde{\bar{x}}_{j-1}, \tilde{\bar{x}}_j)$ is approximately equal to $(1 - \alpha)M_{j-1}$.

The radius of the sphere $s^{(j)}$ is now adjusted to $g\{1-0.01 f_j \chi^2_{j,\alpha}\}^{1/2}$ where g is chosen such that the control chart has the required in-control ARL.

The new control chart for the process mean may now be applied using the following steps:

Declare that the process is out-of-control if the first sample mean $\bar{x}_1$ deviates from the process mean by more than $z_{\alpha/2}\sigma/\sqrt{n}$.

For $2 \leq j \leq N$, declare that the process is out-of-control if

The process is still declared to be in-control in the light of the (j - 1)-th sample mean, and

lies outside the sphere $S^{(j)}$.

For $j \geq N + 1$, transform the value of $\tilde{\bar{\mathbf{x}}}^{[j]} = (\tilde{\bar{x}}_{j-(N-1)}, \tilde{\bar{x}}_{j-(N-2)}, \ldots, \tilde{\bar{x}}_{j-1}, \bar{x}_j)$ to $Z^{(j)}$ in a way similar to the transformation of $\tilde{\bar{\mathbf{x}}}^{(N)}$ to $Z^{(N)}$, and declare that the process is out-of-control if

The process is still declared to be in-control in the light of the (j - 1)-th sample mean, and

$z^{(j)}$ lies outside the sphere $s^{(N)}$.

Comparison of the proposed control chart with cusum and EWMA charts

There are a number of control charts which make use of both the current and previous samples. These charts include the popular charts such as cusum and EWMA charts.

The two-sided cusum procedure may be described via the following Tabular Cusum with the reference value $K = k\sigma/\sqrt{n}$ and decision interval $H = h\sigma/\sqrt{n}$:

$$C_i^+ = \max\left[0, \bar{x}_i - (\mu + K) + C_{i-1}^+\right]$$

$$C_i^- = \max\left[0, (\mu - K) - \bar{x}_i + C_{i-1}^-\right]$$

$$C_0^+ = C_0^- = 0$$

If either C_i^+ or C_i^- exceeds the decision interval H, the process is considered to be out-of-control.

The EWMA chart for monitoring the process mean μ may be described as follows:

$$z_i = \lambda\bar{x}_i + (1-\lambda)z_{i-1}$$

where $0 < \lambda \leq 1$ is a constant and the starting value Z_0 is the process mean μ, the process is declared to be out-of-control by an EWMA chart with width L if

$$z_i > \mathrm{UCL} = \mu + L\frac{\sigma}{\sqrt{n}}\left\{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\right\}^{\frac{1}{2}} \quad \text{or}$$

$$z_i < \mathrm{LCL} = \mu - L\frac{\sigma}{\sqrt{n}}\left\{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\right\}^{\frac{1}{2}}$$

The ARL of the proposed chart together with those of two examples for cusum and EWMA charts are given in Table 1. The observations which can be obtained from the table are as follows. The in-control ARL of the various charts are all near to 371. The out-of-control ARL of the proposed chart for $N = 5$, 10 and 15 are smaller than that of the Shewhart chart for all values of the shift $d\sigma$. For the shift which is larger than 2σ, the proposed chart begins to have the lowest ARL. When N increases, the ARL of the proposed chart for the shift which is smaller than 2σ tends to decrease. This means the performance of the proposed chart for the shift less than 2σ may be improved further by increasing the value of N.

Table 1. ARL of The Various Types of Charts ($n = 1$)

d	Shewhart	cusum	EWMA	Proposed Chart		
		$k = 0.5$, $h = 4$	$L = 2.86$, $\lambda = 0.2$	$N = 5$	$N = 10$	$N = 15$
0	371	336	370.65	372.3039	370.1216	371.1203
0.5	155.22	26.6	36.16	119.0525	98.471	91.7785
1	44	8.38	7.96	22.5314	18.311	17.6169
1.5	14.97	4.75	5.23	7.1936	6.5423	6.7106
2	6.3	3.34	3.58	3.4936	3.4747	3.5087
2.5	3.24	2.62	2.78	2.2173	2.2537	2.2707
3	2	2.19	2.31	1.666	1.6622	1.6724

5. CONCLUSIONS

The multivariate power-normal distribution enables us to transform the original random variables into a set of independent and standard normally distributed random variables. When we have a number of independent and standard normally distributed random variables, the spherical region with centre 0 will form a good acceptance region for testing the null hypothesis that the means of the original random variables are all zero. Thus it is surprising that the proposed chart performs better than the Shewhart chart for all values of the shift.

Further research may be carried out to investigate the performance of the proposed chart for larger values of N. It is hoped that we may be able to extend the range of the shift of the process mean within which the proposed chart performs better than the cusum and EWMA charts.

REFERENCES

Crowder, S. V. (1987). "A Simplified Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts", *Technometrics*, 29(4), 401–407.

Crowder, S. V. (1989). "Design of Exponentially Weighted Moving Average Schemes", *Journal of Quality Technology*, 21(2), 155–162.

Ewan, W. D. (1963). "When and How to Use Cu-Sum Charts", *Technometrics*, 5(1), 1–22.

Gan, F. F. (1991). "An Optimal Design of CUSUM Quality Control Charts", *Journal of Quality Technology*, 23(4), 279–286.

Hawkins, D. M. (1993a). "Cumulative Sum Control Charting: An Underutilized SPC Tool", *Quality Engineering*, 5(3), 463–477.

Hawkins, D. M. (1981). "A CUSUM for a Scale Parameter", *Journal of Quality Technology*, 13(4), 228–235.

Lucas, J. M. (1976). "The Design and Use of V-mask Control Schemes", *Journal of Quality Technology*, 8(1), 1–12.

Lucas, J. M., and Saccucci, M. S. (1990). "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements", *Technometrics*, 32(1), 1–29.

Page, E. S. (1954). "Continuous Inspection Schemes", *Biometrics*, 41(1), 100–115.

Page, E. S. (1961). "Cumulative Sum Control Charts", *Technometrics*, 3(1), 1–9.

Pooi, A. H. (2012). "A Model for Time Series Analysis", *Applied Mathematical Sciences*, 6, 5735-5748

Roberts, S. W. (1959). "Control Chart Tests Based on Geometric Moving Averages", *Technometrics*, 42(1), 97–102.

Woodall, W. H., and Adams, B. M. (1993). "The Statistical Design of CUSUM Charts", *Quality Engineering*, 5(4), 559–570.

Tan, Y. F., and Pooi, A. H. (2005). "Run Length Distribution of Two-Sided CUSUM Procedures for Continuous Random Variables", *Bulletin of the Institute of Mathematical Academia Sinica*, 33(2),

A Stochastic Linear Programming Approach for Solving Production-and-Inventory Problem

Ramya Rachmawati*
Mathematics Department, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Bengkulu, Indonesia
Email: ramya_unib@yahoo.co.id

Abstract

Inventory control is vital for better management. Inventory problem which involves deterministic parameters can be solved using a linear programming approach. In reality, managers often face a difficult situation for making decisions due to the uncertainty of demand and resource availability. In this paper, a Stochastic Linear Programming (SLP) model is applied to minimize the total production-and-inventory cost over a planning horizon under probabilistic environment. This multistage model is then demonstrated using hypothetical data. The model demonstration shows that the model works well to handle this probabilistic production-and-inventory problem.

Keywords: Stochastic linear programming; Production-and-inventory plan; Uncertainty

1. INTRODUCTION

Production and inventory planning play important role in operations management. Managers often face a difficult situation of how to minimize the costs associated with maintaining inventory and meeting customer demand (Hillier and Lieberman, 2001). When demand is certain and at a constant rate each period, Economic Order Quantity (EOQ) models can be used to solve inventory problem (Winston and Goldberg, 2004). Methods such as linear programming is a powerful tool that can be used to deal with the planning problem when product demand and resource availability are known with certainty but no need to be at a constant rate. In the presence of uncertainty, however, the methods may be unsatisfying in determining good production plans. For that reason, we need another approach to solve the inventory problem, such as stochastic linear programming approach.

The objective of this study is to address a multistage production-and-inventory planning problem under uncertainty that minimizes the total inventory cost over a planning horizon.

Model

An important characteristic of multistage models is that random phenomena evolutes over time. The random variables in most cases depend on the values of the precedent random variables (Zanjaniet al., 2010). Therefore, it is convenient to use a tree for illustrating the scenarios, where the outcomes in one stage branch out from the previous stage's outcome (Figure 1). The model to minimize the total production and inventory cost over a planning horizon under probabilistic environment, adapted from Sen and Higle (1999), is as follows.

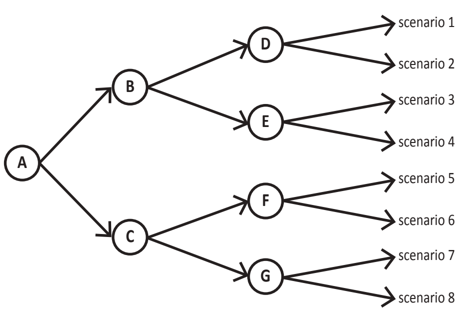


Figure 1 . The Scenario Tree of Three Periods

Indices and sets:
$\mathcal{N}$ = scenario tree
i = product
k = resource
t = period, $t = 1, 2, ..., T$
S = scenario
n = node of scenario tree
$S(n)$ = ancestor of node n in the scenario tree
$t(n)$ = sets of time periods related to node n in the scenario tree

Parameters:
P_s = the probability that scenario S occurs
$C_{i,t}$ = the per-unit cost to produce product i in period t
$h_{i,t}$ = the holding cost for each product i at the end of period t
$\pi_{i,t}$ = the back-order cost for each product i at the end of period t
$\alpha_{i,k}$ = the quantity of units of resource k needed to produce a unit of product i
$b_{k,t}$ = the quantity of resource k available in period t
$b_{k,t,s}$ = the value of $b_{k,t}$ related to scenario S
$d_{i,t}$ = the demand of product i in period t
$D_{i,t,S}$ = the value of $d_{i,t}$ related to scenario S

Decision variables:
$X_{i,t,s}$ = the number of units of product i produced in period t under scenario S
$I^+_{i,t,s}$ = the inventory size of product i by the end of period t under scenario S
$I^-_{i,t,s}$ = the back-order size of product i by the end of period t under scenario S
$I_{i,t,s}$ = the net inventory of product i by the end of period t under scenario S

Minimize

$$\sum_s p_s \sum_i \sum_t c_{i,t} X_{i,t,s} + \sum_s p_s \left[\sum_i \sum_t \left(h_{i,t} I^+_{i,t,s} + \pi_{i,t} I^-_{i,t,s} \right) \right] \quad 1$$

Subject to

$$\sum_i a_{i,k} X_{i,t,s} \le b_{k,t,s}, \quad \forall k,t,s \quad 2$$

$$I_{i,0,s} = A_{i,s}, \qquad \forall i,s \quad 3$$

$$-I_{i,t,s} + I_{i,t-1,s} + X_{i,t,s} = D_{i,t,s}, \qquad \forall i,t,s \quad 4$$

$$I_{i,t,s} - I^+_{i,t,s} + I^-_{i,t,s} = 0, \qquad \forall i,t,s \quad 5$$

$$Y_{i,n} - X_{i,t(n),s} = 0, \qquad \forall s \in S(n), n \in \mathcal{N} \quad 6$$

$$H_{i,n} - \left(I^+_{i,t(n),s} + I^-_{i,t(n),s} \right) = 0, \qquad \forall s \in S(n), n \in \mathcal{N} \quad 7$$

$$X_{i,t,s} \ge 0, I^+_{i,t,s} \ge 0, I^-_{i,t,s} \ge 0, \forall i,t,s \quad 8$$

The objective function (1) minimizes the total production and inventory costs (i.e. holding and backorder costs) throughout a planning horizon for each product under all possible scenarios. Constraints (2) ensure that the production quantity does not exceed the resources available. Constraint (3) set the initial inventory level for each product. The balance equations (4) make sure that the net inventory of product i at the end of period t is the summation of the previous net inventory plus the total production in period

t, minus the demand for that period. Constraints (5) states the net inventory each period. Constraints (6) and (7) are the implementability constraints. These constraints mean that scenarios that have a common history are forced to result in a common production and inventory plan at that time. Here, the variables $\{Y_{i,n}\}$ and $\{H_{i,n}\}$ denote the production-and-inventory plan related to node n. Finally, Constraints (8) ensure that the decision variables take nonnegative values.

3. MODEL DEMONSTRATION

The model presented in the previous section is demonstrated using hypothetical data. The model is solved by ILOG CPLEX 12.6 with the Python 2.7 programming language using PuLP modeler.

Suppose that a company can produce two types of products. Demand is uncertain, however, based on past data, the company believes that demands for the three periods are equally likely. It is 300, 320, 600 (setting 1) or 470, 490, 520 (setting 2) for product 1. and 500, 550, 440 (setting 1) or 450, 450, 590 (setting 2) for product 2. The company has the initial inventory of 100 for each product. Suppose that production and backorder costs per item per period are 2 and 4, while the cost of holding one item per period is 1. Each product uses two types of raw material. Product 1 needs 2 units of material 1 and 6 units of raw material 2, while product 2 needs 3 units of material 1 and 4 units of raw material 2. The availability of the two types of material is uncertain. However, the company believes that the availability of the material for the three periodsis equally likely. It is 4100, 4200, 4400 (setting 1) or 3600, 3650, 3700 (setting 2) for material 1; and 3000, 3100, 2900 (setting 1) or 2700, 2750, 2600 (setting 2) for material 2.

The length of the planning horizon is three periods. We assume that the setting for demand and resource are the same for each period. Therefore, there are eight possible scenarios throughout the planning horizon (Figure 1), $p_s = \frac{1}{8}$, $s = 1, 2, \ldots, 8$. The main question is how many production sizes of each product should be produced each period so as to minimize the production-and-inventory costs. The solution of this problem is as described in Table 1.

Table 1. Optimal Solution for Three Periods

Period	Scenario	Product 1	Product 2
1	1 - 8	200.00	375.00
2	1 - 4	108.33	525.00
2	5 - 8	141.66	475.00
3	1 - 2	56.66	565.00
3	3 - 4	123.33	465.00
3	5 - 6	56.66	565.00
3	7 - 8	123.33	465.00
By the end of period	**Scenario**	**Net Inventory of Product 1**	**Net Inventory of Product**
1	1 -4	-85.00	-25
1	5 -8	-170.00	+25
2	1 -2	-296.67	-50
2	3 -4	+85.00	-381.67
2	5 -6	-433.33	-50
2	7 -8	-518.33	+50
3	1	-755.00	+75
3	2	-715.00	-75
3	3	-33.00	+75
3	4	-818.33	-75
3	5	-891.66	-75
3	6	-851.66	-75
3	7	-995.00	+75
3	8	-995.00	-75

The total production and inventory cost is 9650. From the table, we can see that the company cannot avoid the shortage due in part to the uncertainties of demand and resource availability each period.

4. CONCLUSION

In this paper, we presented a stochastic linear programming approach for solving a production-and-inventory problem. We then illustrated the model in a case study under uncertainties. The problem was to determine the optimal production scheduling to minimize the total production and inventory cost over a planning horizon. The model illustration showed that the model worked well for three-year planning horizon. In the future, practitioners can utilize this promising approach in addressing real life problems that incorporate probabilistic environment.

REFERENCES

Hillier, F. S., & Lieberman, G. J. (2001). Introduction to Operations Research, McGraw Hill. New York.

Sen, S., & Higle, J. L. (1999). An introductory tutorial on stochastic linear programming models. Interfaces, 29(2), 33-61.

Winston, W. L., & Goldberg, J. B. (2004). Operations research: applications and algorithms (Vol. 3). Boston: Duxbury press.

Zanjani, M.K., Nourelfath, M., & Ait-Kadi, D. (2010). A multi-stage stochastic programming approach for production planning with uncertainty in the quality of raw materials and demand. International Journal of Production Research, 48(16), 4701-4723.

Optimal Design of Shewhart-Lepage Type Schemes and Its Application in Monitoring Service Quality

Amitava Mukherjee
XLRI –Xavier School of Management, Jamshedpur, India – amitmukh2@yahoo.co.in

Abstract

We generalize popular distribution-free (nonparametric) Shewhart-Lepage scheme for simultaneously monitoring of location and scale parameters using an adaptive approach. This approach is known as percentile modifications of ranks (or adaptive Gastwirth Score) in statistical literature. This is a powerful tool to improve rank tests for shift detection. The adaptive Gastwirth Score is not much familiar among quality control practitioners and therefore rarely used in practice. However, such scores are very useful in the presence of a few outliers in the reference sample or Phase-I observations and in detecting various types of shifts. Considering its distinct advantages, we develop a new class of Shewhart-type adaptive Lepage-Gastwirth (ALG) scheme. We discuss optimal implementation strategies of the proposed scheme to achieve lower out-of-control (OOC) average run length (ARL) and false alarm rate (FAR). This scheme typically designed to monitor service quality where reference sample may be contaminated. Post signal follow-up procedures of the proposed Shewhart-type optimal ALG chart is discussed. We illustrate the use of optimal ALG charts with a recent data on Vancouver city call centre service quality monitoring.

Keywords: Quality Control; Adaptive Lepage-Gastwirth (ALG) Scheme; Outliers; Service Quality; Simultaneous Monitoring.

This page intentionally left blank

CPS24:
Macroeconomic Statistics (4)

Spatial and Temporal Analysis of Tourism Arrivals and Income Distribution Patterns in Thailand

Anuch Nuaklong
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - anuch.nuaklong@gmail.com

Chotirat Sriwirat
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - Chotirat_Sriwirat@hotmail.com

JunthipTiengtum
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - nutip_4029@hotmail.co.th

Patcharin Chimdit
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - mookchimdit@gmail.com

Wannapisit Thammakul
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - greenreadpean@gmail.com

Assoc. Prof. Raymond J. Ritchie
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - raymond.ritchie@uni.sydney.edu.au

Sangdao Wongsai
Prince of Songkla University, Phuket Campus, Phuket, 83120, Thailand - sangdao.w@phuket.psu.ac.th

Abstract

Tourism is one of the increasingly important components of the economies of many countries worldwide. Thailand is rich in natural resources that attract international tourists from around the world. According to the 2013 Country Brand Ranking of 187 countries, Thailand ranked second globally for its destination image. In 2015, international tourist arrivals were more than 63 million and tourism revenue of the country was more than US$30 billion. Objectives of this paper were to study the spatial and temporal patterns of tourism arrivals and income distributions across 77 provinces of Thailand during 20011-2015, based on the monthly data collection from the Department of Tourism, the Ministry of Tourism and Sports. The mapping results were presented using the ArcGIS version 10.2. We found significant patterns and trends of the number of tourists' distributions across different regions of Thailand over the five-year period. Most tourists visited the Central region and not other parts of the country resulting in an imbalance in the regional economic distribution of tourism. Therefore, the Thai government planners should pay attention to the expansion of tourism growth across the country to ensure a better economic distribution.

Keywords: Tourist receipts; Tourist arrivals; Inbound tourism; Thailand tourism statistics.

1. INTRODUCTION

Thailand comprises several distinct geographic regions, partly corresponding to the provincial groups [1]. The north of the country is the mountainous area of the Thai highlands, with the highest point being Doi Inthanon in the Thanon Thong Chai Range at 2,565 meters above sea level (Figure 1 A). The northeast, Isan region, consists of the Khorat Plateau, bordered to the east by the Mekong River.

Figure 1. Top Destinations of Thailand; (A.) Doi Inthanon Chiang Mai [2] and (B.) Phi Phi Island, Krabi [3]

The Andaman Sea is a precious natural resource as it hosts the most popular and luxurious resorts in Asia. Phuket, Krabi, Ranong, Phang-Nga and Trang, and their islands, all lay along the coast of the Andaman Sea and, they act as a tourist magnet for visitors from around the world (Figure 1B).

Tourism is a major economic factor in the Kingdom of Thailand. Estimates of tourism revenue and number of tourists increased during 2011-2015. Tourism distributes development from economic centers to less developed areas [4]. In many countries the activity of tourism is considered more important than primary industry with regard to economic and social benefits [5]. Tourism has an excellent potential to be a catalyst for economic growth and for this it is a key sector in macro-economic development [6] [7]. The objective of this study was to investigate the spatial and temporal distributions of the number of visitors and the tourism revenue across the country of Thailand over the period of 2011-2015.

2. MATERIAL AND METHOD

2.1. Data

Visitor statistics for all 77 provinces across Thailand have been recorded monthly by the Department of Tourism, Ministry of Tourism and Sports, Thailand. In this study, the number of visitors to each province and its tourism revenue during 2011-2015 were considered. Thailand has 77 provinces, divided into 6 regions including Central Thailand with 12 provinces, Eastern Thailand with 8 provinces, (Isan) Northeastern Thailand with 20 provinces, Southern Thailand with 14 provinces, Western Thailand with 6 provinces and Northern Thailand with 17 provinces (Figure 2A).

2.2. Preliminary data analysis

According to Liu and his colleagues [8], geostatistical analysis has been widely used in geology, agrology, hydrology and climate [9], and recently adopted in analysing data in many socio-economic fields of research as well as tourism studies. In this study, spatial and temporal patterns of the number of tourist arrivals as well as the tourism revenue were examined using the ArcGIS version 10.2.

3. RESULTS AND DISCUSSION

There are two sets of results that we will review. Figure 2B shows annual comparisons of the tourism revenue by province during 2011-2015 and Figure 2C shows the spatial distribution of the average of the tourism revenue over five years. The top three average revenues from tourists are for Bangkok, Phuket and Chiang Mai, respectively.

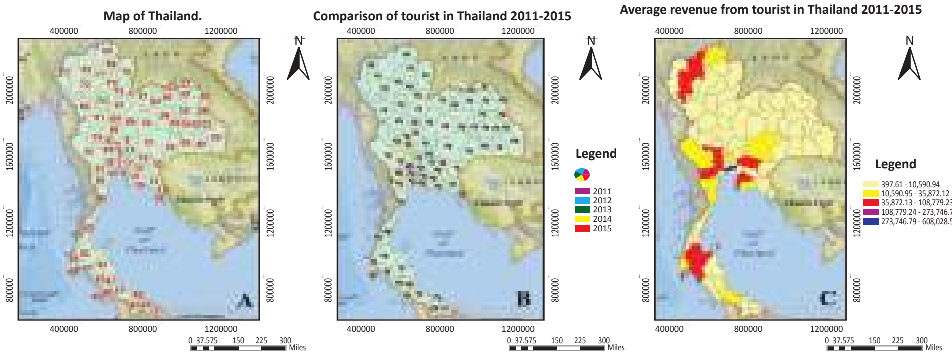


Figure 2. (A) The seventy-seven provinces of Thailand: 1: Bangkok; 2: Angthong; 3: Ayuttaya; 4: Chainat; 5: Lopburi; 6: Nakhonphathom; 7: Nontaburi; 8: Phathumthani; 9: Samutprakan; 10: Samutsakon; 11: Saraburi; 12: Singburi; 13: Chanthaburi; 14: Chachoengsao; 15: Chon Buri; 16: Trad; 17: Nakhonnayok; 18: Prachinburi; 19: Rayong; 20: Sakaew; 21: Amnat Charoen; 22: Bueng Kan; 23: Buri Ram; 24: Chaiyaphum; 25:Kalasin; 26: Khon Kaen; 27: Loei; 28: Maha Sarakham; 29: Mukdahan; 30: Nakhon Phanom; 31: Nakhon Ratchasima; 32: Nong Bua Lam Phu; 33: Nong Khai; 34: Roi Et; 35: Sakon Nakhon; 36: Si Sa Ket; 37: Surin; 38: Ubon Ratchathani; 39: Udon Thani; 40: Yasothon; 41: Krabi; 42: Pattani; 43: Chumphon; 44: Nakhon Si Thammarat; 45: Narathiwat; 46: Phang Nga; 47: Phatthalung; 48: Phuket; 49: Ranong; 50: Satun; 51: Songkhla; 52: Surat Thani; 53: Trang; 54: Yala; 55: Kanchanaburi; 56: Phetburi; 57: Ratchaburi; 58: Samutsongkam; 59: Suphanburi; 60: Prachuapkhirikhan; 61: Kamphaengphet; 62: Chiang Rai; 63: Chiang Mai; 64: Phichit; 65: Nakhonsavan; 66: Tak; 67: Phitsanulok; 68: Phayao; 69: Phetchabun; 70: Phrae; 71: Lamphang; 72: Lamphun; 73: Uttaradit; 74: Uthaithani; 75: Sukhothai; 76; Nan; 77: Maehongson.

The average revenue over the five years of study by province. (B) Spatial distribution of the average revenue of 5-year period (2011-2015) across Thailand. (C) Comparison of the average revenue during 2011-2015 by province.

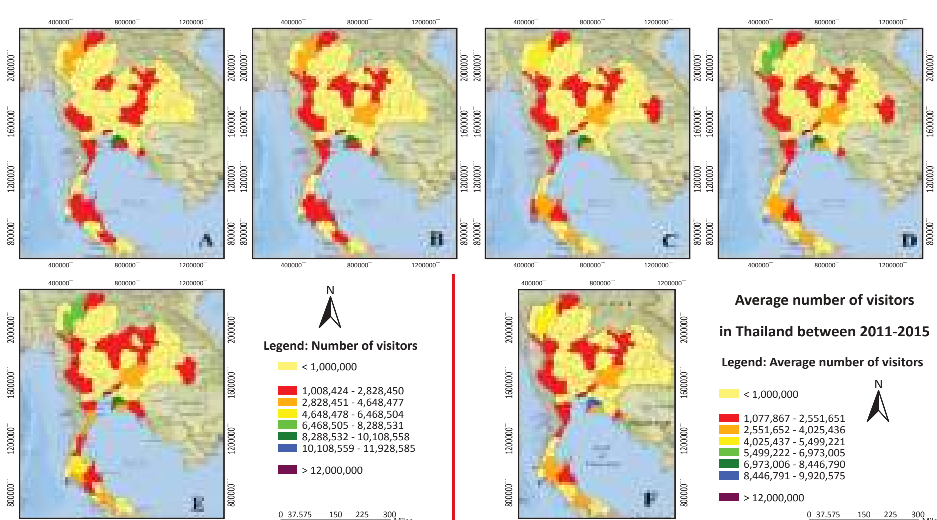


Figure 3. The number of tourist arrivals in Thailand; (A.) 2011, (B.) 2012, (C.) 2013, (D.) 2014, (E.) 2015 and (F.) average over the 5-year period of study.

The second are the results of the number of tourist arrivals in Thailand from 2011-2015. Figure 3 shows the distribution of tourists to the country's 77 provinces. The largest number of tourists visit Bangkok because it's the center of the country. Tourists take a break in order to travel to the provinces such as Chiang Mai and Phuket which are the provinces that are the other popular tourist destinations but earn a lower revenue compared to Bangkok. In Chiang Mai and Phuket there are a lot of natural attractions. In Chiang Mai, the terrain is mountainous and has unique cultural attractions such as Karen, making it attractive to tourists. Phuket is a tropical island with varied topography and has a beautiful sea. The most popular sites in Phuket are Patong Beach and Promthep Cape. The top three average revenues from tourists are for Bangkok, Phuket and Chiang Mai.

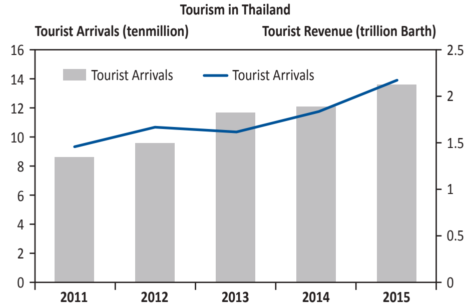


Figure 4. The Number of Tourist Arrivals in Thailand During 2011-2015

Figure 4 is a graph which compares the total amount of tourist revenue. The graph shows that revenues have increased over time. The number of tourists is increasing as time goes on but revenue has increased at a rate slower than arrivals. This means that tourists spend less per capita than before.

4. CONCLUSIONS

This study is showing the tourism arrivals in Thailand are concentrations (not scatter) in Bangkok, Phuket, Chon Buri and Chiang Mai provinces, respectively. These provinces are so-called economic provinces. Because economic provinces have facilities for tourists such as hotels and transport. From Thailand tourism inbound statistic founded the most of the number of tourism inbound is Chinese and other Asian country respectively. And the trend of the number of revenue and tourist arrival appears that increased but number of revenue lower increased when to prepare between tourist arrivals as the Thailand has more promotion for travel promoted.

5. ACKNOWLEDGEMENTS

The authors wish to thank Prince of Songkla University for providing the computer and software facilities and funding for the project. This project was a joint project by the authors as part of a Masters in Science program in The Faculty of Technology and Environment at Prince of Songkla University-Phuket. The co-operation of Department of Tourism, Thailand is gratefully acknowledged.

REFERENCES

P. Ponjan and N. Thirawat, "Annals of Tourism Research Impacts of Thailand ' s tourism tax cut : A CGE analysis," vol. 61, pp. 45–62, 2016.

Weloveandaman. (n.d.). ทัวร์เกาะพีพี เรือเฟอรี เต็มวัน (ทัวร์ราคาประหยัด/เรือออกจากภูเก็ต) [Online]. Available:

http://www.weloveandaman.com/index.php?route=product/product&product_id=50

Van de Tong. **(**2014, Nov. 21**)**. **ดอยอินทนนท์ เสน่ห์ดึงดูดทึงดงามไม่เปลียนแปลง** [Online]. Available:

Travel MThai http://travel.mthai.com/region/north/99627.html

R. Shareef and M. Mcaleer, "Modelling international tourism demand and uncertainty in Maldives and Seychelles : A portfolio approach," vol. 78, pp. 459–468, 2008.

H. Li, J. Li, G. Li, and C. Goh, "Annals of Tourism Research Tourism and regional income inequality : Evidence from China," *Ann. Tour. Res.*, vol. 58, pp. 81–99, 2016.

F. Liu, J. Zhang, J. Zhang, and D. Chen, "Geostatistical Analysis of China Inbound Tourism Spatial Distribution Structure."

D. N. Zurick, "Adventure travel and sustainable tourism in the peripheral economy of Nepal". Annals of the Association of American Geographers., vol 82, pp. 608–628, 1992.

A. Liu and S. Pratt, "Tourisms vulnerability and resilience to terrorism," *Tour. Manag.*, vol. 60, pp. 404–417, 2017.

R. Baggio and R. Sainaghi, "Mapping time series into networks as a tool to assess the complex dynamics of tourism systems," *Tour. Manag.*, vol. 54, pp. 23–33, 2016.

Tourists' Perceptions on Safety and Security: a Case Study of Phuket Island of Thailand

Jaruwan Manui*
Prince of Songkla University, Phuket campus, Phuket, Thailand – jaruwan_j_manui@hotmail.com

Sangdao Wongsai
Prince of Songkla University, Phuket campus, Phuket, Thailand – sangdao.w@phuket.psu.ac.th

Abstract

Phuket Island has globally renowned as the sea-sun-sand destination, resulting in a rapidly increasing number of arrivals reaching 10 million in 2015. Therefore, safety and security are significant to provide quality in the island tourism and to maintain its destination image. The purpose of this study was to evaluate how international tourists felt about safety and security while visiting the island, based on data collection at the Phuket International Airport during March-April, 2016. An exploratory study on tourists' profile and their perceptions of safety and security environment were numerically and graphically summarized to gain insight into the data. Then, multivariate of analysis of variance was carried out to investigate the effects of cultural differences and destination familiarity on the safe and secured perceptions. The findings reveal that the tourists from Oceania and Europe countries felt safer with touristic activities than those from the Asian countries. The more often they visited the destination the less comfortable tourists were with the traffic and road safety on the island. A further study on potential factors associated with the perceptions toward the safe and secure environment on the island is urgently recommended.

Keywords: Tourist behaviour; Destination image; Thai tourism; Repeat visitors

1. INTRODUCTION

Tourism is one of the most important industries for Thailand and thus contributes to the country's economy through a variety of business sectors and investment in the manufacturing and services, such as hotels, accommodation, restaurants, transportation, and trades. These activities have created a number of job opportunities and opportunities to build a career, subsequently spreading prosperity in the tourist regions and the home provinces of migratory workers [1], [2]. Therefore, understanding and measuring the components of tourist destination image is vital for enhancing the development of a tourism-driven economy. We have made a study of the behavior of the tourists as influenced by their individual socio-demographic characteristics and travel behaviors and the different nationalities of tourists [3]–[5]. Numbers of visits was a significant determining factor in attitudes and behavior and on destination image and satisfaction [6] and helps tourists to decide where to travel and places to avoid [7]. Safety and security of tourists is one of the major image components influencing the decision of tourists on destination selection [3], [4], [8]–[10]. Many studies have reported that tourists' perceptions of safety and security were dependent on their demographic differences, mainly gender [11], [12] and age [1], [13]. The purpose of this study was to examine demographic factors that had influenced perceptions of tourists on safety and security in Phuket, the famous tourist destination of Thailand for its local history, natural attractions, and world-renowned diving sites. In 2014, Phuket had welcomed more than 11 million visitors and brought in about 260 billion Baht (nearly 7.3 USD billion), accounting for 50% of the country's tourism revenue (National Statistical Office, 2016). The economy in the region is largely derived from tourism industries.

2. PHUKET ISLAND OF THAILAND

Phuket is the largest island in Thailand of a group that also includes approximately 39 smaller islands and the region has approximately an area of 570 sq. km. (Phuket Provincial Governor's Office, 2001). Figure 1 shows the location of Phuket Island. It is located in the Andaman Sea off the west coast of Thailand. Phuket Island is globally renowned as the sea-sun-sand destination and one of the best world-class tropical beach resorts on the Andaman coast because of its beautiful beaches, crystal-clear blue sea, and mountain greenery [14], resulting in a rapidly increasing number of arrivals reaching 10 million in 2015. According to the travel and tourism competitiveness Index Report in 2015, Thailand ranked 132 out of 141 countries on its Safety and Security. This demonstrates the importance of safety and security issues while visiting Phuket (The Travel & Tourism Competitiveness Report, 2015). Therefore, the study of safety and security is significant to highlight the need for improvements in the quality of Phuket tourism and to maintain its destination image.

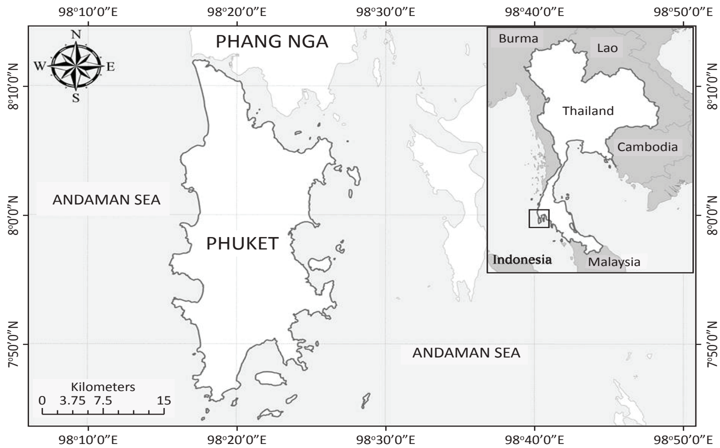


Fig. 1. Location of Phuket Island

3. RESEARCH METHODOLOGY

3.1. Research Hypotheses

Data was randomly collected from 399 international tourists who were visiting Phuket and about to depart from the Phuket International Airport during March-April, 2016. The questionnaire was distributed both on weekdays and weekends to ensure the coverage for all possible target populations. The questionnaire was designed based on the review of the literature on the destination image in tourism, tourists' observations, and recommendations are a respected source of ideas for destination image [15]. The purpose of this study was to evaluate how international tourists felt about safety and security while visiting the Phuket Island. Therefore, a questionnaire survey was designed so as to test the following alternative hypotheses:

Hypothesis 1:

H0: there is no significant difference in the perception of tourists' nationalities felt about safety and security while visiting the Phuket Island.

H1: there is a significant difference in the perception of tourists' nationalities felt about safety and security while visiting the Phuket Island.

Hypothesis 2:

H0: there is no significant difference in the perception of tourists' about number of visits including this trip felt about safety and security while visiting the Phuket Island.

H1: there is a significant difference in the perception of tourists' about number of visits including this trip felt about safety and security while visiting the Phuket Island.

3.2. Research Design

This research is a descriptive and quantitative research to describe characteristics of respondents and their perceptions regarding the variables presented in the conceptual framework. The main objective of descriptive research is to describe characteristics of a population or phenomenon [16]. For this study, the research methodology was based on primary data by using a questionnaire designed by the researchers. The items of the questionnaire were designed for the independent variables in first part and the dependent variables in second part, the first part of the questionnaire, the tourists were asked to disclose personal data including demographic information (e.g., age, gender, nationality, travel party (with whom) and number of visits including this trip). In the second part of the questionnaire, there are 11 variables to measure the satisfaction and perceptions of tourist were listed as follows: (1) touring city in daytime [11], [17], (2) the amount of lighting at night [18], (3) accessibility of the destination, (4) tourist spots and places of visit [19], (5) transportation [20], [21], (6) traffic conditions [22]–[24], (7) quality roads bridge utility services [25], (8) accommodation area [19], [26], (9) fair shopping offer [27], [28], (10) beach lifeguard and services [29] and (11) water safety [27]. A five-point Likert scale ranging from Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5, was utilized as a tool for measuring tourists' perceptions. The R programming [32] was used for statistical analysis. One way ANOVA was used to analyze and assess the differences among tourists in terms of nationalities and the number of visits including this trip. In the ANOVA, Tukey's test for post hoc comparisons was also chosen because all couples of means are being compared [30], [31].

4. RESEARCH FINDINGS

4.1 Sample Profile

This section presents the descriptive analysis of the participants' demographic characteristics including gender, age, nationality, travel party (with whom), and number of visits including this trip of visit to Phuket Island. The frequency and percentage distributions of these characteristics are provided in Table 1 and discussed below.

Table 1. Profile and Trip Characteristics of Respondents

Variables	Frequency (n = 399)	Percentage (%)	Variables	Frequency (n = 399)	Percentage (%)
Gender			**Number of visits including this trip**		
Male	210	52.6	First time	231	58.2
Female	189	47.4	2-3 times	107	27.0
Age			4-5 times	38	9.57
20-29	136	34.3	> 5 times	21	5.29
30-39	121	30.5	**Nationality**		
40-49	68	17.1	Asia	81	20.3
50-59	51	12.9	Australia/Oceania	87	21.8
Over 60	21	5.29	Europe	129	32.3
Travel party (with whom)			North America	30	7.52
Alone	33	8.35	Other	72	18.1
Accompanied	205	51.9			
Family	157	39.8			

Table 1 shows descriptive statistics of profile and trip characteristics of respondents for 399 respondents who visited Phuket Island during March-April 2016. Of a total sample 399 respondents, 32.3% were from countries of Europe (including Russia), 21.8% were from Australia/Oceania, 20.3% were from Asia (including the Middle east), while the remaining 7.5% of the tourists were from North America. Gender of respondents 52.6% (210) were male and 47.4% (189) were female. A large majority of respondents were in age groups biased towards the 20-39 years age group with 34.3% of the respondents, 30.5% of the tourists were then 30-39 years of age, while the remaining 5.3% of the tourists were over 60 years of age, indicating a tendency for young to adult tourists to visit Phuket Island more than older people. Percentages for number of visits including this trip, most of the tourists 58.2% are visited Phuket Island first time, least of the tourists 5.3% are visited Phuket Island more 5 times. The rest of the percentages were; 27.0% visited Phuket Island 2-3 times and 9.6% visited Phuket Island 4-5 times, and with companions or family (91.7%).

4.2. Assessment of Perceptions of Tourists of Safety and Security in Phuket.

Table 2. Perceptions of Tourists of Safety and Security in Phuket.

Perception of safety and security	*N*	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Mean	S.D.
Touring city in daytime	399	1.25	6.02	33.58	37.34	21.80	3.72	0.91
The amount of lighting at night	399	0.75	10.03	32.08	40.35	16.79	3.62	0.90
Accessibility of the destination	397	1.01	6.30	33.75	38.79	20.15	3.70	0.89
Tourist spots and places of visit	395	1.01	5.82	26.08	42.78	24.30	3.83	0.89
Transportation	399	2.51	7.77	28.07	42.86	18.80	3.67	0.95
Traffic conditions	398	4.52	11.06	37.69	32.66	14.07	3.40	1.00
Quality roads, bridges, utility services	398	2.26	11.31	33.92	38.19	14.32	3.51	0.94
Accommodation area	398	0.75	5.03	25.63	42.71	25.88	3.87	0.87
Fair shopping offer	399	0.75	8.27	30.58	38.85	21.55	3.72	0.91
Beachlifeguard and services	398	4.27	12.31	33.92	33.17	16.33	3.44	1.03
Water safety	396	2.27	11.11	32.83	33.59	20.20	3.58	1.00

Table 2 shows the percentages of a five-point Likert perceived scale for each category and the average rating factors influencing the perception of tourists on safety and security in Phuket. The results showed that the 3 factors that influence the highest perception of tourists include Accommodation area factor had an average score of 3.87, Tourist spots and places of visit factor had an average score of 3.83, and Touring city in daytime factor with Fair shopping offer factor both had an average score of 3.72. The results showed that the 3 factors that influence the Lowest perception of tourists include Quality tourism factors (Quality roads, bridges, utility services) had an average score of 3.51, Beach lifeguard and services factor had an average score of 3.44 and Traffic conditions factor had an average score of 3.40 respectively. Analysis of the highest averages found that influence of factors involving accommodation area affected the most affected safety and security concerns in Phuket. Tourists have commented that accommodation areas are very safe in Phuket. However, tourists feel that traffic conditions are unsafe in Phuket and gave traffic safety a very low score.

4.3. Hypotheses Testing

4.3.1. Nationalities

As aforementioned, a series of hypotheses were developed to investigate the influence of national culture on tourists' perception of safety and security while visiting the Phuket Island. Each of these hypotheses was developed to examine the difference in perceptions of safety and security while visiting the Phuket Island between more than two cultural groups of tourists. To test these hypotheses, one-way ANOVA analysis of variance was used. The results of ANOVA analysis of factors composing tourists' perception of safety and security while visiting the Phuket Island are provided in Table 3. The Asian category included people from the Middle East and the Europe category included people from Russia.

Table 3. Perceived Tourism Impacts with Different Nationalities of Respondents

Perception of safety and security	Mean					ANOVA	
	Asia	Australia/ Oceania	Europe	North America	Other	F value	p value
Touring city in daytime	3.50	3.84	3.81	3.80	3.65	1.939	0.103
The amount of lighting at night	3.56	3.61	3.74	3.50	3.57	0.825	0.509
Accessibility of the destination	3.58	3.82	3.74	3.80	3.61	1.054	0.379
Tourist spots and places of visit	3.71	3.91	3.95	3.73	3.72	1.475	0.208
Transportation	3.37	3.90	3.80	3.60	3.57	4.208	0.002*
Traffic conditions	3.51	3.53	3.52	3.31	2.99	4.246	0.002*
Quality roads, bridges, utility services	3.45	3.52	3.60	3.73	3.31	1.664	0.157
Accommodation area	3.78	4.06	3.91	3.70	3.79	1.718	0.144
Fair shopping offer	3.53	3.91	3.78	3.50	3.69	2.396	0.049*
Beach lifeguard and services	3.53	3.52	3.49	3.53	3.17	1.664	0.157
Water safety	3.77	3.53	3.56	3.70	3.43	1.264	0.283

* $p < 0.05$

The results shown in the table above divide into two groups based on the variation of their scores dependency on national culture variable. Within the first group a statistically significant difference was found in tourists' perception of transportation: F = 4.208, traffic conditions: F = 4.246, and fair shopping offer: F= 2.396, for all national groups. These results support the H1 of hypotheses 1 and reject the Null

Hypothesis (H0) of hypothesis 1. Which shows that the tourists from the Oceania and Europe countries felt safer with touristic activities than those from the Asian countries based on the observed average rating (Mean) of factors influencing the perception of tourists on safety and security in Phuket. In the other group, there was no statistically significant difference found in tourists' perception of touring city in daytime: F = 1.939, the amount of lighting at night: F = 0.825, accessibility of the destination: F = 1.054, tourist spots and places of visit: F = 1.475, quality roads bridge utility services: F = 1.664, accommodation area: F = 1.718, beach lifeguard and services and: F = 1.664, and water safety: F = 1.264. These results support the H0 of hypotheses 1 and rejected the alternative H1 of hypotheses 1.

4.3.2. Number of Visits including this Trip

Table 4. Perceived Tourism Impacts with different Number of Visits of Respondents

Perception of safety and security	Mean				ANOVA	
	First time	2-3 times	4-5 times	> 5 times	F value	p value
Touring city in daytime	3.76	3.79	3.61	3.33	1.827	0.141
The amount of lighting at night	3.71	3.64	3.45	3.00	4.655	0.003*
Accessibility of the destination	3.73	3.74	3.61	3.52	0.564	0.639
Tourist spots and places of visit	3.89	3.94	3.66	3.00	7.390	0.000*
Transportation	3.71	3.72	3.79	3.00	3.971	0.008*
Traffic conditions	3.44	3.44	3.42	2.86	2.240	0.083
Quality roads, bridges, utility services	3.61	3.49	3.32	2.86	4.894	0.002*
Accommodation area	3.97	3.85	3.82	3.24	4.702	0.003*
Fair shopping offer	3.74	3.80	3.63	3.38	1.394	0.244
Beach lifeguard andservices	3.56	3.33	3.40	3.00	2.745	0.042*
Water safety	3.70	3.49	3.42	3.10	3.342	0.019*

* $p < 0.05$

Result shown in the table above divide into two groups based on the variation of their scores dependency on Number of visits including this trip. Within the first group the tourists visited Phuket Island first time and 2-3 times a statistically significant difference was found in tourists' perception of The amount of lighting at night: F = 4.655, Tourist spots and places of visit: F = 7.390, Transportation: F =3.971, Quality tourism (Quality roads, bridges, utility services): F = 4.894, Accommodation area: F = 4.702, Beach lifeguard and services: F = 2.745 and Water safety: F = 3.342, for all Numbers of Visits including this trip. These results support the H0 of hypothesis 2 and reject the alternative H1 of hypothesis 2. In the other group, there was no statistically significant difference found in tourists' perception of Touring city in daytime: F = 1.827, Accessibility of the destination: F = 0.564, Traffic conditions: F = 2.240, and Fair shopping offer: F = 1.394. These results support the H0 of hypothesis 2 and reject the alternative H1 of hypothesis 2.

5. CONCLUSIONS

Understanding the perceived safety and security in tourism by the international visitors in relation to their demographic factors is important for tourism-driven economy in the region of Phuket. Our study disclosed that there were statistically significant difference in the perceptions of different nationalities, and the number of visits (including this trip) of respondents influencing the perception of tourists on safety and security in Phuket. The survey shows that the majority of the tourists from Oceania and Europe countries felt safer with touristic activities than those from Asian countries. The majority of the participants were visiting Phuket for the first time, and with companions or family. Therefore, this research has contributed to the tourism management for improving safe and secured environments to enhance the quality of hospitality and tourism for supporting and accommodating the tourist industries.

6. ACKNOWLEDGMENT

The authors would like to thank the Faculty of Technology and Environment, Prince of Songkla University, Phuket Campus, for their financial support and provision of hardware and software. The authors also thank Associate Professor Raymond James Ritchie for his assistance in proof reading. Also, the authors would like to thank the Phuket international airport for their kind cooperation during data collection.

REFERENCES

S. Jariyachamsit, "An Investigation of Safety in Tourism: An Experience of Young Tourists in Bangkok, Thailand," *Procedia - Soc. Behav. Sci.*, vol. 197, pp. 1931–1935, 2015.

C. Chen, S. Chen, H. Lee, and T. Tsai, "Exploring destination resources and competitivenesseA comparative analysis of tourists' perceptions and satisfaction toward an island of Taiwan," *Ocean Coast. Manag.*, vol. 119, pp. 58–67, 2016.

C. Chen, Y. R. Lai, J. F. Petrick, and Y. Lin, "Tourism between divided nations : An examination of stereotyping on destination image," *Tour. Manag.*, vol. 55, pp. 25–36, 2016.

S. Forgas-coll, R. Palau-saumell, J. Sánchez-garcía, and L. J. Callarisa-fiol, "Urban destination loyalty drivers and cross-national moderator effects : The case of Barcelona," *Tour. Manag.*, vol. 33, pp. 1309–1320, 2012.

Y. Boylu, A. D. A. Tasci, and W. C. Gartner, "Education ' s Impact on Cultural Distance Perception : The Case of Turkish Service Providers and European Guests," *Sustain. Qual. Life through Tour.*, pp. 23–27, 2003.

J. B. Garau-vadell, P. Martı, and S. Campo-martı, "Factors influencing repeat visits to a destination : The influence of group composition," *Tour. Manag.*, vol. 31, pp. 862–870, 2010.

T. H. E. Push-pull and M. Framework, "The 'Pull' of Tourism Destinations : A Means-End Investigation," *J. Travel Res.*, vol. 40, pp. 385–395, 2002.

T. Thiumsak and A. Ruangkanjanases, "Factors Influencing International Visitors to Revisit," *J. Econ. Bus. Manag.*, vol. 4, no. 3, 2016.

T. K. Hui, D. W. Ã, and A. Ho, "Tourists ' satisfaction , recommendation and revisiting Singapore," *Tour. Manag.*, vol. 28, pp. 965–975, 2007.

M. G. Larry Yu, "A comparative analysis of international tourists ' satisfaction in Mongolia," *Tour. Manag.*, vol. 27, pp. 1331–1342, 2006.

R. George, "Tourist's perceptions of safety and security while visiting Cape Town," *Tour. Manag.*, vol. 24, pp. 575–585, 2003.

K. A. Boakye, "Tourists' views on safety and vulnerability. A study of some selected towns in Ghana," *Tour. Manag.*, vol. 33, no. 2, pp. 327–333, 2012.

A. F. Amir, M. N. I. Ismail, and T. P. See, "Sustainable Tourist Environment: Perception of International Women Travelers on Safety and Security in Kuala Lumpur," *Procedia - Soc. Behav. Sci.*, vol. 168, pp. 123–133, 2015.

N. Kontogeorgopoulos, "Ecotourism and Mass Tourism in Southern Thailand : Spatial Interdependence , Structural Connections , and Staged Authenticity structural connections , and staged authenticity," *GeoJournal*, vol. 61, pp. 1–11, 2005.

P. Ramseook-Munhurrun, V. N. Seebaluck, and P. Naidoo, "Examining the Structural Relationships of Destination Image, Perceived Value, Tourist Satisfaction and Loyalty: Case of Mauritius," *Procedia - Soc. Behav. Sci.*, vol. 175, no. 230, pp. 252–259, 2015.

W. G. Zikmund, *Business Research Methods*, 8th ed. United State of America: South-Western Collage, 1994.

S. Amir, M. M. Osman, S. Bachok, and M. Ibrahim, "ScienceDirect Sustaining local community economy through tourism: Melaka UNESCO world heritage city," *Procedia Environ. Sci.*, vol. 28, pp. 443–452, 2015.

M. Barker and S. J. Page, "Visitor safety in urban tourism environments: The case of Auckland, New Zealand," *Cities*, vol. 19, no. 4, pp. 273–282, 2002.

S. Lertputtarak, D. Lobo, and T. Yingyong, "Identification of the Factors that Impact Russian Tourists in Thailand," *Procedia - Soc. Behav. Sci.*, vol. 144, pp. 133–142, 2014.

J. G. Choi, T. Tkachenko, and S. Sil, "On the destination image of Korea by Russian tourists," *Tour. Manag.*, vol. 32, no. 1, pp. 193–194, 2011.

G. Sinclair-Maragh, D. Gursoy, and M. Vieregge, "Residents' perceptions toward tourism development: A factor-cluster approach," *J. Destin. Mark. Manag.*, vol. 4, no. 1, pp. 36–45, 2015.

P. Brunt and P. Courtney, "Host perceptions of sociocultural impacts," *Ann. Tour. Res.*, vol. 26, no. 3, pp. 493–515, 1999.

B. King, A. Pizam, and A. Milman, "Social impacts of tourism. Host perceptions," *Annals of Tourism Research*, vol. 20, no. 4. pp. 650–665, 1993.

L. Reid and A. Boyd, "The Social Impacts of Tourism and Their Effects on Attitudes toward a Major Cultural Attraction," *Tourism : Building Credibility for a Credible Industry*. pp. 123–133, 1991.

J. C. Liu and T. Var, "Resident attitudes toward tourism impacts in Hawaii," *Annals of Tourism Research*, vol. 13, no. 2. pp. 193–214, 1986.

P. Kauppila and T. P. Karjalainen, "A process model to assess the regional economic impacts of fishing tourism: A case study in northern Finland," *Fish. Res.*, vol. 127–128, pp. 88–97, 2012.

C. G. Q. Chi and H. Qu, "Examining the structural relationships of destination image, tourist satisfaction and destination loyalty: An integrated approach," *Tour. Manag.*, vol. 29, no. 4, pp. 624–636, 2008.

D. Jeffrey and Y. Xie, "The UK market for tourism in China," *Ann. Tour. Res.*, vol. 22, no. 4, pp. 857–876, 1995.

A. Sada, M. Raheem, and A. Saleem, "Factors Affecting the Quality of Services of the Tourism Companies in Jordan from Tourists Perspective," *Br. J. Econ. Financ. Manag.*, vol. 5, no. September, pp. 92–112, 2012.

F. Bretz, T. Hothorn, and P. Westfall, "Multiple Comparisons Using R." Chapman & Hall/CRC, London, pp. 1–166, 2011.

P. Murrell, *R Graphics*. Chapman & Hall/CRC, London, 2006.

R Development Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org. 2015

Forecasting Foreign Tourist Arrivals to Bali Using Bayesian Vector Autoregression

I Wayan Sumarjaya*
Udayana University, Denpasar, Bali – sumarjaya@unud.ac.id

Abstract

Bali is one of the major tourist destinations in Indonesia. Although, in general, there is an increasing trend in the number of tourist visits, the policy maker often needs to forecast the visits for some period. The aim of this article is to forecast tourist visits from the selected ten countries in Asia Pacific region such as Australia, India, Japan, South Korea, China, New Zealand, Taiwan, Malaysia, Singapore and Thailand. Forecasting multivariate time series allows researcher to see dynamic relationship between variables and to improve forecast accuracy. The author uses Bayesian vector autoregression to model and to forecast this tourist visits and then compares the forecasting results with classical vector autoregression. The author concludes that the Bayesian vector autoregression performs better than classical vector autoregression in forecasting foreign tourist arrivals.

Keywords: Bayesian vector autoregression; Forecasting foreign tourist, Tourist demand modelling.
JEL Classification: C53; L83; Z39.

1. INTRODUCTION

As a major tourist destination in Indonesia, Bali has become popular icon for tourism. Although, in general, there is an upward trend in the number of tourist arrivals, the policy maker often needs to predict the number of these visits. Having information about tourist visits forecasting can help government in tourism planning (Cho, 2001) .

Methods that are often used in forecasting tourist visits are time series analyses. Time series analyes are large class of methods that deal with dependent data which is ordered sequentially. Popular quantitative methology for forecasting time series in tourism can be found in Song and Witt (2000). One of the popular methology in forecasting multivariate time series is vector autoregression. Books such as Lutkepohl and Kratzig (2004), Lütkepohl (2005) and Tsay (2014) discuss in greater detail vector autoregression and its extension. These books mainly discuss classical vector autoregression and only briefly treat Bayesian autoregression.

This article is organized as follows. Section 1 gives motivation for research about tourism forecasting. Section 2 gives basic concept of classical vector autoregression. Section 3 discusses basic concepts of prior and its role in bayesian vector autoregression. In section 4 we forecast tourist visits from ten major countries and compare the forecasts using Bayesian vector autoregression and classical vector autoregression. Section 5 concludes the article.

2. BASIC CONCEPTS OF VECTOR AUTOREGRESSION

In this section we briefly discuss the concept of classical vector autoregression (VAR) as a natural extension of univariat autoregression. The readers are refer to books such as Lütkepohl (2005) and Tsay (2014) for further discussion about vector autoregression and its extension.

Let y_t for $t = 1,\ldots,T$ be a K x 1 vector observation (also called endogenous) on K time series variable, u_t be a K x 1 vector of disturbances, α_0 be a K x 1vector of intercepts and A_i be a K x K matrix of coefficients. Vector autoregressive of order P, denoted VAR(P), can be written as

$$y_t = \alpha_0 + \sum_{i=1}^{p} A_i y_{t-i} + u_t. \tag{1}$$

An important properties of VAR(P) process is stability. This stability ensures the generation of stationary time series with time invariant means, variance and covariances structure (Pfaff, 2008) and can be check by evaluating the characteristic polynomial

$$\det(I_K - A_1 z - \cdots A_p z^p) \neq 0 \tag{2}$$

for $|z| \leq 1$. Further properties of the stability of VAR(P) can be seen in Lütkepohl (2005) and Lutkepohl (2009).

The next step in modelling classical VAR is model specification. This step involves choosing lag order either via sequential testing or model selection criteria (Lutkepohl, 2009). Three information criteria that are commonly used are Akaike information criterion of the form

$$\text{AIC}(m) = \log\det(\hat{\Sigma}_m) + \frac{2}{T} mK^2 \tag{3}$$

Hannan-Quinn criterion of the form

$$\text{HQ}(m) = \log\det(\hat{\Sigma}_m) + \frac{2\log\log T}{T} mK^2 \tag{4}$$

and Schwarz information criterion of the form

$$\text{SIC}(m) = \log\det(\hat{\Sigma}_m) + \frac{\log T}{T} mK^2 \tag{5}$$

where $\hat{\Sigma}_m$ is the residual covariance matrix estimator for a model of order m. After proper lag has been choosen then the VAR(P) model candidate is estimated. Methods for estimating VAR(P) can be done via multivariate least square estim1ation, least square estimation with mean-adjusted data and Yule-Walker estimation, and maximum likelihood estimation. Details about these estimation methods can found in Chapter 3 of Lütkepohl (2005).

The final step in modelling using classical VAR is forecasting. If the process y_t is assummed to generated by a known VAR(P) process then the conditional expectation of y_{T+h} given y_t, for $t \leq T$ then $y_{T+h|T} = E(y_{T+h} \mid y_T, y_{T-1}, \ldots)$ where $y_{T+j|T} = y_{T+j}$ for j $\leq$ 0 If the white noise process u_t is independent and identically distributed then $y_{T+h|T}$ is the optimal mean square error h ahead forecast in period T (Lutkepohl, 2013).

3. BASIC CONCEPTS OF BAYESIAN VECTOR AUTOREGRESSION

Bayesian vector autoregression incorporates prior information in the vector autoregression model in (1). Before discussing prior for Bayesian vector autoregression we note another useful representation of VAR (Koop & Korobilis, 2010)

$$y = (I_K \otimes X)\alpha + u \tag{6}$$

where $u \sim N(0, \Sigma \otimes I_K)$, $\alpha = \text{vec}(A)$, $A = (\alpha_0, A_1, \ldots, A_p)$, $x_t = (1, y'_{t-1}, \ldots, y'_{t-p})$ and $X = (x_1', \ldots, x_1')$. Priors for vector autoregression in (6) have been suggested by many authors. See Geweke and Whiteman (2006) and Karlsson (2013). In this section we briefly discuss the priors for Bayesian VAR.

The VAR form in (6) can be viewed as function of parameters. Thus (6) can be breaks into two parts that is

$$\begin{aligned} \alpha \mid \Sigma, y &\sim N(\hat{\alpha}, \Sigma \otimes (X'X)^{-1}), \\ \Sigma^{-1} \mid y &\sim W(S^{-1}, T - K - M - 1). \end{aligned} \tag{7}$$

Here $\hat{\alpha} = \text{vec}((X'X)^{-1} X'Y)$ and $S = (Y - X(X'X)^{-1} X'Y)'(Y - X(X'X)^{-1} X'Y)$. If Σ is replaced by an estimate, then we have the Minnesota prior. The Minnesota prior assumes $\alpha \sim N(\tilde{\alpha}, \tilde{V})$. Variations on the Minnesota prior can be found in Karlsson (2013).

A natural conjugate prior for Bayesian VAR is Normal Wishart prior which has the form

$$\begin{aligned} \alpha \mid \Sigma &\sim N(\tilde{\alpha}, \Sigma \otimes \tilde{V}) \\ \Sigma^{-1} &\sim W(\tilde{S}^{-1}, \tilde{v}) \end{aligned} \tag{8}$$

with $\tilde{\alpha}, \tilde{V}, \tilde{v}$ and $\tilde{s}$ prior hyperparameters choosen by researcher (Koop & Korobilis, 2010). The advantage of conjugate prior is that analytical results exist. However, according to Koop and Korobilis (2010) the conjugate prior in (8) has several drawbacks. First, it assumes that every equation must have the same set of predictor variables. Second, the prior covariance of the coefficients in any two equations must be proportional to each other.

An extension of prior in (8) is independent Normal-Wishart prior which has the form

$$\begin{aligned} p(\beta, \Sigma^{-1}) &= p(\beta) p(\Sigma^{-1}), \\ \beta &\sim N(\tilde{\beta}, \tilde{V}_\beta), \\ \Sigma^{-1} &\sim W(\tilde{S}, \tilde{v}). \end{aligned} \tag{9}$$

In order to obtain the posterior a Gibbs sampler can be sequentially drawn from the Normal $p(\beta \mid y, \Sigma)$ and the Wishart $p(\Sigma^{-1} \mid y, \beta)$. Details of Gibbs sampling algorithm for (9) can be found in Karlsson (2013).

The last prior we consider in this section is the stochastic search variable selection (SSVS) which specifies a hierarchical prior of the form

$$\alpha_i \mid \gamma_i \sim (1-\gamma_i) N(0, \upsilon_{0i}^2) + \gamma_i N(0, \upsilon_{1i}^2) \tag{10}$$

with γ_i is the dummy variable and υ_{0i}^2 and υ_{1i}^2 denote the first and the second prior variance. Gibbs sampler algorithm can be found in Karlsson (2013).

4. RESULTS DAN DISCUSSION

In this section we discuss and compare tourist visits forecasting using classical vector autoregression and Bayesian vector autoregression. We use monthly foreign tourist visits to Bali from ten countries from January 2010 to December 2015 and use the real 2016 tourist visits to measure the forecast accuracy.

Figure 1 shows time series plot of ten major tourist visits to Bali for period January 2010 to December 2015. These countries are Australia (AU), China (CN), India (IN), Japan (JP), South Korea (KR), Malaysia (MY), New Zealand (NZ), Singapore (SG), Thailand (TH) and Taiwan (TW). It can be seen from the figure that the number of visits clearly fluctuated. Countries such as Australia, China, India, South Korea, Malaysia, New Zealand and Singapore clearly show an increasing trend although seasonal fluctuation apparent in the plot. On the other hand, countries such as Japan, South Korea, Thailand and Taiwan do not seem to have trend, but seasonal fluctuation is evidence.

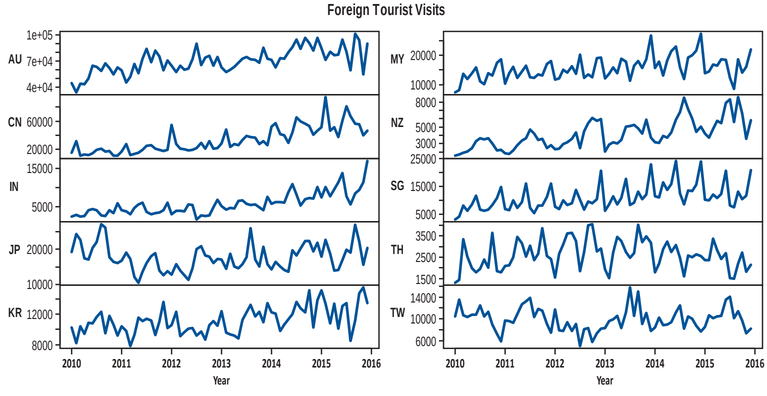


Figure 1. Foreign Tourist Visits to Bali During 2010—2015.

4.1. Forecasting using Classical Vector Autoregression

The first step in vector autoregression model building is lag order determination. Using selection procedure in JMulTi software (Lutkepohl & Kratzig, 2004) we obtained the number of optimal lag (up to 10 lags of levels) is 5 based on Akaike Information Criterion (AIC) value as suggested by Ivanov and Kilian (2005). The next step is diagnostic checking. Since the main aim of this research is forecasting then we need to check for the stability of the estimated VAR. Stability analysis for the estimated VAR(5) model using CUSUM test suggests that the process is stable. Forecasts results can be seen in Table 1.

Tabel 1. Forecasts of Tourist Visits for 2016 Using VAR(5)

Month	AU	CN	IN	JP	KR	MY	NZ	SG	TH	TW
Jan.	60787	35679	6591	15578	14311	-2721	1418	-502	575	12783
Feb.	52521	62704	11738	19868	10730	17718	-882	1176	989	6147
Mar.	64493	18182	8949	7565	4619	12680	1719	5490	1395	9639
Apr.	64656	23510	7130	2460	7102	2046	2009	-236	2444	9670
May	60281	32113	9903	9688	6021	7132	4275	-5092	1683	10291
Jun.	75219	11563	8667	10581	7761	2248	6611	10883	2115	15672
Jul.	86218	52558	6703	20101	11330	3318	5913	-5709	1903	17494
Aug.	37418	24146	4302	14882	10252	-4910	4304	-10432	1665	13657
Sep.	84713	4690	5360	15207	7532	6285	5562	4655	1711	12958
Oct.	74300	13987	5080	16368	12702	1148	3314	-9536	1785	9417
Nov.	16144	-14628	2961	7478	9917	-3329	-1329	-9798	552	9914
Dec.	55164	-8290	9832	16735	6097	13884	-450	7571	1715	5322

It is interesting to note that some forecasts for tourist visits from China, Malaysia, New Zealand and Thailand yield negative values.

4.2. Forecasting using Bayesian Vector Autoregression

In the Bayeasian vector autoregression setup we used Minnesota, independent Normal Wishart, Normal Wishart and stochastic search variable selection-in mean Wishart (SSVS) priors. We followed the recommendation regarding the lag selection based on study by Ivanov and Kilian (2005) and fitted lag 4, 6 and 12 to the model. However, fitting to lag order 6 and 12 resulting in the almost singular matrix. Thus we only fitted the model with lag 4. Due to the space limitation we only provide forecasting results using the Minnesota prior (see Table 2). Forecasting using other priors (not shown) also yields positive values, unlike the classical VAR which yield some negative values (see Table 1).

Tabel 2. Forecasts of Tourist visits for 2016 Using Bayesian Vector Autoregression with Minnesota Prior

Month	AU	CN	IN	JP	KR	MY	NZ	SG	TH	TW
Jan.	95640	75300	9000	24470	14400	16020	8660	18450	3200	9180
Feb.	89730	82440	11120	18410	14220	11390	8730	10930	3760	11580
Mar.	89670	92810	11600	21820	14580	7930	9510	17770	2880	13110
Apr.	92590	74600	12210	18800	13930	15010	9600	13080	2400	12930
May	97490	77120	11870	22900	15720	17040	8280	17290	2280	12530
Jun.	105370	106360	9440	19040	14860	12320	7800	5860	2550	9770
Jul.	101430	82400	10040	17520	15570	13610	8360	13660	2980	11890
Aug.	104060	74640	12170	22720	13780	9300	9400	10420	2770	11320
Sep.	98040	82170	8250	28260	13920	13300	9840	16260	2120	9820
Oct.	83110	70300	9760	15410	12950	7440	8480	6410	2720	14820
Nov.	96670	68070	10140	17230	13280	11620	9490	18430	1700	12130
Dec.	104610	80960	8490	17410	13340	10830	8620	9360	3660	11060

4.3. Comparison of Forecast Accuracy

In order to compare the forecast accuracy we calculated the root mean square error (RMSE) values (see Table 3).

Table 3. RMSE for Classical VAR (CVAR) and Bayesian VAR (BVAR)

Country	RMSE				
	CVAR	BVAR Minnesota	BVAR Independent Normal Wishart	BVAR Normal Wishart	BVAR SSVSMW
AU	37285.189	11513.641	16930.4403	15210.381	15680.1958
CN	63977.982	19219.345	28072.2403	31546.728	21218.8565
IN	9189.912	6664.187	5973.1744	4593.356	7487.8439
JP	8176.757	4157.194	4229.2596	4602.462	4616.4044
KR	3908.838	2544.438	3515.0635	2478.208	2056.6416
MY	11854.278	4688.825	3441.2360	10538.412	4978.4722
NZ	4589.735	2814.771	**5099.4678**	2333.398	2291.0393
SG	13351.872	6326.706	4233.0952	3938.209	3914.5570
TH	1272.557	1074.831	916.7197	**2048.448**	903.9514
TW	3101.909	1952.664	**3643.8197**	**8966.263**	**4010.0970**

As can be seen from Table 3 the Bayesian VAR using Minnesota prior yields smaller RMSE than the classical VAR. Other RMSE for Bayesian VAR using independent Normal Wishart, Normal Wishart and stochastic search variable selection-in mean Wishart show almost similar results with the exception for Taiwan, Singapore and Thailand (highlighted in bold).

5. CONCLUSIONS

In general we obtained that in general forecasting tourist visits using Bayesian vector autoregression yields smaller RMSE than classical vector autoregression. Unlike classical vector autoregression which produce some negative forecasts, Bayesian vector autoregression with priors yields only positive forecasts.

REFERENCES

Cho, V. (2001). Tourism forecasting and its relationship with leading economic indicators. *Journal of Hospitality and Tourism Research,* **25,** 399-420.

Geweke, J. & Whiteman, C. (2006). Bayesian Forecasting. in *Handbook of Economic Forecasting* Eds. G. Elliot, C.W.J. Granger & A. Timmermann. New York: Elsevier.

Ivanov, V. & Kilian, L. (2005). A Practitioner's Guide to Lag Order Selection for VAR Impulse Response Analysis. *Studies in Nonlinear Dynamics and Econometrics,* **9,** 1-34.

Karlsson, S. (2013). Forecasting with Bayesian Vector Autoregression. in *Handbook of Economic Forecasting*: Elsevier.

Koop, G. & Korobilis, D. (2010). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Foundations and Trends in Econometrics,* **3,** 267-358.

Lutkepohl, H. (2009). Econometric analysis with vector autoregressive models. in *Handbook of Computational Econometrics* Eds. D.A. Belsley & E.J. Kontoghiorghes. New Jersey: John Wiley and Sons.

Lutkepohl, H. (2013). Vector autoregressive models. in *Handbook of Research Methods and Applications in Empirical Macroenomics* Eds. N. Hashimzade & M.A. Thornton. Cheltenham, UK: Edward Elgar.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Lutkepohl, H. & Kratzig, M. Eds. (2004). *Applied Time Series Econometrics*. Cambridge: Cambridge University Press.

Pfaff, B. (2008). VAR, SVAR and SVEC Models: Implementation within R Package vars. *Journal of Statistical Software,* **27,** 1-31.

Song, H. & Witt, S.F. (2000). *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches*. Amsterdam: Pergamon.

Tsay, R.S. (2014). *Multivariate Time Series Analysis with R and Financial Applications*. New Jersey: Wiley.

Do FTAs Promote Trade? Evidences from ASEAN Countries' Bilateral and Regional FTAs

Justin Lim Ming Han*[1]
Bank Negara Malaysia, Kuala Lumpur, Malaysia – limminghan@bnm.gov.my

Abstract

Trade integration in East Asia is characterized by a proliferation of ASEAN-centric bilateral and regional FTAs. Using a modified gravity model spanning from 1988 to 2015 and across 60 countries, I examine the effects of these FTAs on trade flows. Results indicate that trade creation effects differ between ASEAN's regional and bilateral FTAs. Since their inception, ASEAN's regional FTAs had, on average, increased trade by at least 30%, three times greater than its bilateral FTAs. Delving deeper, however, I find that most gains are limited to the ASEAN-China FTA, in particular, trade in final goods. Moreover, the impacts of AFTA and the remainder regional FTAs on trade are inconclusive. Hence, paying greater attention towards harmonizing and deepening trade among the ASEAN members, rather than creating new FTAs, may be more effective at promoting trade integration in the region.

Keywords: AFTA, ASEAN+ FTAs, Gravity model, Trade creation
JEI Code: F13, F15

1. INTRODUCTION

This paper is motivated by the following observations. Globally, by 2016, more than 267 bilateral and multilateral free trade agreements (FTAs) are in force compared to just 40 in the early 1990s. Trade integration is especially pronounced in the East Asian region. The number of RTAs involving East Asian economies increased exponentially from less than 10 in 2000 to 67 by 2016. Beginning with the ASEAN Free Trade Agreement (AFTA) in 1992, the regionalization process further gained momentum with the signing of 5 bilateral FTAs with its major regional partners: China (2005), South Korea (2006), Japan (2008), Australia and New Zealand (2008), and India (2009). This trend culminated into the emergence of mega-FTAs in recent years, such as the Trans-Pacific Partnership (TPP) and the Regional Comprehensive Economic Partnership (RCEP). Moreover, this regionalisation process evolved alongside the proliferation of bilateral FTAs, totalling 34 and counting, among the 16 countries examined in this paper.

Several theories explain this global trend[2]. For the ASEAN-centric FTAs, however, a few observed that their formation departs from the early regionalism predecessors, i.e. European Community Treaty (EC), North American Free Trade Agreement (NAFTA), or the Southern Common Market (MERCUSOR). It was observed that some ASEAN countries pursue a dual-track integration approach by choosing to selectively deepen via bilateral FTAs, regionalize, or both[3] (Kleimann, 2014). The rationale for doing so is to alleviate the lowest common denominator dilemma inherent in collective bargaining approach – the scope and depth of policies mutually agreed upon decreases as the number of participant increases.

1 The author is an economist from the Economics Department of Bank Negara Malaysia. I wish to express my gratitude to Steven C M Wong, Witada Anukoonwattaka, Tng Boon Hwa and Mohamad Hasni Sha'ari for their invaluable feedback and helpful comments. The views expressed here are of the author and do not represent those of Bank Negara Malaysia. Any errors and omissions are the sole responsibly of the author.

2 The first is slow multilateralism (Krugman (1991), Bhagwati (1991)), prompting countries to regionalize due to WTO's slow progress of multilateralising trade liberalisation. Some attributed regionalism to increasing trade interdependence, stemming from the development of global value chains (Hamanaka (2012), Baldwin (2006)). Others argued that competing political interests and strong lobbyist groups to be the driving force behind FTA formation (Kleimann (2014), Ravenhill (2010), Grossman and Helpmans (1994)).

3 For instancem the dual track approach is illustrated by Japan's bilateral FTA with Malaysia, Brunei, Philippines, Thailand, Singapore and Vietnam while maintaining the ASEAN-Japan Comprehensive Economic Partnership (AJCEP).

Early trade integration proponents, however, have generally considered FTAs to be second best compared to non-discriminatory liberalisation (Viner, 1950), potentially hampering a more efficient global trading system if excessively utilized (Bhagwati, 1991). In the ASEAN setting, economic gains were considered to be main driver behind the formation of FTAs (Hamanaka, 2012 and Baldwin, 2006). On the other hand, Kleimann (2014) and Ravenhill (2011) argued that these FTAs are politically motivated and hence bear little or no implication on trade flows. Moreover, the continued proliferation of FTAs is not without costs. The "spaghetti bowl" effect of FTAs – overlapping and intertwining tariff structures and overlapping Rules of Origins (ROOs) – do not augur well for businesses which then face increases in administrative and compliance costs (Baldwin, 2006 & Bhagwati, 1991).

As a result, empirical evidence focusing on the effect of AFTA on trade has been mixed[4], at best. Hence, given the confluence of factors affecting East Asia's regionalism, the objective of this paper is to evaluate the effectiveness of all ASEAN-centric FTAs. A series of dummies measures the trade effects of FTAs based on the micro-founded gravity model[5], using annual bilateral import data among 60 countries from 1988 to 2015.

I make the following contributions. Firstly, to examine the trade implications of the dual track strategy, I model the bilateral and regional FTAs in the region. I then compare the performance of the ASEAN regional FTAs against other major CUs and FTAs while distinguishing between AFTA and its regional FTAs. Next, I examine the interaction between the bilateral and regional FTAs. I also consider the effects of each ASEAN+ FTA separately. Lastly, I distinguish between trade in intermediate and final goods. Most empirical studies neither fully specify the related ASEAN FTAs nor provide sufficient economic interpretations of their findings. This study is a first attempt to do so comprehensively for the ASEAN-centric bilateral and regional FTAs.

The key findings are as follows. Results indicate that trade creation effects differ between regional and bilateral FTAs. Since their inception, ASEAN's multilateral FTAs had, on average, increased trade by at least 30%, three times greater than its bilateral FTAs. Evidence of trade gains for ASEAN countries that simultaneously regionalize while selectively deepening via bilateral FTAs is not forthcoming. Delving deeper, however, I find that most gains are limited to the ASEAN-China FTA, in particular, trade in final goods. Thus, the results imply that ASEAN-China FTA did not give way to greater development of production networks in the region, but rather in enhancing market access for final goods. Moreover, the impacts of AFTA and the remainder ASEAN+ FTAs on trade are found to be inconclusive. Broadly speaking, these findings suggest that ASEAN's regionalization endeavours are potentially hindered by the spaghetti bowl FTA conundrum and the lowest common denominator dilemma inherent in the collective bargaining strategy. Nevertheless, the findings corroborate with the view that ASEAN's external orientation via its bilateral and multilateral FTAs appear to be more effective than its intra-regionalization efforts.

The rest of the paper is organized as follows: Section 2 describes the data and methodology used. Section 3 discusses the empirical findings and Section 4 concludes with policy discussions.

2. METHODOLOGY AND DATA

A gravity model typically estimates trade flows as a function of : i) *country-specific time-varying* variables, i.e.: exporter's and importer's GDP; ii) *country-pair time-invariant* variables i.e.: bilateral distance between both economies and binary indicators which assigns a value 1 if both countries share a common border, language or land border and zero otherwise, and; iii) *country-pair time-varying* indicators, i.e.: a value 1 is assigned to a FTA indicator if both economies are part of a particular free trade agreement k at time t and zero otherwise. However, this specification ignores the micro-founded price indices and is subjected to omit variable biases. Bilateral trade does not only depend on bilateral trade costs,

4 Studies by MacPhee and Sattayanuwat (2014), Trotignon (2010) show statistical significance for the AFTA dummy, while Magee (2008) and Elliott and Ikemoto (2004) show otherwise.

5 See Anderson and van Wincoop's (2003) for derivations of a priced-based gravity model.

but also on each partner's average trade costs with the rest of the world - multilateral trade resistance[6] (Anderson and van Wincoop, 2003). What matters is the relative trade barrier. Hence, the time-varying multilateral trade resistance terms are accounted for using *country-and-time fixed effects,* as proposed by Baier and Bergstrand's (2007), under a panel setting[7]. As such, exporter and importer GDP variables can no longer be estimated. Secondly, the *country-pair fixed effects* are necessary because of unobserved time-varying country-pair heterogeneity that is likely to be correlated with the FTA dummies[8]. Alternatively, as proposed by Baier and Bergstrand (2007) and Magee (2008), this issue can also be addressed through first-differencing the panel data[9]. Thus, bilateral time invariant variables and indicators such as distance, common official language, coloniser and border are also dropped from the regression. Hence, only *country-pair time-varying* indicators, for instance, preferential tariff rates, are captured via the RTA dummies[10]. Following their recommendations, the gravity model takes the following logarithmic form:

$$\text{dlnX}_{ij,t-(t-1)} = \beta_0 + \sum_{i=1}^{60}\sum_{t=1988}^{t=2015} \alpha_{i,t-(t-1)} + \sum_{j=1}^{60}\sum_{t=1988}^{t=2015} \alpha_{j,t-(t-1)} + \beta_1 dRTA_{\text{ij},t-(t-1)}{}^{k} + v_{ij,t-(t-1)} \qquad (1)$$

Where $\text{dlnX}_{ij,t-(t-1)}$ is the first-differenced natural logarithm of imports between country *i* and *j*. $\text{dRTA}_{ij,t-(t-1)}$, dummy takes value 1 if both countries are in the same free trade agreement (FTA) *k* and 0 otherwise in year *t*. A positive sign indicates there is intra-regional trade expansion[11]. With 27 years in the series, there are a total of 26 time periods *t*-(*t*-1). Since there are 60 country-pairs, the model introduces 1620 (=60×27) *country-and-time fixed effect* dummies for both exporters and importers to account for the changes in the time-varying multilateral trade resistance terms, denoted by $\sum_{i=1}^{60}\sum_{t=1988}^{t=2015}\alpha_{i,t-(t-1)}$ and $\sum_{j=1}^{60}\sum_{t=1988}^{t=2015}\alpha_{j,t-(t-1)}$ respectively. $v_{ij,t-(t-1)} = \varepsilon_{ij,t} - \varepsilon_{ij,t-1}$ is white noise. The list of FTAs modelled are shown below in Table 1.

Table 1: List of Bilateral and Regional ASEAN FTAs

FTA dummies	Abbreviations[12]
All regional and bilateral FTAs globally	DallFTAs
NAFTA, MERCUSOR, EC, ANDEAN, EFTA	DmajorFTAs
AFTA	DAFTA
ASEAN + Regional FTAs	DASEAN+ FTAs
All ASEAN member bilateral FTAs with key partners	DbilateralASEANFTAs
ASEAN - China FTA	DASEAN + China
ASEAN - Japan FTA	DASEAN + Japan
ASEAN - Korea FTA	DASEAN + Korea
ASEAN - India FTA	DASEAN + India
ASEAN - Australia - New Zealand FTA	DASEAN + ANZ

Gross import data (zero trade flows excluded) is obtained from the World Integrated Trade Solution (WITS) under the Standard International Trade Classification (SITC) Revision 3 for 60[13] countries from 1988 to 2015. The definitions of intermediate and final goods are adapted from Athukorala (2011) at the 3-digi SITC level.

6 Thus, bilateral trade is a function of three types of trade costs: (i) bilateral trade costs between country *i* and *j*, (ii) exporting country *i*'s average trade cost to all countries - inward multilateral resistance, and (iii) country *j*'s average trade cost to all countries - outward multilateral resistance.

7 These variables cannot be observed directly and Anderson and van Wincoop (2003) proposed two methods to estimate them. The alternative approach is based on non-linear least square (NLS) estimation, which is the most efficient method since it utilizes all available information. However NLS is rarely used.

8 For instance, the extent of regulatory reforms (unobservable) due to the potential sizable welfare gains may induce countries to sign FTAs. Hence, FTAs and the extent regulatory reforms may be positively correlated; FTAs dummy is then negatively correlated with the error term, ε, and thus tend to be underestimated.

9 Alternatively, a dummy for each country-pair αij can be introduced (Baier and Bergstrand, 2007).

10 The impact of MFN tariffs, however, is accounted for via the *country-and-time fixed effects.*

11 Here we distinguish between the classical notions of the Vinerian (1950) trade creations effects on welfare, since an increase in bilateral trade is not necessarily welfare improving. Discussion on welfare implications are not further examined in this paper. Instead, trade creation captures the effect of FTAs on trade flows.

12 D prefix refers to dummies in first-differenced panel data.

13 The 60 countries examined cover 90% of global trade. They consists of ASEAN members and other major economies:

3. DISCUSSION OF RESULTS

As shown in Table 2, Model 1 is the baseline gravity model estimated using ordinary least squares (OLS) based on equation (1). The *DallFTA* dummy is found to be insignificant at the 10% level. This finding departs from the empirical literature, but may be the result of having included many ASEAN-centric FTAs that were not previously examined.

Thus, model 2 distinguishes between: i) *DmajorFTAs* – the major CUs and FTAs globally; ii) *DAFTA* – AFTA, and; iii) *DASEAN+ FTAs* – all five regional FTAs. Firstly, the *DmajorFTAs* dummy shows significant positive trade creation effects, although the estimate is smaller compared to the findings in the empirical literature[14]. Secondly, the *DAFTA* dummy does not show any statistical significance - as shown by Magee (2008) despite the rise in intra-region trade shares from 17% in 1990 to 25% in 2015[15]. While not exhaustive or exclusive, I consider four types of explanations for the lack of evidence for trade creation in AFTA. The first is its extremely low utilisation of AFTA's preferential rates (Zhang & Shen, 2011). As alluded to by Bhagwati (1991), the result of the spaghetti bowl effect - multiple preferential tariffs rates and overlapping rules of origins (ROOs) requirements, may had dampened the benefits of regionalisation and hence their overall utilization. Secondly, the promotion of free trade or export processing zones and import duty exemptions also diminishes AFTA's regional trade expansion role (Menon, 2013). Thirdly, the number of non-tariff measures in ASEAN has trended upwards while tariffs were reduced simultaneously (Ing et al, 2016), implying that effective trade barrier have remained relatively unchanged despite on-going regionalization initiatives. Fourthly, the result corroborates with the lowest common denominator dilemma inherent in regional agreements embedded in the collective bargaining strategy – the scope and depth of policies mutually agreed upon decreases as the number of participant increases (Kleimann, 2014).

Table 2: First-Differenced Panel Gravity Equations With Country-and-Time Effects

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
DallFTAs	0.042 (0.04)					
DmajorFTAs		0.120** (0.055)	0.120** (0.055)	0.121** (0.053)	0.098 (0.064)	0.096 (0.063)
DAFTA		-0.111 (0.151)	-0.111 (0.151)	-0.112 (0.148)	-0.170 (0.164)	-0.105 (0.155)
DbilateralASEANFTAs			0.090** (0.0423)	0.091** (0.043)	0.023 (0.029)	0.054 (0.038)
DASEAN+FTAs * bilateralFTA			-0.023 (0.0148)	-0.006 (0.0153)		
DASEAN+FTAs		0.26*** (0.059)	0.27*** (0.060)			
DASEAN+CHINA				0.41*** (0.095)	0.0517 (0.085)	0.122* (0.07)
DASEAN+JAPAN				-0.008 (0.081)	-0.007 (0.091)	-0.008 (0.081)
DASEAN+KOREA				0.0847 (0.107)	0.207** (0.086)	0.0848 (0.107)
DASEAN+INDIA				-0.028 (0.109)	-0.0511 (0.114)	-0.0281 (0.110)
DASEAN+ANZ				-0.034 (0.190)	0.0250 (0.104)	-0.0341 (0.190)
Constant	0.0578 (0.075)	0.0577 (0.075)	0.0549 (0.075)	0.0537 (0.075)	0.40*** (0.096)	0.0555 (0.075)
Number of observations	72,920	72,920	72,920	72,920	69,820	72,444
R-squared	0.138	0.139	0.139	0.139	0.124	0.123
Number of country-pairs	3,299	3,299	3,299	3,299	3,271	3,295

Note: Robust standard errors in parentheses
*** Indicates that the coefficients are statistically significant at 1% ,
** indicates that thecoefficients are statistically significant at 5% ,
* indicates that the coefficients are statistically significant at 10%

14 In comparison, studies by Magee (2008) and Baier & Bergstrand (2007) show higher trade creation effects for FTAs due to the different time period investigated and the number of countries investigated.

15 Clausing (2001), however, cautions that trade share is not indicative of trade creation since it is highly correlated with exchange rate movements.

By contrast, the result for *DASEAN+FTAs* implies that, on average, an ASEAN+ FTA increases trade by 30% (=$e^{0.26}$-1). The result is even more striking given that the ASEAN+ FTAs are relatively newer compared to the original AFTA and does not appear to be affected by the spaghetti bowl FTA conundrum.

Model 3 further introduces the *DbilateralASEANFTAs* and its interaction term with *DASEAN+FTAs* to fully examine the dual track integration hypothesis by Kleimann (2014). Compared to the ASEAN regional FTAs, the impact on trade under the bilateral mode is still positive and significant but smaller at 10% (=$e^{0.09}$-1). However, the interaction term is insignificant and hence ASEAN countries that simultaneously regionalize while selectively deepening trade ties do not appear to have materially benefitted from this strategy. Nevertheless, the findings in model 2 and 3 lend evidence to the notion that gains from ASEAN's trade endeavours are largely attributable to its external orientation rather than from amongst each other.

Model 4 provides disaggregate results for the ASEAN regional FTAs in order to further examine the validity of the spaghetti bowl FTA phenomena. Out of the five, only the ASEAN-China FTA is statistically significant at the 1% level, resulting in an increase in trade flows by 50%. This result, however, should be interpreted alongside China's rapid trade expansion globally - the number of countries with China as a top-5 trading partner increased from only 2 in 1990 to 125 by 2015, of which many do not have a *de facto* trade promoting agreement with China. Moreover, this result is indicative of trade gains between ASEAN and China, rather than among ASEAN members. This is because the internal trade dynamics among ASEAN members remain unchanged, as shown by the *DAFTA* coefficient. As for ASEAN's remainder regional FTAs, the lack of evidence is possibility due to, in part, the aforementioned spaghetti bowl FTA conundrum. However, their trade potential may have yet to materialize due to the extended phase-in periods of the committed tariff reductions - 10 years or more (Baier & Bergstrand, 2007 and Magee, 2008).

Lastly, models 5 and 6 examine the impact of FTAs on intermediate and final goods respectively, as motivated by Yi (2003) who showed that tariff reductions have had large impact on intermediate goods trade. In model 5, only the *DASEAN+KOREA* dummy showed up as significant. However, I do not consider this finding to be robust given its lack of significance for overall trade in model 4. The absence of significant result for *DASEAN+CHINA* is also particularly puzzling given the region's pivotal role in global value chains, but may be attributable to the entrenched intermediate goods export rivalry between ASEAN and China (Holst and Wiess, 2004). On the contrary, model 6 shows an increase in final goods trade, although only at the 10% significance level. Thus, the results imply that ASEAN-China FTA did not give way to greater development of production networks in the region, but rather in enhancing market access for final goods.

4. CONCLUSION

Given the proliferation of ASEAN-centric FTAs over the past two decade, the purpose of this paper is to examine their impact on trade flows, which is nuanced by the multi-pronged trade integration strategies undertaken by ASEAN members and its key trading partners. The key findings are as follows. Results indicate that trade creation effects differ between regional and bilateral FTAs. Since their inception, ASEAN's regional FTAs had, on average, increased trade by at least 30%, three times greater than its bilateral FTAs. Also, evidence of trade gains for ASEAN countries that simultaneously regionalize while selectively deepening via bilateral FTAs is not forthcoming. Delving deeper, however, I find that most gains are limited to the ASEAN-China FTA, in particular, trade in final goods. Thus, the results imply that ASEAN-China FTA did not give way to greater development of production networks in the region, but rather in enhancing market access for final goods. The impact of AFTA and the remainder ASEAN+ FTAs on trade is found to be inconclusive. Broadly speaking, these findings suggest that ASEAN's regionalization endeavours are potentially hindered by the spaghetti bowl FTA conundrum and the lowest common denominator dilemma inherent in the collective bargaining strategy. Nevertheless, the findings corroborates with the view that ASEAN's external orientation via the bilateral and multilateral FTAs appears to be more effective than its intra-regionalization efforts.

Hence, paying greater attention towards harmonizing and deepening trade among the ASEAN members, rather than creating new FTAs, may be more effective at promoting trade integration in the region. The recent emergence of mega-FTAs may provide a pathway for the resolution of the aforementioned issues. Ideally, as described by the theorist (Viner, 2005, Bhagwati, 1991) and the policymakers (Menon, 2013), the better alternative to preferential arrangements is to unilaterally liberalise trade barriers for a more efficient global trading system.

I propose several approaches and issues that can be addressed in future studies. Firstly, Magee (2008) advocated the use of import quantities rather than values as import prices changes after FTAs are formed, which may underestimate the intra-regional trade expansion dummy. Secondly, initial conditions matter and the period of investigation can be extended backwards for robustness purposes. The third is the inclusion of the lagged and anticipatory effects of FTAs on trade, as explored by Baier and Bergstrand (2007) and Magee (2008). Lastly, the number of countries examined may also be expanded and a thorough discourse of the bilateral FTAs may yield additional insights into the dual-track integration strategy in the region.

REFERENCES

Anderson, J. & van Wincoop, E. (2003). Gravity with Gravitas: A Solution to the Border Puzzle. *American Economic Review,* 93(1) , 170-192.

Athukorala, P. (2011). Production Networks and Trade Patterns in East Asia: Regionalization or Globalization? *Asian Economic Papers*, Vol. 10, No. 1, Pages 65-95

Baier. S. & Bergstrand, J. (2007), Do free trade agreements actually increase members' international trade? *Journal of International Economics*, Volume 71, Issue 1, Pages 72-95.

Baldwin, R. (2006). Multilateralising Regionalism: Spaghetti Bowls as Building Blocs on the Path to Global Free Trade. *The World Economy*, Wiley Blackwell, vol. 29(11), pages 1451-1518.

Baldwin, R. (1993). A domino theory of regionalism. NBER Working Paper 4465.

Bhagwati, J. (1991). *The World Trading System at Risk*. Princeton, N.J.: Princeton University Press.

Elliot. R. & Ikemoto. K. (2004), AFTA and the Asian Crisis: Help or Hindrance to ASEAN Intra-Regional Trade? *Asian Economics Journal*, Vol 18, No. 1, pg 1-23

Grossman, G. and Helpman. E. (1995), The Politics of Free-Trade Agreements. *American Economic Review*, 85(4), pp. 667-90.

Hamanaka, S. (2010). Is Trade in Asia Really Integrating? ADB Working Paper Series on Regional Economic Integration.

Holst, D. and Weiss, J. (2004) ASEAN and China: Export Rivals or Partners in Regional Growth? *The World Economy*, 27 (8): 1255-74

Kleimann, D. (2014), Beyond Market Access?: The Anatomy of ASEAN's Preferential Trade Agreements. *Journal of World Trade*, Issue 3, pp. 629–682

Krugman, P. (1991). Increasing Returns and Economic Geography. *Journal of Political Economy, vol. 99(3)* , 483-99

MacPhee, C. & Sattayanuwat, W. (2014). Consequence of Regional Trade Agreements to Developing Countries. *Journal of Economic Integration*, Vol 29 (1), 66-94

Magee. C. (2008), New measures of trade creation and trade diversion. *Journal of International Economics*, Volume 75, Issue 2, Pages 349-362

Menon, J. (2013), Supporting the Growth and Spread of International Production Networks in Asia: How Can Trade Policy Help? Asian Development Bank Working Paper 112

Ravenhill, J. (2010), The "New East Asian Regionalism": A Political Domino Effect. *Review of International Political Economy*, vol. 17, no. 2, pp. 178-208.

Trotignon, J. (2010). Does Regional Integration Promote the Mulitilateralization of Trade Flows? A Gravity Model Using Panel Data. *Journal of Economic Integration*, 25 (2), 223-251

Viner, J. (1950) The Customs Union Issue. New York: Carnegie Endowment for International Peace

Yang, S. & Martinez-Zarzoso, I. (2014). A panel data analysis of trade creation and trade diversion effects: The case of ASEAN–China Free Trade Area. *China Economic Review*, vol. 29, p. 138-151.

Yi, K. (2003). Can Vertical Specialization Explain the Growth of World Trade? *Journal of Political Economy,* 111(1), 52-102.

Zhang, Y., and M. Shen. (2011), The Status of East Asian Free Trade Agreements. ADBI Working Paper 282. Tokyo: Asian Development Bank Institute.

Environmental and Social Assessment of Green Growth in Cairo Governorate (As an Indicator for the Quality of Life)

Hanan Mosad
Department of Geography, Faculty of Arts, Ain Shams University, Cairo, Egypt dr.hananmosad@yahoo.com
Mostafa Mohamed Salah
Department of Environment Statistics, Central Agency for Public Mobilization and Statistics (CAPMAS), Cairo, Egypt.
mostafasalah775@gmail.com

Abstract

Cairo is considered one of the main growing cities in the Arab world today. It is not only the largest city in the Arab world, but also the densest. Rapid population and economic growth of Cairo over recent decades have brought great challenges. Cairo consistently scores low on quality of life indexes, brought down by air pollution, congestion and education. It is worth mentioning that air pollution in Cairo is a matter of serious concern. Actually, Cairo is facing a conflict between economic growth and environmental conservation.

Consequently, green growth or environmentally sustainable economic growth is imperative in light of the current environmental pollution and risks, and the related health costs. One of the main aspects of importance to green growth includes the associated effects on human health and on environmental quality of life.

Quality of life is a multi-dimensional concept. There are two basic approaches to research on the quality of life: a subjective and an objective one. A subjective (or endogenous) approach focuses on feelings, perceptions, opinions and mental states of the individuals or groups studied. An objective (or exogenous) approach tries to do research on the quality of life based on a wide range of measurable or observable indicators in an individual and an environmental dimension. For green growth, one dimension of the quality of life stands out, the direct impact of the environment on people's quality of life.

In this work we aim to examine the relationship between green growth and related environmental quality of life in Cairo Governorate using a set of selected measurable indicators. The selected indicators are related to air and water pollution, the quality of housing and some attributes of the population, such as green growth features and their satisfaction with the surrounding environment and provision of housing utilities in terms of access to clean water and connection to sewage treatment.

The study deduced great variability in quality of life between informal areas (e.g. Manshiyat Nasser, El Haggana), semi-informal areas (El Marg, El Bassatien) and planned areas (e.g. Heliopolis, El Nozha, New Cairo).

Keywords: Green growth, Quality of life, Air quality, Water quality, Housing utilities.

1. INTRODUCTION

Cairo is the largest city in Egypt and the country's capital. Cairo is ´´a primate region´´ and has preserved its urban ascendancy over the last few decenniums. Cairo is located in the northern part of Egypt about 100 miles (165 km) from the Mediterranean Sea. It is also about 75 miles (120 km) from the Suez Canal. Cairo's total area is 175 square miles (453 sq km). Its metropolitan area, which includes nearby satellite cities, extends to 33,347 square miles (86,369 sq km).

Geographically, Cairo Governorate extends over the east bank of the Nile from El-Marg and El-Salam in the north and El-Nozha and Nasr City in the east to Helwan and El-Tebeen in the south and it is divided administratively into 37 districts.

Demographically, while the 2016 population of Cairo Governorate is 12 million; its metropolitan area has over 20 million people. It is the 3rd largest urban area in the Islamic World behind Jakarta and Karachi. Cairo has a population density of 19,376 people per square kilometer (50,180/sq. mi), which ranks 37th in the world. Moreover, as the economic center of Egypt, Cairo faces another problem regarding the daily regional mobility where thousands of commuters travel daily to and from Cairo causing more traffic congestion and air pollution.

Traffic emissions, such as from trucks, buses, and automobiles, also contribute. Airborne by-products from vehicle exhaust systems cause air pollution and are a major ingredient in the creation of smog in some large cities.

The air pollution in Cairo is a matter of serious concern. The air quality in downtown Cairo is more than 10 to 100 times of acceptable world standards. Cairo has a very poor factor because of lack of rain and its layout of tall buildings and narrow streets, which create a bowl effect (bad ventilation and consequent trapping of pollutants). The main air pollution problem in Egypt is the particulate matter. The most notable sources of the dust and small particles is transportation, industry and open-air waste-burning. Another significant source is the wind blown from arid areas around Egypt (e.g. Western Desert).

Cairo is also under what's commonly called "the black cloud": a dense, poisonous mass of smoke caused by the seasonal burning of rice ashes by local farmers.

Air pollution is a widespread problem in megacities. The main pollutant of concern is suspended particulate matter. In addition, the megacities experiencing high SPM levels usually also experience high sulphur dioxide levels. These pollutants can act synergistically, exacerbating health problems. Pollutants arising from vehicle emissions are also a major cause of poor air quality in megacities.

Seven of the megacities had three or more pollutants which exceeded WHO health protection guidelines, Mexico City, Beijing, Cairo, Jakarta, Los Angeles, Sao Paulo and Moscow. Mexico City was classified as having serious problems for sulphur dioxide, suspended particulate matter, carbon monoxide and ozone plus moderate to heavy pollution for lead and nitrogen dioxide.

The World Health Organization (WHO) released its 2014 report into global air pollution with some concerning finds. Egypt is one of the 10 most polluted countries in the world.

The average resident of Cairo breathes in more than 20 times the acceptable level of air pollution every day according to WHO. The growing number of cars, factories and power plants, and the use of old heating methods such as burning coal and wood are considered to be the main man-made sources of air pollution.

Traffic congestion is a serious problem in Cairo with large and adverse effects on both the quality of life and the economy. In addition to the time wasted standing still in traffic, time that could be put to more productive uses, congestion results in unnecessary fuel consumption, causes additional wear and tear on vehicles, increases harmful emissions lowering air quality, increases the costs of transport for business, and makes the GCMA an unattractive location for businesses and industry.

Cairo has seen rapid growth over the last few decades, in large part because of a huge gap between the birth rate (30 births/1,000 people) and the death rate (7 deaths/1,000 people.) Cairo struggles with many health issues, including serious air pollution. Today, Cairo has one of the highest air pollution levels in the world and is believed to cause 2% of all deaths[1].

Informal urban sprawl as a physical outcome of pseudo-urbanization has manifested in Cairo Region. In Cairo region more than 7 million inhabitants dwell in informal areas; 80 % are on privately owned arable lands (UNDP 2004). The annual population growth in Cairo's informal areas reached to 0.8 %. (Denis and Sejourne 2002). It is estimated that more than 1.5 million Acres of arable lands have been wasted because of the expansion of informal Sprawl areas over the past five decenniums.

2. RESEARCH INCENTIVES

2.1. Theoretical framework

The meaning of the phrase quality of life differs a good deal as it is variously used but, in general, it is intended to refer to either the conditions of the environment in which people live, (air and water

1 http://worldpopulationreview.com/world-cities/cairo-population (Accessed on 21.01.2017)

pollution, or poor housing, for example), or to some attribute of people themselves (such as health or educational achievement) (Pacione, 1982; Hills, 1995; Benzeval et al., 1995)[2].

The urban QoL concept gains more importance when it is considered that the world population is expected to reach somewhere between 7.6-9.4 billion, and the urban population is expected to reach 50% by the year 2020. In many countries, the increase of urban population is followed by economic growth which is usually complemented with environmental depletion such as pollution of water, air and other related problems[3]. Accordingly, Green growth or environmentally sustainable economic growth is imperative in light of the current environmental crises and resource depletion. Green Growth" has been used to cope with the current environmental challenges such as energy conservation and renewable generation, pollution and waste reduction and more efficient use of resources. It represents an opportunity to simultaneously improve people's quality of life[4].

2.2. Applied and Environmental dimension

A group of selected measurable indicators have been used as an attempt to examine the relationship between green growth and environmental quality of life within Cairo Governorate.

The indicators specified and measured and the research outputs constitute a referential framework for decision makers and provide analysis and guidance to promote and monitor a sustainable environmental development at the national level.

3. METHODOLOGY

The study depends mainly on two approaches, descriptive and analytical approach. Descriptive approach was used to mainly clarify air and water quality both qualitatively and quantitatively. Moreover, analytical approach was considered to break the research problem down into elements, in order to analyze accurately each element.

An exploratory study was conducted as a starting point to collect data about the topic of the research. A stratified random sample of 250 was surveyed in three selected areas representing the most common housing patterns in Cairo Governorate. The main purpose of the questionnaire was to measure respondents' satisfaction with the quality of life in their neighborhoods, presence of green features in each neighborhood and degree of access to housing utilities.

SPSS was used as a useful tool for entering, processing and analyzing statistical data collected by questionnaires.

4. DATA SOURCES

The research depends on different data sources including official statistics and field study. The official statistics include CAPMAS publications, 2006 census, and annual report on environmental statistics for the year 2014. Moreover, an exploratory field study was conducted in December 2016.

5. INDICATORS RELATED TO PHYSICAL ENVIRONMENT

5.1. Air pollution

According to WHO Global Urban Ambient Air Pollution Database 2016, more than 80 % of people living in urban areas that monitor air pollution are exposed to air quality levels that exceed the World

---

2 Pacione, M.,(1982). The use of objective and subjective measures of quality of life in human geography. Prog. Hum. Geogr. 6 (4), pp.495–514.
Hills, J., (1995). Inquiry into Income and Wealth, vol. 2. Joseph Rowntree Foundation, York.
Benzeval, M., Judge, K., Whitehead, M., (1995). Tackling Inequalities in Health. Kings Fund, London.

3 Saufi, N., et al.,(2016), Green Growth and Corporate Sustainability Performance, Procedia Economics and Finance (35), pp. 374-378, p. 374.

4 Saufi, N., et al., 2016, p. 375.

Health Organization (WHO) limits. While all regions of the world are affected, populations in low-income cities are the most impacted.

According to the latest urban air quality database, 98% of cities in low- and middle income countries with more than 100 000 inhabitants do not meet WHO air quality guidelines. However, in high-income countries, that percentage decreases to 56%[5].

Air pollution can harm human health and the environment. the consequences of air pollution on public health are measured not only in terms of sickness the consequences of air pollution on public health are measured not only in terms of sickness and death, but also in terms of lost productivity and missed educational and other human development opportunities and death, but also in terms of lost productivity and missed educational and other human development opportunities[6]. The Six major types of materials that released directly into the atmosphere in their unmodified forms and in sufficient quantities to pose a health risk are carbon monoxide, hydrocarbons, particulates, sulphur dioxide, and nitrogen compounds and lead[7].

Egyptian environmental affairs agency (EEAA) is the main responsible for monitoring air pollutants in Egypt through two networks. The first is designed to monitor suspended dust and lead in ambient air over Cairo. The second is the National network for monitoring industrial emissions, and it covers the rest of the governorates of the country especially industrial areas. The latter is one of the basic tools to collect data and information on emissions quality and its quantity and the analysis of these data contributes effectively to the control process in industrial pollution, and also contribute in giving a clear picture of air quality in places located nearby those companies through analysis of pollution loads. Moreover, there is another national network to monitor air quality over residential areas.

Fig. (1) presents Annual average concentrations of SO_2, NO_2 and PM_{10} in Cairo Governorate during the period (2010-2014). The figure reveals that over the five years the concentrations of PM_{10} exceed the air quality limit value*.

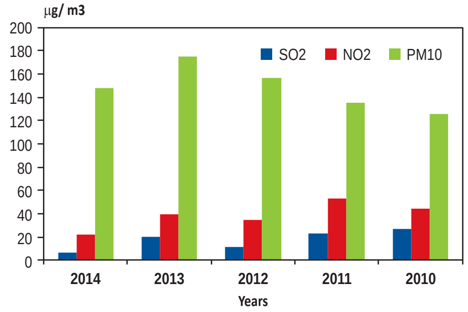


Figure 1. Annual Averages Concentrations for (SO2) (NO2) and (PM10) for Cairo During the Period (2010-2014)

It is worth mentioning that PM10 is the major air pollution problem in Egypt. Annual average concentrations of PM10 range between 100 and 200 µg/m3 in urban and residential areas and between 200 and 500 µg/m3 near industrial areas[8]. The concentration levels of PM10 have also been observed to exceed the Air Quality Limit values in El Fayoum Governorate followed by El Gharbiya Governorate

5 http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/ (Accessed on 25.01.2017).

6 United Nations (UN) – Department of Economic and Social Affairs, Population Division. 2001. World Population Monitoring 2001 – Population, environment and development. New York: UN.

7 Enger, E. D. & Smith, B. F. (2000). Environmental science: A study of interrelationships (7th ed.). Boston, Mass: McGraw-Hill.

* Air quality limit value of PM_{10} is 70 $\mu g/m^3$

Air quality limit value of NO_2 is 40 $\mu g/m^3$

Air quality limit value of SO_2 is 40 $\mu g/m^3$

8 Bjarne Sivertsen, B., & Aboud El Seoud, A.,(2004), The air pollution monitoring network for Egypt, Norwegian

where big industrial zones are found such as Mahalla and Tanta. Such concentrations enforce the need to develop a warning system and a strong emphasis on improving the management of air quality based on well-prepared policies and strengthened institutional arrangements toward reducing PM10 concentrations in Egypt as a whole and in Cairo in particular.

The concentration levels of SO2 have also been observed to exceed the Air Quality Limit values during the years 2010 and 2011. SO2 concentrations exceeding the Egypt air quality limit values occurs in the city center of Cairo and downwind from industrial areas[9].

5.2. Water pollution

The Nile is the predominant source of fresh water in Egypt. It represents 73% of the water renewable resources. The second sources are perception along the Northern coast, ground water and wastewater treatment. Due to the high costs involved, desalination plays only a minor role in supplying water in Egypt. Unfortunately, the Nile is subject to unsustainably and shockingly high levels of industrial, agricultural and domestic wastewater pollution.

Egypt has been suffering from severe water scarcity in recent years. Given the value of water poverty is 1000 m^3, Egypt has entered the era of water poverty where per capita fresh water decline to 76,9 m^3 in 2014 and it is expected to decline further by passing of time.

It is worth mentioning that pollution and environmental degradation are decreasing water availability in the country.

Water of Nile is being polluted by municipal and industrial waste, with many recorded incidents of leakage of wastewater, the dumping of dead animal carcasses, and the release of chemical and hazardous industrial waste into the river.

Industrial waste has led to the presence of metals in the water which pose a significant risk not only on human health, but also on animal health and agricultural production.

It is worth mentioning that the main water pollutants in Egypt are TDS, COD, BOD and DO.

Although water quality limit value of TDS is 500 mg/L, figure (2) reveals that concentrations of such material in the water in Cairo Governorate are within acceptable limits. Such low concentrations can be attributed to the absence of the primary source of TDS in Cairo Governorate which is agricultural wastewater. Contrariwise, Concentrations of COD, BOD and DO exceeded minimum allowed limit[*10] in Cairo Governorate during the period (2011-2014) as shown in figure (3).

At the national level, the concentration levels of COD have also been observed to exceed the minimum allowed limit in El Fayoum Governorate followed by El Qalioubya Governorate. It ranges between 26,9 and 31,9 mg/L. On the other side, COD concentrations recorded the lowest level in Luxor and Minya Governorates, where they range between 5 to 6 mg/L.

The concentration levels of BOD have also been observed to exceed the minimum allowed limit in El Qalioubya and Cairo Governorates. It ranges between 6,5 and 7,4 mg/L. Contrarily, BOD concentrations recorded the lowest level in Suez, where it recorded 1,5 mg/L.

The concentration levels of DO have also been observed to exceed the minimum allowed limit in three governorates in Upper Egypt which are Sohag, Minya and Asyout. It ranges between 7,7 and 9 mg/L. In contrast, Alexandria and two governorates of Lower Egypt recorded values lower than permissible limit. It ranges between 4 and 5,9 mg/L.

9 Bjarne Sivertsen, B., & Aboud El Seoud, A., (2004), The air pollution monitoring network for Egypt, Norwegian Institute for Air Research (NILU), Kjeller, Norway.

10 * Minimum allowed limit of TDS is 500 mg/L
Minimum allowed limit of COD is 10 mg/L
Minimum allowed limit of BOD is 6 mg/L
Minimum allowed limit of DO is 6 mg/L

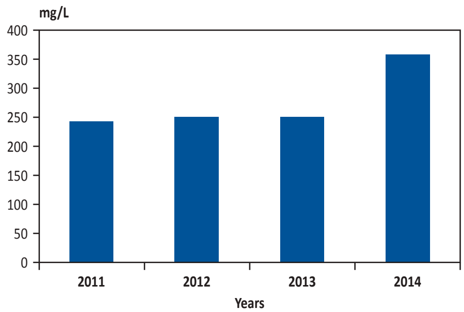


Figure 2. Annual Averages for Concentrations of TDS in the Nile River in Cairo Governorate During the Period (2011-2014)

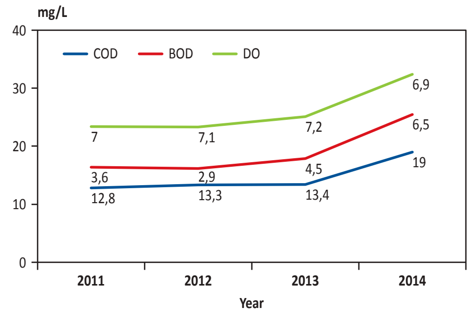


Figure 3. Annual Averages for Concentrations of COD,BOD and DO in the Nile River in Cairo Governorate During the Period (2011-2014)

6. HOUSING AND PERSONAL INDICATORS (CASE STUDY)

On the basis of the sample social survey questionnaire and the statistical analysis of the data, it can be concluded that quality of life and green growth vary strongly among the different housing patterns in Cairo Governorate, namely informal areas (such as Ezbet El Haggana and Manshiet Nasser), semi informal areas (such as El Marg and El Basateen) and planned areas (such as El Nozha, Heliopolis and Madinet Nasr) where $\alpha = 0.05$. Figure 4, 5 and 6 demonstrate the results of the survey as follows:

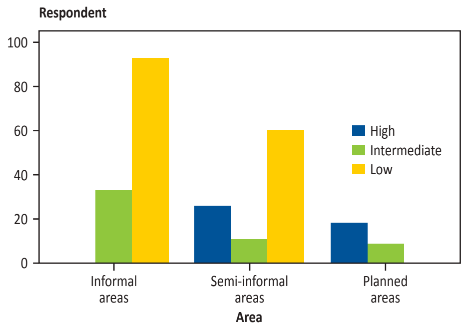


Figure 4. The Degree of QoL Provided by the Government

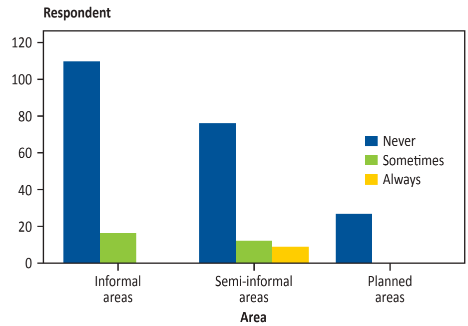


Figure 5. Presence of Green Growth Features

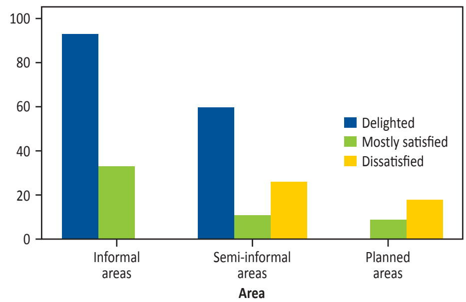


Figure 6. Degree of Satisfaction with the Official Availability of Housing Utilites e.g. (safe drinking water and sewage)

It can be summarized from the previous statistical and diagrammatic presentation that informal and semi-informal areas in Cairo Governorate characterized by low degree of quality of life which positively correlates with disappearance of green growth features and high degree of satisfaction with the modest provided housing utilities.

7. CONCLUSIONS AND RECOMMENDATIONS

The results of analyzing physical environment indicators reveal that PM_{10} is the major air pollution problem in Egypt as a whole and in Cairo in particular. Its concentrations exceed Air quality limit value. For water quality, it was observed that concentrations of COD, BOD and DO exceeded minimum allowed limit. Such results enforce the need to direct the academic research towards addressing green growth indicators as a starting point towards reducing the potential health risk from such pollutants. The study also recommends promoting awareness of protecting environment especially in urban areas. Studying housing and personal indicators confirmed that the lower quality of life, the lower the features of green growth from one hand, and the lower accessibility to housing utilities. Land use planning is recommended to achieve high standard of living.

REFERENCES

Benzeval, M., Judge, K., Whitehead, M., (1995). Tackling Inequalities in Health. Kings Fund, London.

Bjarne Sivertsen, B., & Aboud El Seoud, A., (2004), The air pollution monitoring network for Egypt, Norwegian Institute for Air Research (NILU), Kjeller, Norway.

Enger, E. D. & Smith, B. F. (2000). Environmental science: A study of interrelationships (7th ed.). Boston, Mass: McGraw-Hill.

Hills, J., (1995). Inquiry into Income and Wealth, vol. 2. Joseph Rowntree Foundation, York

Pacione, M., (1982). The use of objective and subjective measures of quality of life in human geography. Prog. Hum. Geogr. 6 (4), pp.495–514.

Saufi, N., et al., (2016), Green Growth and Corporate Sustainability Performance, Procedia Economics and Finance (35), pp. 374-378, p. 374.

United Nations (UN) – Department of Economic and Social Affairs, Population Division. (2001). World Population Monitoring 2001 – Population, environment and development. New York: UN. http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/(Accessed on 25.01.2017). http://worldpopulationreview.com/world-cities/cairo-population (Accessed on 21.01.2017)

CPS25:
Macrofinancial Statistics (4)

Maximum Likelihood Estimation of Non-stationary Variance

Jetsada Laipaporn*
Prince of Songkla University, Pattani, Thailand – laipaporn.j@gmail.com
Phattrawan Tongkumchum
Prince of Songkla University, Pattani, Thailand – phattrawan@gmail.com

Abstract

The general autoregressive conditional heteroscedasticity – GARCH (1, 1) model is widely used for estimating non-stationary variance of financial time series, but it gives results that need smoothing in order to show structural changes over the period of interest. Its estimation also requires iterative optimization of a function involving derivatives that need to be computed numerically, and is not guaranteed to converge. We consider an alternative method using maximum likelihood to estimate parameters in a natural cubic spline function. This method gives similar results to GARCH (1, 1) when applied to exchange rates of currencies in term of special drawing right (SDR) for Thailand, Singapore, Malaysia and Indonesia over the last two decades, and simulation studies suggest that it is more accurate than GARCH (1, 1).

Keywords: Heteroscedasticity; natural cubic spline function; GARCH (1, 1); exchange rates.
Journal of Economic Literature (JEL) Classification: C13; C14; C22; G15

1. INTRODUCTION

The non-stationary variance or heteroscedastic condition is critical issue in financial modeling. With respect to this condition, most of financial analysts and economists apply General Autoregressive Conditional Heteroscedasticity or GARCH model introduced by Bollerslev (1986) to estimate the financial volatility. The analysts usually apply the GARCH model to daily returns by using maximum likelihood estimation. Unfortunately these estimates of daily volatility are fluctuated and unclear to illustrate the fundamental movement of the non-stationary variance.

Generally, the natural cubic splines (NCS) function is practical to extract the non-stationary variance, because the continuous piecewise cubic polynomials is efficient in capturing the flexible trend among noisy data (see Hastie *et al.*, 2009). However, Fitting the NCS function to the estimates of the GARCH model will be a good estimate of the non-stationary variance, if only the GARCH model is satisfactory to exhibit the behavior of the observed returns variation.

Therefore, this study proposes an alternative approach to estimate the non-stationary variance of the exchange rates in term of special drawing rights (SDR) for 4 currencies comprised of Thai baht, Singapore dollar, Malaysia ringgit and Indonesia rupiah, and the generated returns with known signal from Monte Carlo simulation by using a maximum likelihood estimation to fit a NCS function directly to the absolute returns and compares to NCS fits to the GARCH (1,1) which is adopted as the representative of the GARCH model.

The next parts of this paper are organized as follows. The second section informs the term of exchange rate data used for calculating returns and how to generate returns from the Monte Carlo simulation. The third section provides methodologies of two approaches to estimate non-stationary variance. The estimated results are reported and discussed in the fourth section. The last section concludes this study.

2. DATA

Two kinds of returns series are employed. The first one is the returns from the exchange rates which are officially used in the international market. The second one is generated by Monte Carlo simulation. The details of these data are described as follows.

2.1. Exchange Rate

Basically, many literatures employ the effective exchange rate to investigate the impact of its volatility on the national economy rather than using the bilateral exchange rate (for detail discussion, see Clark *et al.*, 2004 and McKenzie, 1999), because the effective rate efficiently reflects the uncertainty of the national currency pricing. The effective exchange rate is the average of a national currency relative to an index or basket of the bilateral exchange rate weighted by the trade volumes. These trade volumes indicate the relative importance of that each bilateral rate to the valuation of the national currency (for detail, see Turner *et al.*, 1993).

Calculating the effective exchange rate is comprehensive and need sufficient data for producing the index. However, there is an alternative way to price the national currency. Eugenio (2016) introduces to use the commodity's price in term of the Special Drawing Rights units (SDR). Hence, the SDR is the weighted average of major national currencies traded in the world market. The price in term of SDRs subsequently becomes the international price of that commodity. By this concept, the exchange rates of national currency in term of SDRs are employed as the international price relative to the major currencies of that national currency.

This paper employs the exchange rate of four national currencies in term of SDRs per currency unit which are obtained from the website of the International Monetary Fund or IMF (http://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx comprise of Indonesia rupiah (IDR), Malaysian ringgit (MYR), Singapore dollar (SGD) and Thai baht (THB) during January 3rd 1994 to December 30th 2016. These exchange rates are plotted in Figure 1.

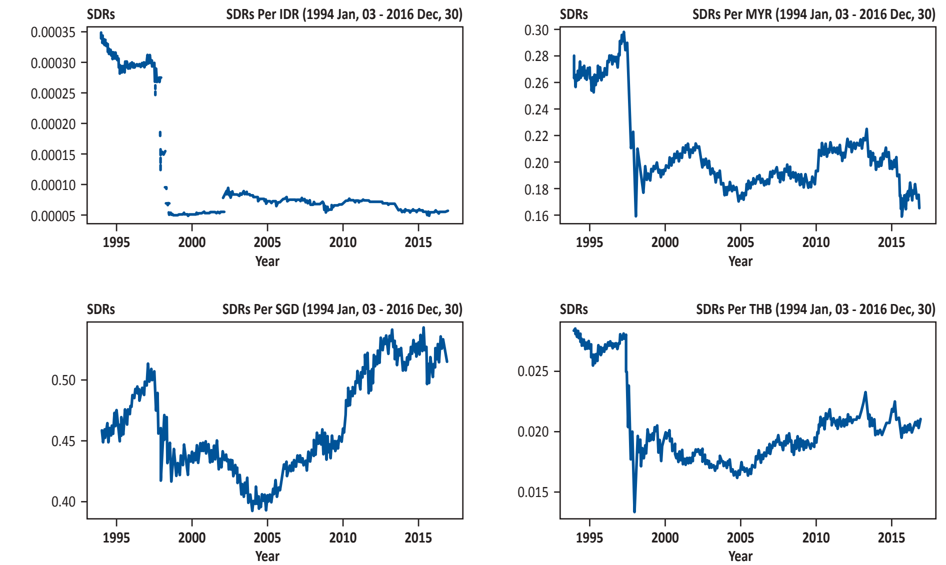


Figure 1. The Exchange Rate of Indonesia Rupiah (IDR), Malaysian Ringgit (MYR), Singapore Dollar (SGD) and Thai Baht (THB) During 1994-2016

The daily variation of the exchange rate is measured by daily returns (u_t) which is a rate of the proportional increasing or decreasing of the exchange rate from preceding day as this following equation,

$$u_t = \frac{E_t - E_{t-1}}{E_{t-1}} \qquad (1)$$

E_t and E_{t-1} are exchange rate on the present day and the preceding day, respectively. Figure 2 graphs the returns of Thai baht, Singapore dollar, Malaysia ringgit and Indonesia rupiah during given period.

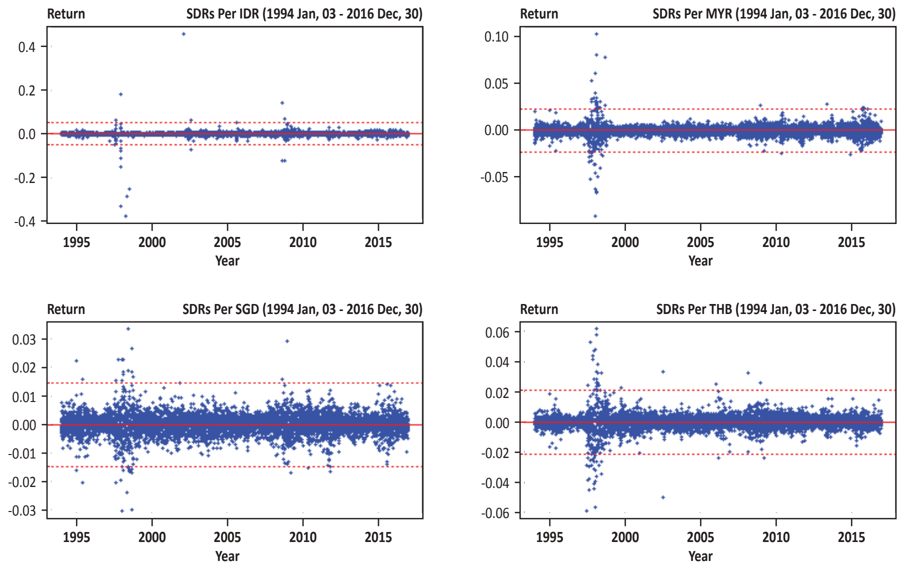


Figure 2. Returns on one Trading Day of Indonesia Rupiah (IDR), Malaysian ringgit (MYR), Singapore Dollar (SGD) and Thai Baht (THB) During 1994-2016

2.2. Generated Returns from Monte Carlo Simulation

The generated returns are simulated following the stylist facts of financial asset returns as discussed in Cont (2001) and Engle (2003). The generated returns are separated into two group based on different assumption.

The first group comprises of four datasets of the returns. Each dataset contains 5,000 random data with zero mean and inconstant variance with typical known signal. These returns are consequently generated as

$$u_t=\mu+s_t z_t \tag{2}$$

The u_t is a return on day t which is a function of expected return (μ) and residual as white noises (z_t) with a known signal (s_t). Since the simulated returns have zero mean, so the expected return (μ) is equal to zero.

The second group is assumed the same as the first group, except the white noises (z_t) become the fat-tailed residual (ε_t) following this equation

$$u_t=\mu+s_t\varepsilon_t \tag{3}$$

The fat-tailed residuals (ε_t) are the transformation of white noises (z_t) on day t. This transformation is to stretch the tails of the white noise's distribution between two critical points, -c and c with the stretching factor equal to a. The transformation is followed this formula

$$\varepsilon_t=\begin{cases} c+a(z_t-c) & ,c > z_t \\ z_t & ,-c < z_t < c \\ -c+a(z_t+c) & ,z_t < -c \end{cases} \tag{4}$$

c and a are equal to 1.25 and 2, respectively.

This study applies four known signals (s_t) represented four different movement of non-stationary variance in the long-run. These signals are constructed as a function of time (t) as shown in Figure 3, consisted of constant (s_0), linear (s_1), quadratic (s_2), and cubic (s_3) signals.

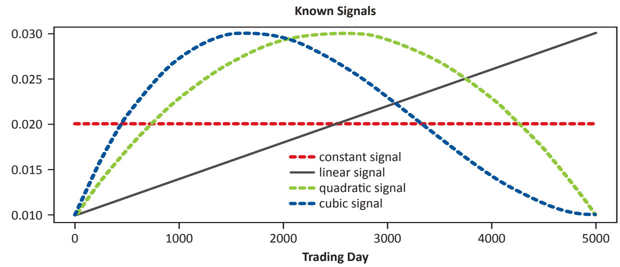


Figure 3. The Four Types of Known Volatility Signal (s_t)

3. METHODOLOGY

This paper utilizes two approaches for estimating the non-stationary variance using NCS function. The first approach is fitting NCS function to the estimated volatility of the GARCH (1, 1) model. Meanwhile, the second approach is fitting NCS function directly to the absolute value of the observed returns. The details of each approach are described as follows.

3.1. NCS Fits to GARCH (1, 1) Estimates.

This approach begins with estimating the daily volatility by using the GARCH (1, 1) model and then fitting NCS function to the estimated volatility for non-stationary variance.

The GARCH (1, 1) assumes that the observed returns (u_i) are independent and the distribution of returns is normal with mean equal to zero and conditional on its variance σ_i^2. This model is usually parameterized as

$$\sigma_i^2 = (1 - \alpha - \beta)V_L + \alpha u_{i-1}^2 + \beta\sigma_{i-1}^2 \tag{5}$$

V_L is a long term variance which is constant over the period of interest. u_i and σ_i^2 are observed return and estimated conditional variance on day i, respectively. The parameter α is a measure of the influence of the most recent return value and β is a smoothing constant (for detail, see Bollerslev, 1986; Engle, 2001 and Brooks, and 2008).

This model can be fitted to the observed returns by maximizing the likelihood of the n observations. Using the formula for the probability density function of this normal distribution, the likelihood (L) is thus

$$L = \prod_i^n \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{u_i^2}{\sigma_i^2}\right) \right] \tag{6}$$

where the parameters α and β in the GARCH (1, 1) model are substituted for maximizing the likelihood function by using the Newton-Raphson method with Marquardt damping factor followed this iteration process,

$$\theta_{j+1} = \theta_j - dH_j^{-1} \times w_j \tag{7}$$

At iteration j, θj and ωj are 2×1 vectors containing estimates of α and β and their first derivatives, respectively. H_j is the corresponding 2×2 matrix of the second derivatives. The Marquardt damping factor (d) is constant and in the range between 0 and 1. This factor is designed to decrease the changes at each iteration and thus prevent overshooting maximum values, which are constrained within the conditions, $0 < \alpha < 1$, $0 < \beta < 1$ and $0 < \alpha+\beta < 1$. Confidence intervals for the parameters α and β can be obtained by using the statistical theory of maximum likelihood estimators.

After fitting the GARCH (1, 1) model, the squared root of the estimated conditional variance (σ_i^2) becomes the daily volatility (σ_i) which is used for estimating the non-stationary variance.

The non-stationary variance is estimated in term of the daily deviations (σ_i) which are assumed as a NCS function. This function is express as following equation,

$$\sigma_i = a + bi + \sum_{k=1}^{p} c_k\left(i - t_k \right)_+^3 \tag{8}$$

where i denotes time period (day). σ_i is estimated deviation on day i. The p knots are placed at $t_1<t_2<\ldots<t_p$. Function $(i - t_k)_+$ is a plus function that equal to $i - t_k$ for $i > t_k$ and 0 for otherwise. Since cubic spline function is linear in the distant past and future, the coefficients of quadratic and cubic are 0 for $i < t_1$ and $i > t_p$. To satisfy these constraints, the cubic spline functions becomes

$$\sigma_i = a + bi + \sum_{k=1}^{p-2} c_k\left[\left(i - t_k \right)_+^3 - \frac{t_p - t_k}{t_p - t_{p-1}}\left(i - t_{p-1} \right)_+^3 + \frac{t_{p-1} - t_k}{t_p - t_{p-1}}\left(i - t_p \right)_+^3\right] \tag{9}$$

The parameters a, b and c_k (k = 1, 2… p-2) are obtained by fitting the NCS function with least square estimation (Venables *et al.*, 2002).

3.2. NCS fits to absolute returns

This approach estimates the non-stationary variance by fitting the NCS function directly to the absolute value of the returns series. The returns are assumed independent and normally distributed with zero mean and inconstant standard deviation σ_i (i = 1, 2, 3… n).

Since the absolute returns have a long memory property that keeps effect of fluctuation continuing persistent over a long time period, so they can be used as the proxy of daily standard deviation (for details, see Ding *et al.*, 1993).

Note that the parameters of natural cubic spline function can be estimated by using the ordinary least squares method, but it may not give preferable results, because the returns have inconstant variance. So the maximum likelihood method is alternatively applied (Greene, 2002). The log likelihood function is followed this equation,

$$L = \sum_{i-1}^{n}\left[- \log(\sigma_i) - \frac{u_i^2}{2\sigma_i^2}\right] \tag{10}$$

The daily deviation σ_i is estimated as equation (9) on day i and u_i is return on day i. The parameters of NCS function is obtained by maximizing the log likelihood function with the Newton-Raphson method with Marquardt damping factor. The iteration process is followed the equation (7) where θ_j is p × 1 matrix that contains the estimate of parameters a, b and c_k (k = 1, 2… p-2). The vector of first derivative and second derivative of likelihood functions according to each parameter expressed as simple algebraic are contained in p × 1 matrix (ω_j) and p × p matrix (H_j), respectively. The standard errors of parameters a, b and c_k can be obtained from the square root of - *diagonal* (H^{-1}).

4. RESULT AND DISCUSSION

Following two estimating approaches described above, the non-stationary variances in term of daily deviation of four exchange rates are plotted in Figure 4. These deviations are estimated by fitting a NCS function with 45 equi-space knots. The space between each pair of knots equals to 130 days which are trading days in a half year.

Using maximum likelihood method to fit the NCS to the absolute value of the returns of these four exchange rates is remarkably better to trace the variation of returns series than fitting the NCS to the daily volatility. Moreover, as shown at the legend of each graph, fitting NCS function to the absolute returns provides the higher likelihood value than the other approach, especially in case of Indonesia Rupiah (IDR).

The estimate daily deviations of daily volatility of Indonesia Rupiah (the top left panel of Figure 4) are much higher than the absolute value of the actual daily returns during low variation period and lower during high variation period. This evidence shows that the GARCH (1, 1) model is inappropriate to estimate the daily volatility of the Indonesia Rupiah.

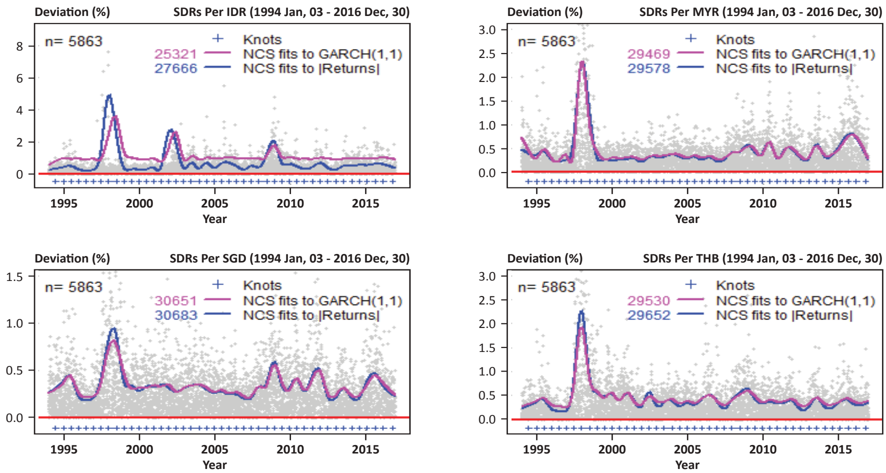


Figure 4. The Estimated Deviation of Indonesia Rupiah (IDR), Malaysian Ringgit (MYR), Singapore Dollar (SGD) and Thai Baht (THB) During 1994-2016

The same as the real data, the NCS function are fitted to two groups of generated returns by using a half year equi-space knots. The daily deviations of these two groups which are the generated returns with white noise residual and fat-tailed residual are shown in Figures 5 and 6, respectively. These estimated results show that the both estimating approached are efficient to trace the known signal of the non-stationary variance. These estimated results show the same typical trends as the specified signals.

Again, both graphic and the likelihood values shown at the legend of each graph indicate that using maximum likelihood method to fit the NCS function to the absolute returns provides the daily deviations which are fitted to the variation of daily returns better than fitting the NCS function to the GARCH (1, 1) estimates, especially in case of the generated returns with fat-tailed residuals.

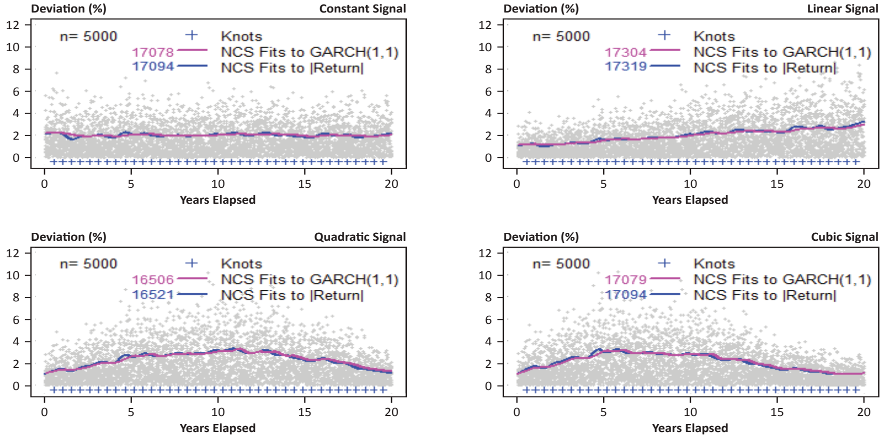


Figure 5. The Estimated Deviation of Four Dataset of the Generated Returns with White Noise Residuals and Each Known Signal

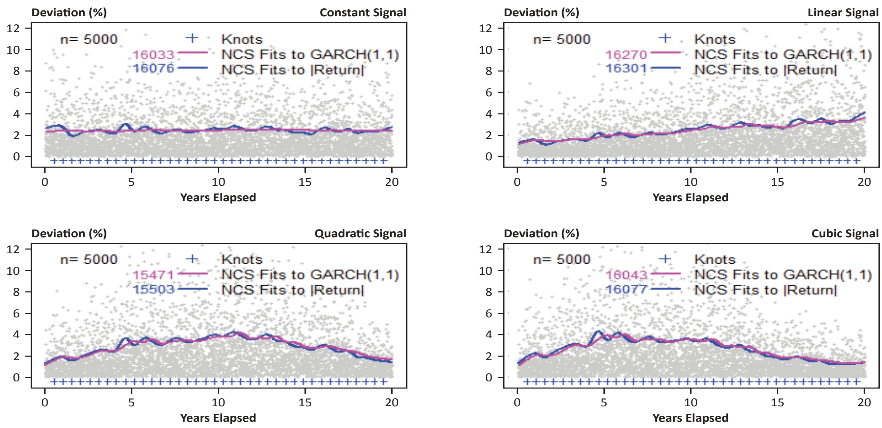


Figure 6. The Four Types of Known Volatility Signal (s_t)

5. CONCLUSIONS

This paper investigates two approaches for estimating the non-stationary variance of the financial time series data. The graphical results show that these two approaches are efficiency to trace the structural variation of daily returns, but the likelihood values show that using maximum likelihood estimation for fitting the NCS function to the absolute returns provides more accurate estimated deviations than fitting the NCS function to the estimates of GARCH (1, 1). Furthermore, this paper also shows that this maximum likelihood estimation of non-stationary variance can be utilized as the baseline for comparing to the other financial volatility models.

6. ACKNOWLEDGEMENTS

Most of all, we would like to thank emeritus professor Don McNeil for his continuous advice and encouragement.

REFERENCES

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307-327.

Brooks, C. (2008). *Introductory econometrics for finance.* Cambridge: Cambridge University Press.

Clark, P. B., Tamirisa, N., Wei, S.-J., Sadikov, A., & Zeng, L. (2004). *A New look at exchange rate volatility and trade flows.* Washington, D.C.

Cont, R. (2001). Empirical properties of assets returns: stylized facts and statistical issues. *Quantitative Finance,* **1**, 223-236.

Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance,* **1**, 83-106.

Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives,* **15** (4), 157-168.

Engle, R. (2003). Risk and Volatility: Econometric Models and Financial Practice. *Nobel Lecture, December 8, 2003.*

Eugenio, D. B. (2016). Volatile volatility: Conceptual and measurement issues related to price trends and volatility. In *Food price volatility and its implications for food security and policy, eds.* (pp. 35-57).

Greene, W. (2002). *Econometric Analysis.* New Jersey: Prentice Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

McKenzie, M. D. (1999). The impact of exchange rate volatility on international trade flows. *Journal of the economic survey,* **13** (1), 71-106.

Turner, P., & Van 't dack, J. (1993). Measuring international price and cost competitiveness. *BIS Economic Papers,* **39**.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer-Verlag.

Financial Stability Modelling for Establishing Financial Integration in ASEAN

Ferdinand David Aritonang
Institute of Statistic (STIS), Jakarta, Indonesia – 13.7620@stis.ac.id

Abstract

Financial integration is one of the important aspects of economic integration. In the beginning of January 2016, ASEAN Economic Community (AEC) has just commenced where must be supported by regional economic integration with which one of them is protecting financial stability. ASEAN economies are still growing over the past year to 4.4% with a Gross Domestic Product (GDP) of 2.4 trillion US dollars which become one of the world's major economic blocs. Therefore, this research aimed to analyse factors that affect the stability of the financial system in five ASEAN countries in the period of 2008-2014. The method used is panel regression with Fixed Effect Model (FEM) modified with General Least Square (GLS). The result shows that return on assets, Gross Domestic Product (GDP) per capita, stock market total value traded to GDP and bank deposit to GDP have significant upon financial stability. Thus, the increase in return of asset, Gross Domestic Product (GDP) per capita, stock market total value traded to GDP and bank deposit to GDP can become the solution to maintain financial stability (Bank NPL to gross loan) on the determined and future year.

Keywords: Fixed Effect Model; Economic Integration; Financial System.
JEL Classification: C5, E6, G1

1. INTRODUCTION

Recent years, economic integration and regional financial have been becoming prominent priority to strengthen economic in every regional integration around the world. It is supported by similarity of geographic, historical, and economic relation. The purpose of economic integration in a regional integration is to elevate economic development and social prosperity in the region. However, economic integration is different with financial integration which focused in the cooperation to protect financial stability and exchange rate including financial crisis prevention. Moreover, financial integration's act is removing barrier in financial traffic among countries in the regional bloc and elaborate regional financial infrastructure to support liquidity and maintain financial stability in the region.

Financial integration process requires financial stability in the regional integration. Financial stability comprises soundness and reliability of national banking as intermediary institution to enhance economic activity. Bank Indonesia (2016) state financial stability is supported by the resilience of capital, Bank's liquidity escalation and safe payment system performance as one of the financial system infrastructure to promote financial stability and expedite economic activity.

Monetary crisis happened in Southeast Asia during 1997 as the impact of financial crisis in Thailand that affected exchange rate, stock price and another asset prices. It gave a domino effect for countries in ASEAN which also made high inflation, thus investors perceived reluctant to invest their money in ASEAN. That problem is occurred by financial instability in every country so that brought about effect in their economy entirely.

In the beginning of 2016, the ASEAN Economic Community (AEC) was established whose purpose is to create economic regional integration to ease trade access and financial. Free trade in ASEAN is expected give positive impact for financial integration with stable intraregional exchange rate and also trade expansion among ASEAN countries. It is becoming a challenge of ASEAN members after establishing the AEC. In the financial ministries and central bank governors meeting in Vientiane, Laos an initial action in 2025 was approved to strengthen financial integration, financial inclusion and regional financial stability for supporting integration, openness, and stability in every financial sector in

ASEAN. In the middle of global economic slowdown, ASEAN economic grew up to 4.4% after AEC establishment, ASEAN is recognized as the greatest world economic bloc which considers 600 million population with 320 million labor force and a gross domestic product (GDP) of 2.4 billion US dollar.

The European Central Bank (2012) defines financial stability as a condition in which the financial system comprises of financial intermediaries, markets and market infrastructures capable of withstanding shocks, thereby reducing the likelihood of disruptions in the financial intermediation process which is severe enough to significantly impair the allocation of savings to profitable investment opportunities.

Furthermore, the European Central Bank (ECB) defines three particular conditions associated with financial stability, i.e. the financial system that should be able to efficiently and smoothly transfer resources from savers to investors, financial risks that should be assessed and priced reasonably accurately and should also be relatively well managed, and the financial system that can comfortably absorb financial and real economic surprises and shocks.

Anjom, W., &Karim, A. M. (2016) stated that financial stability is considered as a pre requisite for the sustained and rapid economic progress for any economy. Among various indicators of financial stability, bank's non- performing loan (NPL) is the most critical because it reflects the asset quality, credit risk and efficiency in the allocation of resources to productive sectors.

Therefore, study about financial stability in ASEAN is well performed as a reference in establishing economic integration in a bloc. The purpose of this paper is to find out profile of financial stability in ASEAN which use NPL as a proxy and to find out which variables affect significantly financial stability in ASEAN. The variables comprises of Bank deposit to GDP, Bank return on asset (ROA), GDP per capita and stock market total value traded to GDP and use panel regression method to explain financial stability model and the variables that have a significant effect to financial stability.

2. THEORY AND LITERATURE REVIEW

Panel data is combination between cross section and time series. Therefore, the observation is large that becomes one of the advantage of using panel data. Based on Gujarati, D. N., &Porter, D. C. (2009), a several benefit of using panel data are coping heterogeneity explicitly, giving more information, more degree of freedom and more efficient, an appropriate model to study about dynamic change, the best data to detect and measure the invisible impact rather than cross section and time series, simplify complexity model and minimize bias as a consequence of individual aggregate.

Baltagi, B. H. (2005) arranged the general form of panel regression model

$$y_{it} = \alpha + X'_{it}\beta + u_{it}$$

$$u_{it} = \mu_i + \upsilon_{it}$$

where i denotes cross section and t denotes time-periods with i = 1,2,…,N and t = 1,2,…,T. α is scalar, β is K x 1 and X_{it} is the it-th observation on K explanatory variables. u_{it} is error components where μ_i's are cross section specific component and υ_{it} are reminder effects.

Estimating model's parameter of panel data consist of common effect model, fixed effect model and random effect model. Meanwhile, selection of panel regression model is conducted to choose a proper model for estimating parameter of panel regression. Several test for model selection are Chow-test, Hausman-test and Lagrange-Multiplier test. Chow-test is used to choose between common effect model or fixed effect model for estimating parameter. Moreover, Hausman test is used to choose between fixed effect model or random effect model. And Lagrange-Multiplier is used to choose between common effect model or random effect model. The last, classical assumption test is conducted to verify whether the model generate the best linear unbiased estimator (BLUE) or not.

Bank Indonesia. (2007) defines financial stability as a stable financial system which capable to allocate fund sources and absorb shock to prevent disruption in real sector and financial system. Awareness of financial stability is needed as surveillance and regulation in macro-prudential policy. Generally, instability in financial system can generate disadvantage condition such as monetary policy transmission become inefficiency, stagnating intermediary function that make slowdown economic growth, public incredulity in financial system and high cost saving if occur crisis in systemic.

2.1. Non-Performing Loan (NPL)

Non-performing loan (NPL) is one of the indicators of asset quality to measure the quality of bank productive asset to bad credit, thus the lower the ratio shows the better the asset quality. Among the various indicators of financial stability, NPL can reflect asset quality, credit risk and efficiency to allocate resources to productive sector. Declining NPL continuously in banking balance gives a benefit for economic from micro-prudential and macro-prudential perspective.

Escalation of non-performing loan proportion in banking sector indicate poor financial health and incapability of government as regulator in banking sector. Woo. , &David. (2000) explained that NPLs are viewed as typical byproduct of financial crisis: they are not main product of the lending function but rather an accidental occurrence of the lending process, one that has enormous potential to deepen the severity and duration of financial crisis and to complicate macroeconomic management.

2.2. Bank Return on Asset (ROA)

Return on Asset is one of the profitability indicators that used to measure bank capability for producing profit during certain period, the higher ratio means the best profitability is produced in its management. ROA also indicate bank efficiency in producing income that use asset and equity. The higher diversification of investment, the bigger benefit that gotten by equity and asset using. Efficiency in financial management effect lending facility policy that afford to increase benefit and decrease non-performing loan. The highest benefit make the lowest incentives that used to risk activity. Thus, ROA show a negative relation with non-performing loan.

Berger, A. N., &DeYoung, R. (1997) explained that bad management in financial institution will generate banking inefficiency that will affect lending process. It indicates non-performing loan enhancement, ROA is also the indicator to measure management effectiveness to execute bank operation thus it has the opposition direction between ROA and NPL.

2.3. Gross Domestic Product (GDP) per Capita

Anjom, W., &Karim, A. M. (2016) defined gross domestic product (GDP) as the measurement of the total market value of the goods or services produced by the economy of a particular country as well as total income earned by the people living at that country. Meanwhile, GDP per capita is defined as average income of population in one country where total GDP is divided to total population.

The highest GDP indicate stable economic in one country and also show increasing people's income. People with high trend of income who lend money from bank indicate that they are capable to pay back the lending cost.

Empirical study by Khemraj, T., &Pasha, S. (2009) show negative relation between GDP and non-performing loan. It can be explained that GDP growth usually increase income which also increase the capacity of lender to pay the lending cost that make non-performing loan declining.

2.4. Stock Market Total Value Traded to GDP

Stock Market is defined as a set of entity including seller and buyer that are doing transaction activity of stock in a country. Twu, M. (2005) explained that stock market as part of financial market has several function in economic activity such as raising capital for businesses, mobilizing savings for investment, facilitating company growth, redistribution of wealth, corporate governance, creating investment opportunities for small investors, raising capital government for development projects and barometer of the economy.

In financial system, fund flow is gathered in financial market through traditional banking activity like lending and deposit operation. However, in modern financial system, people switch their fund to stock investment that can give a huge advantage. Stock market total traded to GDP more than 1 indicate that stock market are liquid, stable and more integrated with international. Jakubik, P., &Reininger, T. (2014) explained the stock market is leading indicator for financial and economic development in a country that has direct negative effect to non-performing loan.

2.5. Bank Deposit to GDP

World Bank. (2014) defined bank deposit to GDP as total asset in money deposit to GDP ratio in a country. Asset includes nonfinancial real domestic asset, nonfinancial public firm and private firm. Money deposit comprises of commercial bank and another financial institutions that have transaction activities of deposit demand.

Credit lending is conducted by banking industry in saving and deposit. The lowest Bank deposit to GDP ratio indicate banking industry hard to distribute credit that impact declining profitability. One of the banking industry treatment to increase liquidity is to enhance interest rate which triggers investor to invest their money, but this gives domino effect which makes high credit interest rate. Thus, lender will be difficult to payback the lending cost that increases non-performing loan.

Moreover, the high NPL is not just indicated by banking industry mechanism in applying high deposit interest rate but also by poor quality of lender. Yigitbas, S. B., &Cambazoglu, B. (2014) stated that credit or lending cost which is distributed from money deposit becomes a function of economic activity which parallel with monetary policy. Ozurumba, B. A., &Onuorah, C. (2013) explained that there is a positive relation between lending cost with economic growth. Economic growth will increase people's income that increases capacity to pay lending cost, so this contributes to decrease non-performing loan.

2.6. Literature Review

Several studies about financial stability in previous time with different researchers of different countries. Skarica, Bruna. (2013) conducted a study about determinants of non-performing loans in central and eastern European countries. The aim of his study is to find out the variables that affected non-performing loans. The method uses panel regression with fixed effect model that concludes GDP growth is a main stimulus for the NPL declining for last 5 years, so the government can apply monetary expansion to increase GDP and aggregate demand. Another study by Farhan et al. (2012) was about economic determinants of non-performing loans in Pakistan that have a purpose to identify economic determinant that influenced NPL based on Pakistani perception. This study uses multiple regression that concludes interest rate, crisis energy, unemployment, inflation and exchange rate have a positive relation with NPL, meanwhile GDP have negative relation. An the last, conducted by Anjom, W., &Karim, A. M. (2016) about relationship between non-performing loans and macroeconomic factors with bank specific factors in South Asian Association for Regional Cooperation to know macroeconomic factors and banking factors which resulting in non-performing loans. The method uses multiple regression that concludes high NPL that will decrease capital adequacy ratio (CAR) and debt ratio to GDP, lending growth, return on equity, return on asset, total loan to total asset ratio, total loan to total deposit ratio and operating expense to operating income ratio have a significant relation.

3. METHODOLOGICAL

3.1. The Type and Sources of Data

This paper is using panel data that is defined as combination of cross section and time series. The cross section consist of 5 countries in ASEAN such as Indonesia, Malaysia, Singapore, Thailand and Philippine while time series consist of 7 annual period during 2008-2014. The data is gathered form World Bank include World Development Indicator (WDI) and Global Financial Development Database (GFDD). In addition, data processing in this paper is using Eviews 8 and Microsoft 2010.

Table 1. Sources of Data Summary

No	Data	Unit	Sources
1	Non-performing Loan	%	WDI
2	Bank Deposit to GDP	%	GFDD
3	GDP per Capita	USD	WDI
4	Bank Return on Asset	%	GFDD
5	Stock Market Total Traded Value To GDP	%	GFDD

3.2. Method of Analysis

Analysis method used in this paper consists of descriptive analysis and inference analysis. Descriptive analysis is used to know the profile of financial stability in ASEAN 5 while inference analysis is used to generate a model to know a variables that effect dependent variable. Inference analysis is using static panel regression with General Least Square (GLS) estimation and fixed effects model approach. GLS is a robust model for autocorrelation and heteroscedastic.

The initial step to estimate panel data is model formulation. After that, the next step is selecting the best model which use several test such as Chow test, Lagrange Multiplier test and Hausman test. The last, verify the criteria test that consist of statistic test, econometric test and economy test.

Dependent variable in this paper is non-performing loan as financial stability proxy while the independent variables are bank deposit to GDP, LN GDP per capita, bank return on asset, and stock market total traded value to GDP. Thus, the panel regression is:

$$NPL_{it} = \alpha + ROA_{it}\beta_1 + Deposit_{it}\beta_2 + LgGDPP_{it}\beta_1 + Stock_{it}\beta_1 + u_{it}$$

Where:

NPL_{it} : Financial stability proxy for i counties and t period
ROA_{it} : Ratio between income after tax to total asset for i counties and t period
$Deposit_{it}$: Ratio between bank deposit to GDP for i counties and t period
$LgGDPP_{it}$: LN GDP per capita for i counties and t period
$Stock_{it}$: Ratio between total value stock traded in market to GDP for i counties and t period

4. MODEL SPECIFICATION AND RESULTS

4.1. Descriptive Analysis

According to the above consideration, the following calculation of dependent and independent variable are below:

Table 2. The Average Indicators of Five ASEAN Countries During 2008-2014

Country	NPL	ROA	LgGDPP	Stock	Deposit
Indonesia	2.385714286	2.007016	3281.261975	12.39251	31.75531
Malaysia	2.842857143	1.348525	9463.827795	42.45114	118.8294
Singapore	1.228571429	1.133342	47100.34094	113.8815	117.4374
Thailand	3.528571429	1.164357	5222.814811	57.36856	101.2917
Philippine	2.957142857	1.237533	2234.090023	11.71719	52.41392

Source: World Bank (edited)

Based on Table 2, Singapore has the most stable financial system in ASEAN 5 indicated by the lowest non-performing loan (NPL). It is caused by the lower credit interest rate in Singapore approximately 5-5%. Moreover, it is also caused by a lower net interest margin (NIM) of averagely 1.75%, thus the creditors have a capability to pay the lending cost which also supported by the high income per capita of Singapore's people. Meanwhile, Thailand has the less stable financial system in the ASEAN 5 indicated by the highest non-performing loan (NPL) which is caused by the poor quality of creditor in Thailand. Therefore, in 2015 the government of Thailand had set aside high level of loan provision and maintained investment grade credit rating.

From five countries in ASEAN, Singapore has the highest level of GDP per capita that is supported by the high GDP of averagely 334 billion dollar Singapore. In other side, the lowest GDP per capita is Philippine, although services sector contributes value added about 55% of GDP, but agricultural sector only shares 12.34% of GDP that is opposite with agricultural contribution of about 41% of land area. Moreover, inefficiency of forest conversion to unproductive land leads to a failure land rehabilitation that exert bad effects for climate, precipitation, and sedimentation which cause environment degradation. The degradation has a correlation with poverty because the province with high degradation of environment doesn't have enough land for agricultural activities.

The highest Bank Return on Asset occurred in Indonesia, it was caused by the percentage of net interest margin in Indonesia that reached the highest level of about 5.99% which influenced net income of banking industry and also effect return on asset (ROA). In contrast, Singapore become the lowest one producing return on asset in banking industry as impact from low net interest margin of about 1.75%.

The next variable is stock market total value traded to GDP, Singapore has the highest level of it of about 113% from GDP this indicates that the flow of fund in Singapore is not just in traditional banking system but also in stock investment such as Strait Times Index (STI). Meanwhile, the lowest one is Philippine with only 11.72%, this means Philippine's people still depend on traditional banking system.

And the last variable is bank deposit to GDP, Malaysia has the highest level of bank deposit of around 118%, it also indicates that banking industry in Malaysia has a huge liquidity fund to conduct for lending policy, but the highest level of non-performing loan in ASEAN 5 indicates that lending activities conducted by banking industry didn't face an appropriate creditor. In contrast, Indonesia as large population country in ASEAN has the lowest level of bank deposit of about 31.76%. It is caused by the high level of un-bankable people in Indonesia that didn't save their money in banking industry.

4.2. Inference Analysis

The initial step to determine an appropriate model is selecting model based on several test. The first is conducting Chow test, the result shows that the probability value < 0.05 which means reject H_0 and conclude fixed effect model as estimation model. Moreover, the Hausman test is applied to ensure the best model, the result shows that the probability value < 0.05 which means reject H_0 and conclude that fixed effect model as the best model. The last, to cope heteroscedasticity and autocorrelation is using General Least Square estimation. R-Squared of this model is 0.9674 which means 96.74% the variance of dependent variables can be explained by independent variable. Moreover, simultaneous test to determine goodness of fit also significant in alfa 5% with F-statistic 96.45. Finally, the result shows:

Table 3. Estimating Financial Stability Model

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ROA	-0.327028606	0.120098217	-2.723009668	0.01
STOCK	-0.015602406	0.002128787	-7.329245542	0.00
DEPOSIT	-0.06756524	0.010634957	-6.353127861	0.00
LNGDPP	-7.841834902	0.572997516	-13.68563508	0.00
C	78.91088623	5.308791934	14.86418892	0.00

Bank return on asset (ROA) has a negative relationship with non-performing loan (NPL) and significant in alfa 5%. Based on table 3 as estimation result, increasing 1% of ROA will decrease 0.32% of NPL, ceteris paribus. Berger. , &DeYoung. (1997) indicate that enhancement in ROA will increase financial stability that caused by increasing the profitability and banking efficiency in managing their fund that also effects the lending process.

Stock market total traded value to GDP (STOCK) has a negative relationship with non-performing loan (NPL) and significant in alfa 5%. Based on the estimation result, increasing 1% of STOCK will decrease 0.016% of NPL, ceteris paribus. It indicates that enhancement in STOCK will increase financial stability. Jakubik, P., &Reininger, T. (2014) conclude that STOCK plays a role as early proxy indicator for the value of loan collateral.

Bank deposit to GDP (DEPOSIT) has a negative relationship with non-performing loan (NPL) and significant in alfa 5%. Based on the estimation result, increasing 1% of DEPOSIT will decrease 0.068% of NPL, ceteris paribus. It indicates that enhancement in DEPOIST will increase financial stability because the high bank deposit to GDP ratio, the high lending activities that given by banking industry that also increase economic activity, thus Economic growth will increase people's income that increase capacity to pay lending cost, so contribute to decrease non-performing loan.

GDP per capita (LNGDPP) has a negative relationship with non-performing loan (NPL) and significant in alfa 5%. Based on the estimation result, increasing 1% of LNGDPP will decrease -7.84% of NPL, ceteris paribus. Allen et al. (2012) indicates that enhancement in LNGDPP will increase financial stability because increasing GDP per capita will increase formal account in bank. This increasing also increase saving basis and intermediary process.

Interpretation of individual effect indicate that in constant independent variable, Singapore has the highest intercept which reflect the highest non-performing loan (NPL) when be assumed all independent variable didn't effect dependent variable while Philippine has the lowest NPL.

5. CONCLUSION

The variables that have a significant effect to decrease non-performing loan (NPL) are bank return on asset (ROA), stock market total traded value to GDP, bank deposit to GDP and GDP per capita. Indeed, increasing all those variables will keep financial stability in a safety level. Moreover, the average of global NPL during 2008-2014 was about 3.87% based on World Bank data, thus countries in ASEAN actually have stable financial system. As a suggestion, ASEAN countries should increase aggregate demand to elevate GDP per capita, the government also should apply low interest rate to help creditor run their business. In addition, bank asset diversification and facilitating stock market traded should also become the main attention to create financial stability as the core of financial integration.

REFERENCE

Allen, F., Kunt, A. D., Klapper, L., &Peria, M. S. M. (2012). The Foundation of Financial Inclusion: Understanding Ownership and Use of Formal Accounts. *World Bank Policy Research Working Paper.*

Anjom, W., &Karim, A. M. (2016). Relationship between Non-Performing Loans and Macroeconomic Factor with Bank Specific Factors: A Case Study on Loan Portfolios – SAARC Countries Perspective. *ELK Asia Pacific Journal of Finance and Risk Management.*

Baltagi, B. H. (2005). *Econometric 4th edition.* Berlin: Springer-Verlag Berlin Heidelberg.

Bank Indonesia. (2016). Financial Stability Report. Jakarta: Bank Indonesia.

Bank of Thailand. (2015). Annual Report 2015. Bangkok: Bank of Thailand.

Berger, A. N., &DeYoung, R. (1997). Problem Loans and Cost Efficiency in Comercial Banks. *Journal of Banking and Finance, Vol. 21, pp. 849-870.*

European Central Bank. (2012). Financial Stability Review. What is Financial Stability? Frankfurt: European Central Bank.

Farhan et al. (2012). Economic Determinants of Non-Performing Loans: Perception of Pakistani Bankers. *European Journal of Business and Management.*

Gujarati, D. N., &Porter, D. C. (2009). *Basic Econometric 5th edition.* New York: McGraw-Hill.

Jakubik, P., &Reininger, T. (2014). What Are the Key Determinants of Nonperforming Loans in CESEE. *Faculty of Social science, Institute of Economic Charles University Prague Working Paper No. 2014/26.* Prague.

Khemraj, T., &Pasha, S. (2009). The Determinants of Non-Performing Loans: An Econometric Case Study of Guyana. *Presented at the Carribean Centre for Banking and Finance Bi-annual Conference on Banking and Finance, St. Augustine, Trinidad.*

Morgan, P., &Pontines, V. (2014). Financial Stability and Financial Inclusion. *Asian Development Bank Institute Working Paper No. 448.* Tokyo.

Ozurumba, B. A., &Onuorah, C. (2013). Bank credit: An aid to economic growth in Nigeria. *Information and Knowledge Management, Vol. 3, No. 3, pp. 41-50*

Skarica, Bruna. (2013). Determinants of Non-Performing Loans in Central and Eastern European Countries. *Faculty of Economics and Business University of Zagreb Working Paper No. 13-07.* Zagreb.

Twu, M. (2005). Unified Stock Market for the Promotion of Business Activities in West African Monetary Zone (WAMZ). Taiwan: National Chengchi University.

Woo. , &David. (2000). Two Approaches to Resolving Nonperforming Assets during Financial Crisis. *IMF Working Paper 00/33, March: 2-5.*

World Bank. (2014). *World Development Report Risk and Opportunity.* Washington, DC: World Bank.

Yigitbas, S. B., &Cambazoglu, B. (2014). Modeling the Relationship between Economic Activity and Bank Loans to Private Sector: A Case of Turkey. *International Journal of Trade, Economic and Finance, Vol. 5, No. 5, October 2014*

Analysis of Japanese Financial Market by Text Mining with Both Time Series and Hierarchical Structures of Topics

Shunsuke Yuasa*
Kwansei Gakuin University, Sanda, Japan– etw99266@kwansei.ac.jp
Takayuki Morimoto
Kwansei Gakuin University, Sanda, Japan – morimoto@kwansei.ac.jp

Abstract

In this study I analyze articles in newspaper with a topic model, and study a relation between transitions of topics and time series of stock prices. When changes of stock prices or articles happen for short terms, vanilla topic models cannot reflect them. Therefore topics are estimated by Dynamic Stacked Topic Model (DSTM), which analyzes the hierarchical structure and the time evolution of topics.
Using DSTM, changes of topics on time series can be analyzed in a broad grouping. If analyzing them with each topic, its occurrences may vary greatly. However it is expected that occurrences of super-topics which are collections of topics are not that. I use Automatic Differentiation Variational Inference(ADVI) for decreasing calculating time with respect to estimating parameters of DSTM.

Keywords: topic model, hierarchical structures, DSTM, Japanese financial market.

I. INTRODUCTION

Recently there have been many reports about analyzing events in the real world not only with numerical data, but also with nonstructural data for which it is difficult to define data structures. One of these data is text data. For examples optimizing operational, getting information for improving products, or analyzing change popular culture, text mining is studied activity in order to obtaining new knowledge from a lot of text data. In text mining we must do natural language processing, such as Morphological Analysis before extract information by approaching with stochastic. It is more difficult for Japanese texts to do that than English these, because Japanese language is not separate each words. However we can do that to common texts, such as newspapers with accuracy. An example of text mining studies for analyzing economic is Vlastakis and Markellos (2012). Its study is for analyzing stock volatility with headline of news and google search volume index (SVI). There are some problems, for instance results is change by selected words or notation variability because it uses the frequency of specific indexes. A study by Ymauchi use not indexes but topics in documents. Its study cannot respond to abrupt economic changes or a rapid increase of news. In this study I analyze with abstract topics which is super topic (in the rest of this document referred to as the section).

2. DYNAMIC STACKED TOPIC MODEL

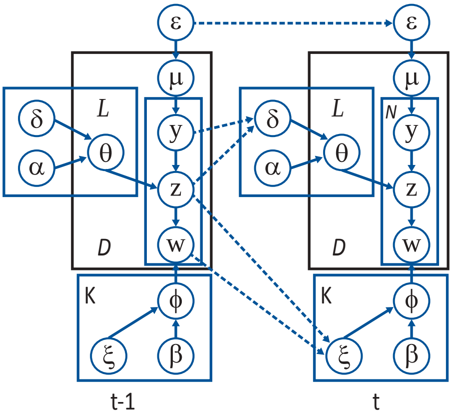
In this section I explain about Dynamic Stacked Topic Model (DSTM). The model is one of models expended from Latent Dirichlet Allocation (LDA). Especially it expended from both Pachinko Allocation Model (PAM) and Multiscale Dynamic Topic Model (MDTM). PAM is a model concerned relation between topics by introduced hierarchal structures to topics. For example, a topic of Japan and a topic of Indonesia are included in a super-topic of Asia. MDTM is a model concerned multiscale time series which depend not only on just before time. Using DSTM you can analyze documents with time series and hierarchal structures, and know transition of extensive topics. I show parameters of DSTM (a) and a graphical model of it (b).

Parameters of DSTM

Parameters	Explain
Dt	The number of documents at the time of t
Nt,d	The number of words at the time of t, document of d
W	The number of vocabularies
w,d,n	The n-th word at the time of t, document of d,
K	The number of topics
zt,d,n	The topic of w,d,n
L	The number of section
yt,d,n	The section of zt,d,n
θt,y	The topic section y at the time of t
φt,z	The word distribution topic z at the time of t
μt,d	The section distribution document d at the time of t
εt	The prior distribution of the section distribution at the time of t
ξt,z	The word distribution topic z at the time of t
δt,y	The topic distribution section y at the time of t

A graphical model of DSTM

α and β is hyper parameters of dirichlet distribution



Time evolution of sections flows from Gamma distribution.

$$\text{Gamma}\,(\,\varepsilon_t \mid \gamma\varepsilon_{t-1}, \gamma\,) = \frac{\gamma^{\gamma\varepsilon_{t-1}}}{\Gamma(\gamma\,\varepsilon_{t-1})}\,\varepsilon^{\gamma\varepsilon_{t-1}}\exp(-\,\gamma\,\varepsilon_t)$$

Time evolution of word distributions and topic distributions follow from each past distribution.

$$\xi_{t,z,w} = \frac{N_{t,z,w}+1}{\sum_w N_{t,y,z}+V}$$

$$\delta_{t,y,z} = \frac{N_{t,y,z}+1}{\sum_w N_{t,y,z}+K}$$

The generative process of DSTM is thus as follow:

1. for section $l \in \{1 \ldots L\}$:
   - generating prior distributions of section
     $\boldsymbol{\epsilon}_{t,l} \sim Gamma(\boldsymbol{\gamma}\boldsymbol{\epsilon}_{t-1}, \boldsymbol{\gamma})$
2. for topics $k \in \{1 \ldots K\}$:
   - generating distributions of word
     $\boldsymbol{\phi}_{t,k} \sim \text{Dirichlet}(\boldsymbol{\beta}_{t,z}\xi_{t-1,z})$
3. for documents $d \in \{1 \ldots D_t\}$:
   - generating distributions of section
     $\mu_{t,d} \sim \text{Dirichlet}(\boldsymbol{\epsilon}_t)$
   - generating distributions of topic
     $\boldsymbol{\theta}_{t,y} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{t,y}\boldsymbol{\delta}_{t-1,y})$
   - for words $i \in \{1 \ldots N_{t,d}\}$:
     (a) generating sections
     $l_{t,i} \sim \text{Multinomial}(\mu_{t,d})$
     (b) generating topics
     $z_{t,i} \sim \text{Multinomial}(\theta_t l_{t,i})$
     (c) generating words
     $w_{t,i} \sim \text{Multinomial}(\phi_{z_{t,i}})$

3. ANALYZING TEXT DATA AND FINANCIAL MARKET

Using Topic Model for Japanese documents, documents must be Bag of Words (BOW). In this study I use MeCab, which is Morphological Analysis machine for that. I use the NIKKEI (Nihon Keizai Shimbun; it means Japanese Economical Newspaper) as text data and use only noun for appearing characteristics of documents. Section distributions is estimated from that data by DSTM. The Score calculating a transition of sections follow from the next formula. $\mu_{t,j,i}$ is means the ratio of section i at the time of t, and document of j.

$$s_{t,i} = \sum_{j=1}^{D_t} \mu_{t,j,i}$$

For example, added the score to AR(1), we can get the next formula with regression coefficient ζ, 0≤ζ≤1.

$$\log(RV_t) = \boldsymbol{\alpha} s_{t,i} + \zeta \log(RV_{t-1}) + \boldsymbol{\omega}_t$$
$$\boldsymbol{\omega}_t \sim N(0, \sigma^2)$$

Like that formula, added the score to realized volatility as exogenous variables, financial time series can be analyzed.

REFERENCES

Shimizu Takuya, and Okadome Takeshi Dyamic Stacked Topic Model: A Dynamic Topic Model for Topic-Correlated Documents. The 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2015

Nikolaos Vlastakis and Raphael N. Markellos. Information demand and stock market volatility. Journal of Banking & Finance, 2012

Kaito Yamauchi, Takayuki Morimoto. Jikeiretsu topic model wo mochita kabushiki shijo no bunseki. SIG-FIN-010-01, 2013

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res. 2003.

Tomoharu Iwata, Takesi Yamada, Yasushi Sakurai, and Naonori Ueda. Online multiscale dynamic topic models. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010

Li, W. and A. McCallum: Pachinko allocation: DAG-structured mixture models of topic correlations. Proceedings of the 23rd International Conference on Machine Learning, 2006

Li, W., D. Blei and A. McCallum: Nonparametric Bayes pachinko allocation. Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, 2007

This page intentionally left blank

CPS26:
Food Security Statistics

Improving Storage and Distribution of Agricultural Produce in Bangladesh: Can Cluster Analysis Help?

Tamanna Howlader and Fatema Fazrin
1,2 Institute of Statistical Research and Training, University of Dhaka, Dhaka 1000, Bangladesh

Abstract

The growth of an agrarian economy depends not only on the efficient allocation of resources but also on the preservation and efficient distribution of its agricultural produce. Unfortunately, in developing countries such as Bangladesh, inadequate storage facilities and inefficient distribution systems lead to substantial wastage and loss. One of the most important reasons for this is poor planning. The purpose of this study is to improve planning by segmenting the regions of Bangladesh into groups according to their agricultural production/yield. By identifying high yield and low yield districts, planners are able to better decide on the number and types of storage facilities required, where such facilities need to be provided and in which regions road networks and other modes of transportation need to be strengthened. The study uses secondary data collected from the 2013-14 Yearbook of Agricultural Statistics of Bangladesh published by the Bangladesh Bureau of Statistics. Hierarchical cluster analysis is performed to classify the regions according to the yield of three major crops, namely, rice, fruits and vegetables. Results indicate the existence of five clusters for rice production and three clusters for fruit and vegetable production. Interestingly, districts or regions within the same cluster are geo- graphically well dispersed and in some cases single member clusters are obtained. The paper concludes with recommendations to inform policy.

Keywords : Clustering.

1. INTRODUCTION

According to the Food and Agriculture Organization, upto one-third of all food produced for human consumption is lost or wasted globally. Global quantitative food losses and waste per year are roughly 30% for cereals and 40-50% for root crops, fruits and vegetables. It is estimated that even if just one-fourth of the food currently lost or wasted globally could be saved, it would be enough to feed 870 million hungry people in the world (Food and Agricultural Organization, 2017). In agrarian economies, such as Bangladesh, wastage of agricultural produce is of particular concern due to its effects on the economy. In Bangladesh, agriculture is the single largest producing sector of the economy comprising about 17 percent of the country's GDP and employing around 45 percent of the total labor force (Bangladesh Bureau of Statistics, 2016). Thus, the performance of this sector has an overwhelming impact on major macroeconomic objectives like employment generation, poverty alleviation, human resources development and food security.

Although there are several constraints to the agricultural sector in Bangladesh, such as lack of proper land use planning, insufficient capital for agricultural activities, decrease in the availability of agricultural land, frequent natural calamities and inadequate use of appropriate technology, a more important constraint to the industry's growth is wastage of agricultural produce. Agricultural food loss and wastage is a consequence of having an average capacity of 2500 tonnes. The capacity of refrigerated warehouses in the country is about 8 million cubic meters (Board of Investment Bangladesh, 2014). However, almost all cold storages in Bangladesh are primarily used for storing potatoes and are inadequate for storing high value items like fruits and vegetables which are rapidly perishable. Moreover, lack of adequate transportation facilities, poor planning and inefficient management impede quick distribution of agricultural produce to market places all over the country.

The overall production in Bangladesh conceals considerable regional difference because of differences in farming practices, availability of irrigation facilities, attitude of the farmer, etc. in different parts of the country. The differences in agricultural productivity among the regions can also be attributed to natural phenomena, such as, rainfall, temperature, humidity and some other agri-ecological features which are relatively less avorable in the lagging regions. As a result, there exists clusters of highly productive and low productive regions or districts. The objective of this study is to identify these clusters by performing a hierarchical cluster analysis of the districts or regions of Bangladesh according to the yield of three major crops, namely, rice, fruits and vegetables. The analysis is based on secondary data collected from the 2014 Yearbook of Agricultural Statistics published by the Bangladesh Bureau of Statistics. Identification of these clusters would be useful in planning and strengthening the already existing National Agriculture Policy. For instance, information on high yielding clusters would enable planners to decide on the number and location of cold storage facilities required and how these facilities should be distributed to minimize post-harvest loss due to wastage. It would also help planners to decide in which regions road networks and transportation services need to be improved and expanded to ensure rapid distribution of perishable products to markets. Knowledge regarding high yield clusters could also be useful in deciding where to establish agro-processing and agro-based industries to minimize wastage of agricultural produce. Furthermore, the government could chalk out programmes to ensure profitable and sustainable agricultural system for members of high yield clusters. Identification of low yield clusters is equally important to policy makers for developing new measures to boost agricultural production in these areas.

The paper is organized as follows. Section 2 gives a brief overview of the agricultural industry in Bangladesh. Data and variables are described in Section 3 while details regarding the analysis are provided in Section 4. The results of the analysis are presented in Section 5 and conclusions and policy implications are discussed in Section 6.

2. THE AGRICULTURAL INDUSTRY IN BANGLADESH

As the world's largest delta, the vast plains of Bangladesh are washed by the mighty rivers - the Padma, Meghna, Jamuna and Karnaphuli, and are drenched by tropical monsoon rains that lead to flooding of low and outlying areas every year. Hence the majority of the country's thirty agro ecological zones have between medium to high levels of fertility for which a variety of crops grow abundantly in this country.

Rice is the single most important food grain and is grown in about 75% of the cultivated land in Bangladesh. The cultivation of rice in Bangladesh varies according to seasonal changes in the water supply. The largest harvest is aman, occurring in November and December and accounting for more than half of annual production (Bangladesh Bureau of Statistics, 2016).

The second harvest is aus, involving traditional strains but more often including high-yielding, dwarf varieties. Rice for the aus harvest is sown in March or April, so that it benefits from April and May rains and then matures during the summer rains and is harvested during the summer. With the increasing use of irrigation, another variety that is grown during the dry season from October to March is boro rice. Total rice production in Bangladesh was about 343 lakh metric tons during 2013-14 (Bangladesh Bureau of Statistics, 2016).

Bangladesh abounds with a large variety of tropical and subtropical fruits. About 30 species of fruits are common in Bangladesh. Among them mangoes,bananas, jack fruits, pineapple, litchi, guava and green coconut are the most common. Total fruit production was 4587 thousand metric tons during 2013-14. In 2010-11 there were exports worth Tk. 610 million of fresh or chilled fruits. Over the years export earnings from fruits have declined to Tk. 19 million in 2013-14 (Bangladesh Bureau of Statistics, 2016).

Vegetables are grown both in the summer and winter seasons. Main vegetables grown are brinjal, tomato, cabbage, green banana and papaya. During 2013-2014, total vegetable production was 3357 thousand metric tons. Although not an important export commodity, Tk. 3170 million worth of fresh or chilled vegetables were exported in 2010-11. However, export earnings from vegetables have declined to Tk. 108 million in 2013-14 (Bangladesh Bureau of Statistics, 2016).

Bangladesh is yet to attain self sufficiency in food production. Hence, the overall objective of the national agricultural policy is to make the nation self sufficient in production of all crops. This objective is difficult to realize unless Bangladesh minimizes wastage of its agricultural produce.

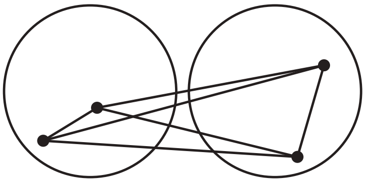


Figure 1. Illustration of Average Linkage within Groups Method of Clustering.

3. DATA AND VARIABLES

The Bangladesh Bureau of Statistics (BBS) is Bangladesh's national statistical agency providing the Government with reliable statistics to guide decision making and development. The organization compiles agricultural data which is published annually in the Yearbook of Agricultural Statistics. The publication contains data on crops area and production, crop damage, weather indicators, agricultural inputs, livestock, fisheries as well as export-import data. This study analyzes secondary data on crop production given in the 2013-14 Yearbook of Agricultural Statistics (Bangladesh Bureau of Statistics, 2016). Data on crop production are collected by the BBS via clustered sample survey. Selected sample clusters are visited four times in a year and crop-cutting experiments are conducted in due time to estimate the major crop productions in Bangladesh. Both household based interview and direct observation methods are used to obtain crop estimates. Further details regarding data collection and crop estimation system are available in the 2013-14 Yearbook of Agricultural Statistics (Bangladesh Bureau of Statistics, 2016).

Bangladesh is divided into 64 districts. The Yearbook provides district-wise data on production for major cereals and jute. However, for fruits, vegetables and other types of crops, the country has been divided into 23 regions and region-wise data on production are available. This study focuses on the production of three types of perishable crops, namely, rice, fruits and vegetables.

4. METHODS AND ANALYSIS

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups or clusters. It is an exploratory data mining process used in many fields for discovering patterns in data. Cluster analysis is generally of two types: hierarchical and non-hierarchical. Non-hierarchical methods use a distance based metric to cluster the objects based on their similarity and produce one-level partitioning and non-overlapping spherical shaped clusters. The most popular example is K-means clustering. Hierarchical methods on the other hand do not provide a single partitioning of the data but instead provide an extensive hierarchy of clusters that merge with each other at certain distances (Johnson and Wichern, 2007). An advantage of the hierarchical method is that unlike the K-means algorithm, it does not require the number of clusters to be specified in advance. Hierarchical clustering can be done in either bottom-up (agglomerative) fashion or top-down (i.e. divisive) fashion (Johnson and Wichern, 2007).

This study uses agglomerative hierarchical clustering which begins with every object being a cluster. At successive steps, similar clusters are merged and the algorithm ends with all objects ending up in one cluster (Landau and Everitt, 2004). The objects are clustered into groups on the basis of crop production so that objects in the same cluster have similar levels of production that differ significantly from that of other clusters. Clustering is performed separately for rice, fruits and vegetables. The objects in this study are either districts or regions depending on whether the clustering variable is rice production or fruit or vegetable production.

Dissimilarity between the districts is measured in terms of the Euclidean distance since the clustering variable 'production' is continuous. When forming clusters with more than one case it is necessary to define distance between pairs of clusters. This study uses the average linkage within groups method illustrated in Figure 1. According to this method, the distance between two clusters is defined as the averages of the distances between all pairs of objects, including pairs from the same cluster. That is, given $X1, X2, \cdots, Xn$ and pairwise distances dij , the average linkage score for groups G and H is given by (Hastie et al., 2008)

$$d_{avg}(G,H) = \frac{1}{(n_G + n_H)(n_G + n_H - 1)} \sum_{i \in G \cup H} \sum_{j \in G \cup H, i \neq j} d_{ij} \tag{1}$$

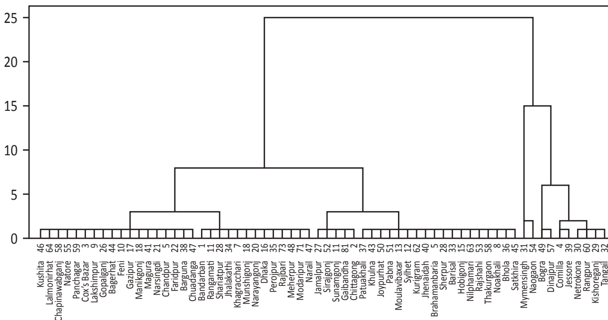
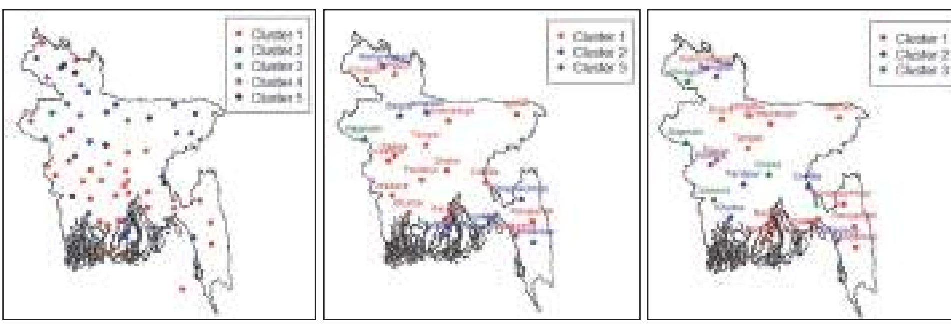


Figure 2. Dendogram of Districts in Bangladesh According to Production of Rice.



(a) (b) (c)

Figure 3. Spatial Distribution of Clusters According to Production of (a) Rice (b) Fruits (c) Vegetables.

where *nG* and *nH* are the number of points in clusters *G* and *H*, respectively. Thus, the average linkage within groups method combines clusters *G* and *H* so that the average distance between all cases in the resulting cluster *G [H* is as small as possible. By looking at all object-object similarities within *G [H*, including those that occur within *G* and those that occur within *H*, this method ensures that the resulting cluster is coherent.

5. RESULTS

An important objective of this study is to identify the number of clusters among the districts or regions of Bangladesh based on the productivity of rice, fruits and vegetables. In hierarchical cluster analysis, a useful visual representation that provides information on the number of clusters is the dendogram (Johnson and Wichern, 2007). This section provides separate dendograms for the three types of crops and summarizes results regarding number of clusters, cluster membership of the districts or regions, the spatial distribution of the clusters as well as the average productivity per cluster.

Table 1. Cluster Membership of Districts According to Rice Production

Cluster 1		Cluster 2		Cluster 3	Cluster 4	Cluster 5
Kushtia	Chuadanga	Jamalpur	Barisal	Mymensingh	Bogra	Comilla
Lalmonirhat	Bandarban	Sirajgonj	Hobigonj	Naogaon	Dinajpur	Jessore
Chapinawabganj	Rangamati	Sunamgonj	Nilphamari			Netrokona
Natore	Shariatpur	Gaibandha	Rajshahi			Rangpur
Panchagar	Jhalakathi	Chittagong	Thakurgaon			Kishoreganj
Cox's Bazar	Khagracchari	Patuakhali	Satkhira			Tangail
Lakshimpur	Munshigonj	Khulna	Bhola			
Gopalganj	Narayangonj	Joypurhat	Noakhali			
Bagerhat	Dhaka	Pabna				
Feni	Perojpur	Moulavibazar				
Gazipur	Rajbari	Sylhet				
Manikgonj	Meherpur	Mymensingh				
Magura	Modaripur	Sylhet				
Narsingdi	Narail	Kurigram				
Chandpur		Jhenaidah				
Faridpur		Brahamanbaria				
Barguna		Sherpur				

5.1. Rice

The dendogram is a diagram that shows the distance at which clusters are combined and is often read from left to right. Figure 2 shows the dendogram of the 64 districts where the clustering variable is rice production measured in metric tons. The horizontal lines show joined clusters and the position of the line on the scale indicates the distance at which the clusters are joined. The original distances are rescaled to fall within the range of 1 to 25. We wish to determine the stage at which the distance between clusters combined is large because such large gaps arguably indicate natural clusterings. In other words, we cut the dendogram where there is a sudden large jump in the rescaled distances (D) at which the clusters combine. Following this strategy, we cut the dendogram in Figure 2 at D = 5. This yields five clusters of the districts according to rice production.

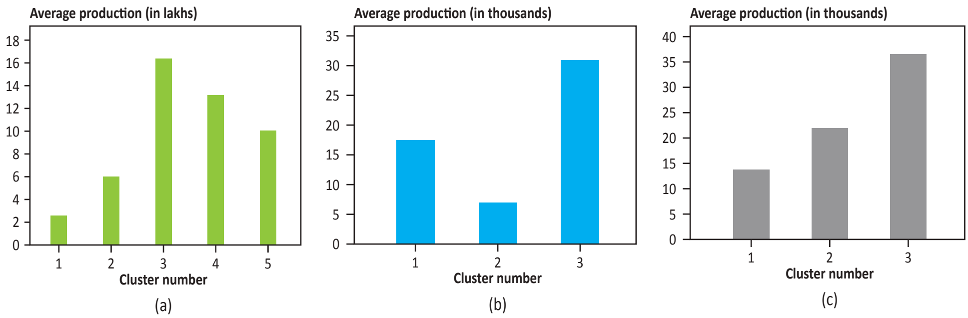


Figure 4. Average Production in a Cluster (in Metric Tonnes) for (a) Rice (b) Fruits (c) Vegetables .

Table 1 shows cluster membership of the districts. Clusters 1 and 2 are the largest consisting of 31 and 25 districts, respectively. Cluster 5 contains five districts while clusters 3 and 4 each contain two districts. Figure 3(a) shows the spatial distribution of these clusters. Cluster 1 runs in a West- Easterly direction occupying much of the central region of the country. Cluster 2 spans the north of the country with the exception of a few coastal districts in the south. Unlike clusters 1 and 2, the five districts in Cluster 5 are widely dispersed spatially. The cluster ranges from Rangpur, which is located in the north-west tip of Bangladesh, to as far as Jessore in the south and Comilla in the east. On the other hand, the geographical distances between the members of cluster 1 and also between the members of cluster 2 are relatively small.

Figure 4 (a) shows the average production of rice in each cluster. On average, cluster 3 has the highest productivity followed by cluster 4. In contrast, cluster 1 is the least productive. From these results and theresults provided in Table 1, it appears that the average productivity of a cluster is inversely proportional to the cluster size.

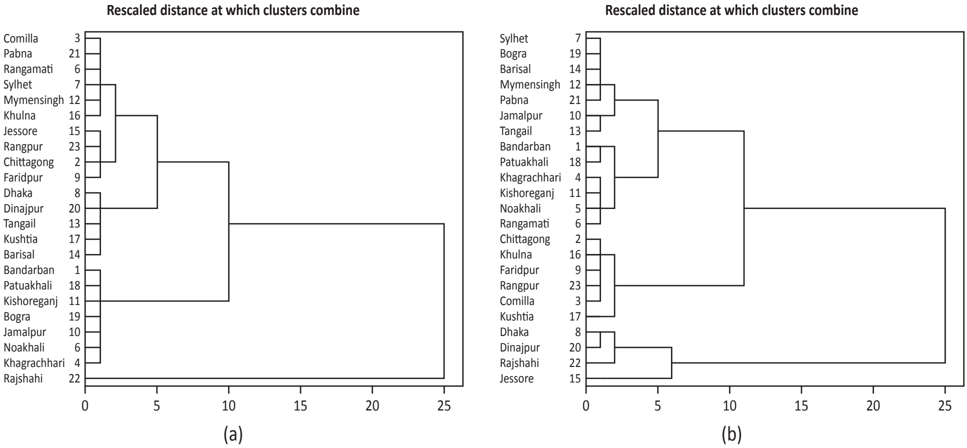


Figure 5. Dendogram of Regions in Bangladesh According to Production of (a) Fruits (b) Vegetables.

5.2. Fruits

Figure 5 (a) shows the dendogram for fruit production. There appears to be a significantly large jump beyond D = 5, which indicates the existence of three clusters. Regions belonging to each cluster for fruit production are listed in Table 2. Cluster 1 is of the largest size consisting of 15 regions while cluster 3 is the smallest being a single member cluster. In terms of fruit production, Rajshahi stands out from the rest of the regions and forms cluster 3. It is evident from Figure 3 (b) that cluster 1 covers a large geographical area. It extends from Rangpur and Dinajur in the north to Chittagong in the south, and from Kushtia and Jessore in the west to as far as Sylhet and Rangamati in the east. Cluster 2 is divided between the north and southern regions of Bangladesh. Figure 4 (b) indicates that the average production of fruits in a region is highest for cluster 3 compared to the regions in the other clusters. Thus, Rajshahi can be singled out as the most productive region for fruit production in Bangladesh.

5.3. Vegetables

The results of cluster analysis of the regions according to vegetable production are summarized in the dendogram given in Figure 5 (b). We cut the dendogram at D = 7 because the gap between two successive distances at which the clusters combine is largest beyond this point. This suggests that the regions of Bangladesh can be divided into three groups based on vegetable production. Table 2 summarizes cluster membership of each of the regions. Cluster 1 is the largest and cluster 3 the smallest. According to Figure 4 (c), the average production of vegetables in a cluster is largest for cluster 3 and smallest for cluster 1. From Figure 3 (c) it is evident that most of the regions in cluster 3 lie along the western belt of the country.

6. CONCLUSION

In an agro-based developing economy such as Bangladesh, wastage of agriculture produce is something that cannot be afforded. Yet every year tonnes of crops are being wasted due to lack of infrastructure, adequate storage facilities and development of agro-based industries. To facilitate improvement on each of these fronts, this study has identified whether clusters of districts or regions exist that require special attention. Results revealed two important clusters in terms of rice production, namely, (Mymensingh, Naogoan) and (Bogra, Dinajpur) that should be targeted for development of rice grain storage facilities.

In terms of fruit production, Rajshahi region deserves special attention being a single-member cluster and having the highest production. The government should prioritize Rajshahi by developing a greater number of cold storage facilities in the region, and taking measures to ensure efficient handling, packaging and marketing of fruits and quick dispatch to other regions. A more effective strategy for improving the overall economy of Rajshahi is to encourage the development of fruit-based industries in this area. The highest yielding cluster for vegetables consisted of only 4 regions, namely, Dhaka, Dinajpur, Rajshahi and Jessore, which are geographically well dispersed. The government needs to prioritize these regions when allocating resources for development of new cold storage facilities and transport systems.

Table 2. Cluster Membership of Regions According to Fruit and Vegetable Production

Fruits			Vegetables		
Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Comilla	Bandarban	Rajshahi	Sylhet	Chittagong	Dhaka
Rangamati	Jamalpur		Bogra	Khulna	Dinajpur
Sylhet	Kishoreganj		Barisal	Faridpur	Rajshahi
Pabna	Patuakhali		Mymensingh	Rangpur	Jessore
Faridpur	Noakhali		Pabna	Comilla	
Khulna	Bogra		Tangail		
Rangpur			Bandarban		
Jessore			Patuakhali		
Chittagong			Khagrachhari		
Tangail			Kishoreganj		
Dhaka			Noakhali		
Dinajpur			Rangamati		
Kushtia					
Barisal					

REFERENCES

Bangladesh Bureau of Statistics, 2016. Yearbook of Agricultural Statistics 2014, Series 26, Dhaka.

Board of Investment Bangladesh, 2014. Doing Agribusiness in Bangladesh: Flourishing Agriculture for Blossoming Bangladesh, Dhaka.

Food and Agricultural Organization, 2017. Key facts on food loss and waste you should know!,

< http://www.fao.org/save-food/resources/keyfindings/en/>.

R.A. Johnson and D.W. Wichern (2007). Applied Multivariate Statistical Analysis, 6th edition, Pearsons Education Inc.

S. Landau and B.S. Everitt (2004). A Handbook of Statistical Analysis Using SPSS, 6th edition, Chapman & Hall Inc.

T. Hastie, R. Tibshirani and J. Friedman (2008). The Elements of Statistical Learning: Data mining, Inference and Prediction, 2nd edition, Springer.

Spline Interpolation for Forecasting World Tuna Catches

Boonmee Lee*
Faculty of Science and Technology, Prince of Songkla University, Pattani, Thailand – bm_lee2001@yahoo.com
Don McNeil
Faculty of Science, Macquarie University, Sydney, Australia – don.mcneil@mq.edu.au
Apiradee Lim
Faculty of Science and Technology, Prince of Songkla University, Pattani, Thailand – apiradee.s@psu.ac.th

Abstract

The future of sustainable tuna fisheries is important to the global economy. To date, vast statistical fishery research has focused on stock assessment and policy evaluation contributing to the management of tuna fishing. We propose a univariate time series analysis to develop a dynamic statistical tool for monitoring annual trends in fluctuating tuna catches, using available data from the UN's Food and Agriculture Organization (FAO). Our proposed approach involves fitting an autoregressive-spline model to annual tuna catches and determining which degree of spline functions provides the most plausibility for reliable forecasts based on the data. A quintic, or fifth-degree, natural spline function, is chosen because it corresponds to the linearly decreasing trend that existing data shows in the rates of tuna catches and allows for a more practical model, provided the knots are selected to ensure that conditions necessary for long-term sustainable forecasts are satisfied.

Keywords: Sustainable tuna fisheries; Univariate time series forecasting; Cubic spline, Quintic spline.
Journal of Economic Literature (JEL) Classification: C-Mathematical and Quantitative Methods

1. INTRODUCTION

Tuna fisheries have operated on many coasts in the world since ancient times. These coastal fisheries have caught tuna seasonally as tunas are warm-temperate and highly migratory finfish. In the 1950s, an increasing demand for tuna to supply canneries drove the industrialization of many tuna fisheries, especially Japan's longline fisheries in the Pacific Ocean. These Japanese longliners rapidly expanded their fishing area and reached the Atlantic Ocean in the late 1950s (FAO, 2004). Within three decades, many new and large-scale longliners and purse seiners from South Korea, Taiwan, Spain, France, the United States and South America entered into this tuna fishery industry. Meanwhile Japan had continuously developed super-cold storage and deep longlines to enhance its capacity to supply its own newly established sashimi market. These growing fishing efforts caused the number of annual catches of 20 species of tuna and tuna-like catches to jump from 0.6 million to 3.5 million at the end of the 1980s (FAO, 2016). The increase was due not only to the greater number of multi-national fishing vessels concentrating on catching tuna, but also to the enormous growth in fishing efficiency that resulted from the use of auxiliary equipment such as bird radar, helicopters and fish-aggregating devices (FADs). FADs can locate and attract fish with minimal search time and operating costs, and their use was widespread from the Atlantic to the Indian and Pacific Oceans in the 1990s. By the end of that decade, the total catches reached 5 million tonnes. There has been an overabundance of purse-seine fishing documented in all oceans (Reid et al., 2005), and tuna canneries have experienced much smaller-sized tuna landed during past years. It is obvious that in the aggregate tuna fisheries are facing a problem of overexploited stock. The International Seafood Sustainability Foundation (ISSF) reported that 39 percent of tuna stocks in 2014 was overfished (ISSF, 2016), compared to 35 percent in 2013 (ISSF, 2015). In other words, 4 percent more of tuna stocks was exploited within a year's time.

Tuna is a highly traded commodity in the global seafood industry. The value-at-landing of tuna catches is more than 10 billion USD a year. The principal commercial tuna species account for 8 percent of world's international seafood trade (FAO, 2013). Tuna consumption is growing significantly as a source of nutritious food, especially in U.S. and EU markets. Economically, the tuna supply-chain has ties with many of both developed and developing countries. The tuna industry encompasses more than 100 fishing countries, more than 200 tuna processing companies, and a large number of tuna traders around the world. With awareness of the economic importance and food security potential this nutritious fish species provides, the UN's Food and Agriculture Organization (FAO) has for years put great effort into improving tuna fishery management. This includes facilitating international cooperation in the establishment of regional fishery management organizations (RFMOs), especially the Indian Ocean Tuna Commission (IOTC) in 1996 and the Western and Central Pacific Fisheries Commission (WCPFC) in 2004. In promoting the future sustainability of tuna fisheries over the years, the FAO and these RFMOs have published numerous periodic reports, developed statistics, and built academic networks and significant capacities in tuna research. Of 88 academic articles related to tuna fisheries published during 2000-2016, 76 percent were published in three major journals: Marine Policy (33 articles), Fisheries Research (21 articles), and Fisheries Oceanography (13 articles). These tuna-research articles focus mainly on 1) stock assessment and the biology of specific species in specific oceans, 2) fishing efforts and the capacity of specific fishing gear or gear related to specific countries, and 3) fishery development and fishery policy evaluation in specific countries or specific oceans. The research results offer scientific information and many exclusive findings useful for fishery policy makers like RFMOs. Currently, several fishing regulations and policies are ongoing, such as catch quota, three-month FAD fishing bans in purse seine fishing operated in the Western and Central Pacific, and the international fight against illegal, unreported and unregulated (IUU) fishing. However, policy implementation is not simple, as it involves multiple fleets fishing over distant waters under different national jurisdictions.

In parallel with those ongoing big efforts, there is a smaller effort to utilize existing statistics from the FAO and RFMOs for developing a dynamic and interactive model to enable the monitoring of trends in world tuna catches that have implications for managing the sustainability of one or more fisheries. Such a model would be expected to offer an alternate method of providing a quick view, especially for non-academic stakeholders, of the plausible status of sustainable tuna fisheries in the tuna industry. As part of this model development, this paper aims to analyze patterns of world tuna catches and to propose statistical methods of predicting a trend in sustainable tuna catches.

2. MATERIAL AND METHODS

2.1. World Tuna Catch Time Series

Time series data of world tuna capture are publicly available in global fishery and aquaculture statistics called "FishstatJ", which has been developed and maintained for more than 60 years by the Fisheries and Aquaculture Department of the UN's Food and Agriculture Organization (FAO). These datasets contain annual tonnages of both capture and aquaculture recorded by species, fishing countries and fishing areas. The world tuna capture used in this analysis are retrieved from the latest datasets released in March, 2016 (FAO, 2016) and cover annual catches of 65 years, from 1950 through 2014, for an aggregation of commercially important species, including skipjack, yellowfin, bigeye, albacore and bluefin caught by all fishing countries from all fishing areas.

2.2. Autoregressive-Spline Model

A simple linear regression model was first fitted to the annual time series tuna-catch data, and different power transformations of the series were tried, to stabilize variance. As an autoregressive integrated moving average (ARIMA) model is widely used to forecast fishery landings and catches (Koutroumanidis et.al., 2006; Lloret et.al., 2000; Pierce and Boyle, 2003; Stergiou et.al., 1997), diagnosing an autocorrelation of the noise is thus also included by plotting both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of residuals from the model.

In the 65-year historical time series of world tuna catches, there is a change point where the annual catches had been increasing rapidly in earlier periods but then began slowing down. For this particular non-linear time series, the ARIMA model alone is not enough to express this point since it assumes the same structural relationships across the observation and forecasting periods. Other researchers' time series analyses in fish landings faced similar non-linearity, and these researchers improved their model performance differently. For example, Koutroumanidis et al. (2006) incorporated fuzzy logic into a univariate forecasting model that fit the monthly volume of landed catches of anchovy, hake and bonito in Greece. Instead of modeling time-series data with a simple ARIMA model alone, Gutiérrez-Estrada et al. (2007) compared several computational neural network (CNN) models to forecast monthly anchovy catches in the north area of Chile, and their results showed that combining the ARIMA model with CNN provides a more precise forecast for a short- to medium-term time period. Accordingly, to improve the performance of a model that ultimately will provide a tool for *long-term* sustainable tuna fishery management, we propose to integrate a natural spline function into commonly used linear regression models. Such integration has enabled us to interpolate change points within time series documenting annual catches and extrapolate for longer periods as $Y_t=S(t)+Z_t$ where t is time (year), Y_t is the total tuna catches at time t; $S(t)$ is a natural spline function and Z_t is the random component, assumed to follow an independent, normally distributed process with mean 0 and constant variance σ^2. A spline function contains piecewise polynomials of degree n connecting together at so-called *knots*. Thus the spline function in our proposed model enables one to see the continuity between polynomial segments joined together forming a smooth curve, ultimately enabling practical approximations for real-world applications (Rice, 1969).

In general, using a spline function in a data analysis allows statisticians to fit data into a curve that provides reliable results with no loss of information and no bias due to oversimplification of the model (Wold, 1974). The natural polynomial smoothing splines of degree *2m-1* have m^{th} derivatives on an interval join data at specified knots where they minimize the integral of their squared m^{th} derivative over the interval in order to make the smoothest curve. When m is 1, the 1st degree spline simply gives straight lines joining values at a specific knot. When m is 2, the calculated value yields the 3rd degree, or *cubic*, spline. And when m is 3, the resulting spline derives from the 5th degree, or *quintic* spline. Granting that a cubic spline is most common in smoothing a forecasting curve (Lukas et.al., 2016), both cubic spline and quintic spline are experimented with and compared in our specific integrated autoregressive-spline model applied to tuna catches, and these two spline functions are rewritten and their properties described.

Case 1: Cubic spline. The 3rd degree polynomial spline function minimizes the integral of the squared second derivative over the interval between knots, and it provides linear extrapolation outside the knots. The natural cubic spline function can be expressed in this mathematic form:

$$S(t) = a + bt + \sum_{k=1}^{p-2} c_k \left[(t-t_k)_+^3 - \frac{(t_p - t_k)}{(t_p - t_{p-1})}(t - t_{p-1})_+^3 + \frac{(t_{p-1} - t_k)}{(t_p - t_{p-1})}(t - t_p)_+^3 \right]$$

where p is number of knots, $t_1 < t_2 < \dots < t_p$ are specified time knots and $(t-t_k)_+$ is $t-t_k$ for $t>t_k$ and 0 otherwise, and a, b and c_k are parameters to be estimated.

Case 2: Quintic spline. The 5th degree polynomial spline function minimizes the integral of the squared third derivative over the interval between knots, and it provides quadratic extrapolation outside the knots. The natural quintic spline function can be expressed in this mathematic form:

$$S(t) = a + bt + ct^2 + \sum_{k=1}^{p-3} d_k \left[\begin{aligned} &(t-t_k)_+^5 - \frac{(t_{p-1}-t_k)(t_p-t_k)}{(t_{p-1}-t_{p-2})(t_p-t_{p-2})}(t-t_{p-2})_+^5 \\ &+ \frac{(t_{p-2}-t_k)(t_p-t_k)}{(t_{p-1}-t_{p-2})(t_p-t_{p-1})}(t-t_{p-1})_+^5 \\ &- \frac{(t_{p-2}-t_k)(t_{p-1}-t_k)}{(t_p-t_{p-2})(t_p-t_{p-1})}(t-t_p)_+^5 \end{aligned} \right]$$

where a, b, c and d_k are parameters to be estimated.

The R statistics program is a tool used for developing the statistical model in this study since it is programmable and it has an *arima* function capable of modeling simultaneously both the signal (linear regression with spline) and the noise (autoregression) of tuna-catch time-series as needed.

2.3. Model Selection and Performance Validation

Using spline functions to fit curves in the model offers multiple forecasting outcomes because positions of the knots are free parameters. Thus, our study is designed to implement a required minimum number of knots for each application of the spline function, e.g.: 3 knots for a cubic spline and 4 knots for a quintic spline. Additionally, knot selection is based on two considerations: 1) *goodness of fit* – both the adjusted r-squared and mean-absolute-percentage-error (MAPE) are used to verify how well the model fits past data. A small program was developed to compute the statistical measures for the combination of knots used in the model; 2) *plausibility of forecast* – plausibility is arrived at from joint considerations of updated tuna fishery practices, stock assessment, and the implementation of fishery management. The answer may include some subjective considerations that complement statistical modelling.

3. RESULTS

The *volume* of world tuna catches shows a growing trend over 65 years (1950-2014). The tonnage of the principal tuna species accounted for 81 percent of all tuna and tuna-like species and reached 5.2 million tonnes in 2014 (Fig. 1). However, average *growth rates* in the number of these catches decreased (Fig. 2), particularly in the 2000s, when the average growth rate of catches was 1.4 percent, greatly lower than rates in 1990s (3.4 percent) and 1980s (5 percent). As depicted in Fig. 3, tuna fisheries have been expanded to all oceans. A total of 66.4 percent of these catches, over a 65-year span, were from the Pacific Ocean, mainly (35.6 percent) from the West Central Pacific area. Among the commercially important tuna species, skipjack is the most caught species (47.5 percent), about 1.1 million tonnes a year in average; yellowfin next (30.5 percent), and big eye next (10.6 percent).

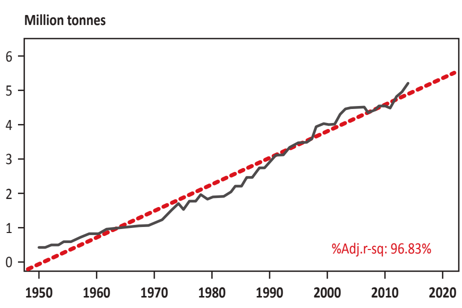


Figure 1. Annual Tuna Catches During 1950-2014 (Solid Line) and Linear Trendline (Dashed Line).

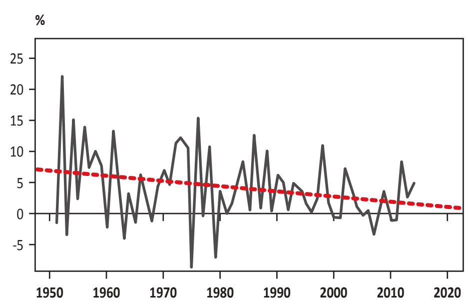


Figure 2. Percent Changes of Annual Tuna Catches (Solid Line) with a Linear Trendline (Dashed Line).

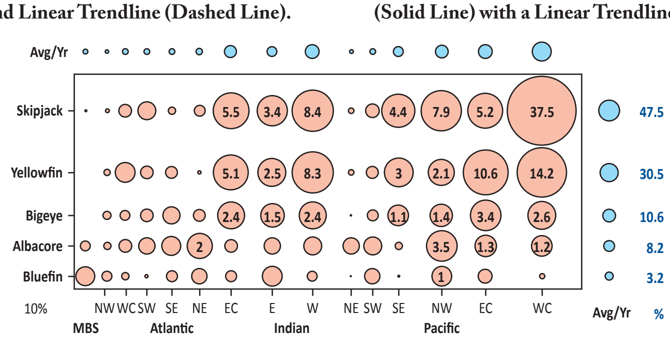


Figure 3. Bubble plot of Total Tuna Catches During 1950-2014 in a Matrix of Tuna Species by Fishing Area.

This particular trend was first assessed through an examination of its normal distribution using different types of data transformations, including logarithms, square roots and cube roots. The normal Quantile-Quantile plots of residuals from a simple linear regression model of annual tuna catches demonstrated that a cube-root transformation provides more symmetrical distributions than others. The examination of autocorrelation (ACF) and partial autocorrelation (PACF) coefficients against the number of lags for the cube-root transformed data on the volume of tuna catches indicates one-year lagged correlation (Fig. 4b). AR(1) process removed the significant autocorrelation (Fig. 4b) and provided a good fit to the error component.

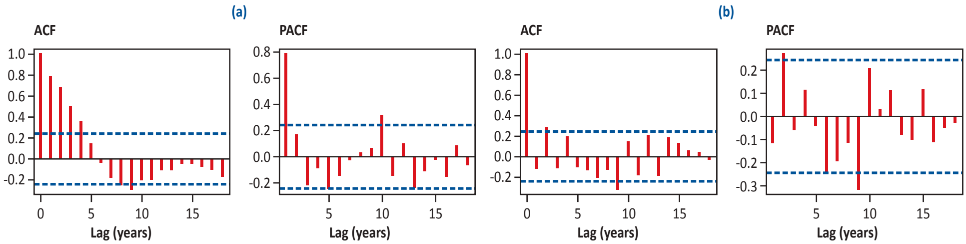


Figure 4. Diagnostic Plots for Autocorrelation of The Model's Residuals: ACF and PACF before (a) and after (b) Fitting an Autoregressive Model to Tuna Catches.

As noted above in section 2.3, the AR(1)-spline model generates many possible forecasting results, since knot positions are free parameters. Table 1 illustrates four examples of the result of incorporating a cubic spline into our model and another five examples of incorporating a quintic spline. Both MAPE and adjusted r-squared are two key statistical measures used to screen possible results. In the case of the cubic spline integration, the combination of knots in No. 4 has the lowest MAPE (1.615). No. 3 has the highest adjusted r-squared (0.99403). But the distance between knots in both options represents only one period in the time series recorded. The knots occupy almost the same positions in both instances and do not benefit from the curve-fitting properties of the cubic spline. Next, equispaced knots is considered and offers two more choices (No. 1 and No. 2), which have very high adjusted r-squared and very low MAPE. No. 1 with three knots (1992, 2003, 2014) is preferred as representative of results from applying the AR(1)-cubic spline since it has better statistical measures, 0.99372 adjusted r-squared and 1.652 MAPE. The model in option No. 1 displays a linear trend in future catches of tuna, continually growing with an annual rate of 1.2 percent as shown in Fig. 5.

Table 1. Comparitive Results from The AR(1)-Spline Models

No.	Knot position				Distance between knots			Optimal catches		% change of catches after the optimal	%Adj. r-squared	MAPE	Forecasts		
	k1	k2	k3	k4	k2-k1	k3-k2	k4-k3	year	tonnes				2017	2018	2019
Model options with a cubic spline															
1	1992	2003	2014		11	11					99.3718	1.6520	5195	5258	5321
2	1994	2004	2014		10	10					99.3692	1.6546	5160	5215	5270
3	2001	2002	2003		1	1					99.4033	1.6188	5214	5284	5356
4	2002	2003	2004		1	1					99.4001	1.6151	5172	5235	5299
Model options with a quintic spline															
5	1977	1998	2000	2009	21	2	9	2023	5272	-0.0130%	99.3334	1.6866	5160	5192	5219
6	1977	1999	2001	2009	22	2	8	2023	5240	-0.0990%	99.3333	1.6864	5150	5180	5203
7	1978	1990	2002	2014	12	12	12	2024	5307	-0.0507%	99.3298	1.6898	5174	5209	5239
8	1978	1989	2000	2011	11	11	11	2026	5408	-0.0149%	99.3273	1.6930	5198	5240	5278
9	1979	1990	2001	2012	11	11	11	2024	5321	-0.0032%	99.3326	1.6871	5175	5212	5243

In the case of the AR(1) model with quintic spline incorporated, the same conditions were executed to filter possible results, and option No. 6 resulted in the lowest MAPE (1.686), No. 5 has the highest adjusted r-squared (0.99333), and Nos. 7-9 have equally spaced knots. Option No. 9 is chosen as representative of results from the AR(1)-quintic spline, since both adjusted r-squared (0.99332) and MAPE (1.687) are most satisfied. As depicted in Fig. 6, the model with four knots (1979, 1990, 2001 and 2012) predicts annual catches of those principal important tuna species in a quadratic curve. The AR(1)-quintic spline with these four specific chosen knots indicates the volume of catches will continue increasing to the optimal point at 5.3 million tonnes in the year 2024 and thereafter remain almost constant with a very small change in annual catches, -0.0032 percent.

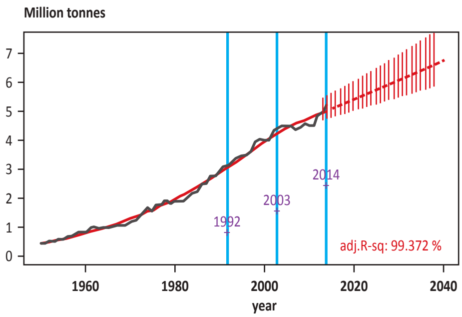


Figure 5. Tuna Forecasts from AR(1)-Cubic Spline Model (dot line) with 95% Confidence Interval

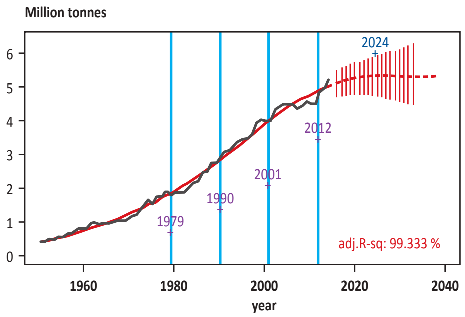


Figure 6. Tuna Forecasts from AR(1)-Quintic Spline Model (dot line) with 95% Confidence Interval

4. DISCUSSION

Tuna fisheries contribute significantly to our global economy. This project was initiated to add a new dimension to forecast modeling. It is hoped that the results will ultimately contribute to ongoing efforts to manage both the supply and demand critical to sustaining tuna fisheries the world over. To help moderate the tuna supply, many fishing regulations and practices have been implemented, including controlling a number of fishing vessels, banning the use of FADs for a certain period, and opposing IUU fishing. Meanwhile, some environmental strategies have been put in place to shape the behavior and cooperation of consumers in these management efforts. Such strategies include eco-labelling and promoting awareness of the importance of sustaining tuna as a natural resource. However, stabilizing tuna stock at a sustainable level is very challenging and needs close monitoring. Principal steps toward stabilization include encouraging and enforcing related stakeholders' commitment to the regulations and recommendations in the FAO's and RFMOs' management plans. Because data on tuna catches can reflect changes in fishing efforts, management measures, consumer behaviors, and the abundance or depletion of tuna stock, a simple, dynamic and interactive modeling tool to interpret this data has been needed. Our goal has been to establish a reliable and realistic picture of trends in catches that can be compared to estimations of sustainable levels of volume and rates needed for the perpetuation of the industry.

In the proposed autoregressive-spline model, a natural spline function is applied to interpolate the univariate time series data on hand in order to project a trend in world tuna catches. This fishery data is aggregated into annual tonnages and is thus less volatile than daily or monthly time series. The number of knots within the minimum required for a spline function is easily chosen. As noted above in section 2.3, both goodness of fit and plausibility of forecast determine the positions of the knots. Generalized cross-validation (GCV), not included in the model yet, should be explored and added to further studies since it is one of the most popular parameter selection criteria in knots selection (Ruppert, 2002; Lukas et.al., 2016).

5. CONCLUSIONS

Our integration of an autoregressive model with a spline function facilitates time-series, data-driven forecasting. The developed model takes advantage of natural spline smoothing not only to interpolate change points within time series of tuna catches but also to extrapolate future catches. By comparing AR(1)-cubic spline and AR(1)-quintic spline models, we observe that the function of a cubic spline may be simpler than applying a quintic spline, but forecasts resulting from the application of quintic spline are more accurate and better represent long-term changes in volume and rates. This is because a quintic spline function provides quadratic extrapolation outside the knots, corresponding to the linearly decreasing trend in the rates of tuna catches that existing data show (Fig. 2). This integrated model, developed with the R program, provides flexibility in its application. For example, we can use it to project catch volume and rate for individual fishing areas or individual species. Another example is its use as a convenient and quick alternative tool for measuring trends in tuna fishing efforts against the baseline of maximum sustained catches regulated by related RFMOs. Above all, the forecasting results expressed with the proposed model will be useful for planning to ensure the sustainability of the tuna industry.

REFERENCES

FAO. (2013). Tuna: A global perspective. http://www.fao.org/docrep/017/ap939e/ap939e.pdf (Accessed 09 March 2016).

FAO. (2016). Fishery and Aquaculture Statistics. Global production by production source 1950-2014 (FishstatJ). In: FAO Fisheries and Aquaculture Department [online]. Rome. Updated 2016. http://www.fao.org/fishery/statistics/software/fishstatj/en (Accessed 16 April 2016).

Georgakarakos, S., Koutsoubas, D., & Valavanis, V. (2006). Time series analysis and forecasting techniques applied on loliginid and ommastrephid landings in Greek waters. *Fish. Res,* 78(1), 55-71.

ISSF (2015). ISSF Tuna Stock Status Update, 2015: Status of the world fisheries for tuna. ISSF Technical Report 2015-03A. International Seafood Sustainability Foundation, Washington, D.C., USA.

ISSF. (2016). ISSF Tuna Stock Status Update, 2016: Status of the world fisheries for tuna. ISSF Technical Report 2016-05. International Seafood Sustainability Foundation, Washington, D.C., USA.

Koutroumanidis, T., Iliadis, L., & Sylaios, G.K. (2006). Time-series modeling of fishery landings using ARIMA models and Fuzzy Expected Intervals software. *Environ. Model. Softw,* 21, 1711-1721.

Lloret, J., Lleonart, J., & Sole, I. (2000). Time series modeling of landings in Northwest Mediterranean Sea. *ICES J. Mar. Sci,* 57, 171-184.

Lukas, M.A., Hoog, F.R., & Anderssen, R.S. (2016). Practical use of robust GCV and modified GCV for spline smoothing. *Comput. Stat,* 31, 269–289.

Pierce, G.J., Boyle, & P.R. (2003). Empirical modeling of interannual trends in abundance of squid (Loligo forbesi) in Scottish waters. *Fish. Res,* 59 (3), 305-326.

Reid, C., Kirkley, J.E., Squires, D., & Jun Y. (2005). An analysis of the fishing capacity of the global tuna purse-seine fleet. Proceedings of the Second Meeting of the Technical Advisory Committee of the FAO Project Management of Tuna Fishing Capacity: conservation and socio-economics. Fisheries Proceedings No.2. FAO, Rome.

Rice, J.R. (1969). The approximation of functions, vol. II. Addison-Wesley, Massachusetts.

Ruppert, D. (2002). Selecting the Number of Knots for Penalized Splines. *J. Comp. Graph. Stat,* 11(4), 735–757.

Stergiou, K.I., Christou, E.D., & Petrakis, G. (1997). Modeling and forecasting monthly fisheries catches: comparison of regression, univariate and multivariate time series methods. *Fish. Res,* 29, 55-95.

Wold, S. (1974). Spline Functions in Data Analysis. *Technometrics.* 16(1), 1-11.

Short and Long Term Relationships among the Prices of East Kalimantan Fresh Fruit Bunches, CPO & World Crude Oil Price

Sri Wahyuningsih*
Study Program of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, East Kalimantan, Indonesia – swahyuningsih@gmail.com, swahyuningsih@fmipa.unmul.ac.id
Abdussamad
Study Program of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, East Kalimantan, Indonesia – abdussamad.stat@gmail.com
Memi Nor Hayati
Study Program of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, East Kalimantan, Indonesia – meminorhayati@yahoo.com

Abstract

This paper aims to obtain the best VECM models of the price of fresh fruit bunches from the East Kalimantan farmers, East Kalimantan CPO price, world CPO price, and the price of world crude oil. In addition, the analysis of short-term relationships and long term between the price of fresh fruit bunches from the East Kalimantan farmers, East Kalimantan CPO prices, world CPO price, and the price of crude oil is done using VECM approach. Forecasting the price of fresh fruit bunches from the East Kalimantan farmers, East Kalimantan CPO price, world CPO price, and the price of crude oil was obtained using the best VECM models. The test results on the causality cointegration equation shows there is a long-term relationship between the price of East Kalimantan CPO as the dependent variable with the price of fresh fruit bunches from the East Kalimantan farmers, East Kalimantan CPO price and crude oil prices. There is a short-term relationship in one direction on the price of fresh fruit bunches from the East Kalimantan farmers to the price of East Kalimantan CPO. This also applies to the world CPO price to the price of fresh fruit bunches from the East Kalimantan farmers; CPO world price on the East Kalimantan CPO price; East Kalimantan CPO prices on the world CPO price; and the crude oil price on the price of East Kalimantan CPO. Short-term relationship bidirectional happen to the price of fresh fruit bunches from the East Kalimantan farmers with crude oil prices. Forecast results obtained using the model VECM (1) suggests that the price of fresh fruit bunches from the East Kalimantan farmers, the East Kalimantan CPO prices, world CPO price and the price of crude oil in December 2016 until May 2017 increased over the forecast period.

Keywords: CPO price; Forecast; Relationship; VECM.

CPS27: Demography & Social Welfare Statistics (4)

Heteroscedasticity in Grouped Data: a Case of Infant Mortality Data in Indonesia

Ray Sastri
Statistics College, Jakarta, Indonesia – raysastree@gmail.com
Khairil Anwar Notodiputro
Bogor Agricultural University, Bogor, Indonesia – khairilnotodiputro@gmail.com

Abstract

In modeling using ordinary least square estimators, the presence of heteroscedasticity could result in misleading conclusions. In regression analysis where the data are obtained from grouped of individual data, heteroscedasticity problems are often undetected when the test is conducted using the ordinary heteroscedasticity test for individual data. Meanwhile, the government often requires grouped data when a policy, such as reducing infant mortality, is being made. The infant mortality grouped data is obtained by aggregating the binary categorical individual data for every district. In this paper we discuss how to detect the presence of heteroscedasticity on infant mortality data and how to develop models that can be used to explain the effects of explanatory variables on infant mortality in Indonesia. Our analysis showed that the heteroscedasticity was evidence and it was due to the differences in the number of samples in each district. The weighted logistics regression model fit the data with R square equals to 54.80 percent. The significant variables affected the infant mortality were found to be the percentage of childrens with birth order four or more, the ratio of health facilities per 1000 population, and the infant mortality in neighbor district.

Keywords: Heteroscedasticity, IDHS, Infant mortality, Weighted regression.

1. INTRODUCTION

Infant mortality is the incidence of death in infants under one year old (BPS 2013, Sastri *et.al* 2015). In Indonesian Demographic and Health Survey (IDHS), infant mortality is a binary individual data with two outcomes, 1 and 0. The outcome is 1 if the infant die under one year old, and 0 if the infant is still alive. An appropriate model for binary response variable is logistic regression.

If we are interested more in the analysis of district level rather than the individual level, we could group the individuals into district levels. So the response variable was no longer binary categorical variables but proportion of infants mortality in district levels. However, grouping the individual data could cause problems of assumptions violation in regression model especially the assumption of homoscedasticity in error variance. Violations on this assumption will produce incorrect variance estimates. Hence, a proper approach is required in modeling.

In this paper we discuss heteroscedasticity in grouped data, especially for infant mortality data in Indonesia, as well as present the best model for infant mortality data in Indonesia.

2. RESEARCH AND METHODS

2.1. Data

This research used data from Indonesian Demographic and Health Survey (IDHS) 2012. In addition, this rsearch also used linked data from the Ministry of Health.

Heteroscedastity and Weighted Regression Model

The response variable (Z) in this research was two binary categorical variables, Z = 1 if the infant died before one year old and Z = 0 if the infant still alive. If we taken n_j random sample in districts j, then we have $Z_{1j}, Z_{2j}, \dots Z_{n_j j})$ that independent of each other. Y_j is the number of infant who died in j districts, then $Y_j = \sum_{i=1}^{n_j} Z_i$ statistically have a binomial distribution with parameter (n_j, P_j). The probability of infant mortality in j district can be write as $P_j = Y_j / n_j$.

The logistic regression models are formulated as follow:

$$ln\left(\frac{P_j}{1-P_j}\right) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_{k-1} X_{(k-1)j} + \beta_k X_{kj} + \varepsilon_j$$
$$L_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_{k-1} X_{(k-1)j} + \beta_k X_{kj} + \varepsilon_j \quad (1)$$

L_j is response variable, **X** is a set of regressors and β is a set of unknown parameter.

In this paper we use OLS since it is simple and if the assumptions are met then OLS produce estimates known as BLUE. One of OLS assumption is homoscedasticity, a condition in which Var(L_j |X_j is constant. This condition can be written as follow:

$E(\varepsilon_j^2)=\sigma_j^2$ j=1,2,....n

According to Theil (1970), in logistic regression for grouped data, if the number of sample is large enough and if response variable follows the binomial distribution, then the residual will follow the normal distribution with vary variance or does not fix for all observations.

$$\varepsilon_j \sim N\left(0, \frac{1}{n_j p_j (1-p_j)}\right)$$

There are several methods that can be used to detect violation in homoscedasticity assumption. Plot between resual and fitted value could detect it. Usual formal test for that is park test, white test, and bartlet test.

Violation in homoscedasticity can make inaccurate conclusion in hypothesis testing. Suppose the regression model with one explanatory variables:

$$L_j=\beta_0+\beta_1 X_j+\varepsilon_j$$

The OLS estimators for β_1 has the following variance:

$$\text{var}(\hat{\beta}_1)=\frac{\sigma_j^2}{\sum X_j^2}$$

If the variance of error is homogenous, then the variance of $\hat{\beta}_1$ becomes:

$$\text{var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum X_j^2}$$

$\hat{\beta}_1$ is the best linear unbiased estimator for β_1 when variance of error fulfilled homoscedasticity assumption. If not, $\hat{\beta}_1$ is still linear unbiased estimator, but no longer the best because variance is not the minimum. Because its variance not the minimum, confidence interval of parameter estimators will be shorter than it should be and tend to make us to not reject the null hypothesis in hypothesis testing.

An estimation method called generalize least squares (GLS) use some information about heteroscedasticity in getting the best linear unbiased estimator.

If the variance σ_j^2 are known, and if the model (1) divided by σ_j^2 then the equation becomes :

$$\frac{L_j}{\sigma_j} = \beta_0 \frac{1}{\sigma_j} + \beta_1 \frac{X_{1j}}{\sigma_j} + \beta_2 \frac{X_{2j}}{\sigma_j} + \cdots + \beta_k \frac{X_{kj}}{\sigma_j} + \frac{\varepsilon_j}{\sigma_j}$$
$$L_j^* = \beta_0^* + \beta_1^* X_{1j}^* + \beta_2^* X_{2j}^* + \cdots + \beta_k^* X_{kj}^* + \varepsilon_j^* \quad (2)$$

L_j^* is the new response variable and X^* is new regressor. The model in equation (2) called weighted regression model. β_0^*... β_k^* are the parameters of the model after weighted. Variance of error after weighted becomes:

$$\text{var}(\varepsilon_j^*)=E(\varepsilon_j^*)^2= E\left(\frac{\varepsilon_j}{\sigma_j}\right)^2= \frac{1}{\sigma_j^2} E(\varepsilon_j^2)=\frac{1}{\sigma_j^2}\sigma_j^2=1$$

Variance of error in a weighted regression model is constant and the homoscedasticy assumptions is fulfilled, so that $\beta_0^*\dots\beta_k^*$ are best linear unbiased estimator for $\beta_0\dots\beta_k$.

2.3. Model

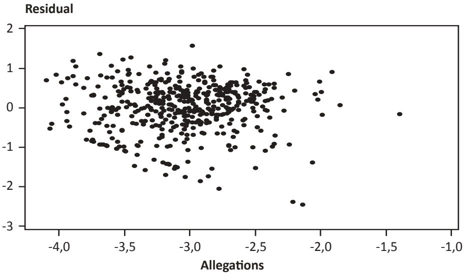
We run a logistic model with response $ln\left(\frac{p_j}{1\text{-}p_j}\right)$ being a logarithm of odds of infant mortality and various explanatory variables including:

X_1 : Sex ratio of men to women
X_2 : Percentage of children whom birth order number 4 or more (%)
X_3 : Percentage of mothers who graduated from junior high school or below (%)
X_4 : The percentage of maternal age <20 years or > = 40 years (%)
X_5 : Percentage of households with Welfare Index middle and lower(%)
X_6 : Type Region (1: District, 0: City)
X_7 : Helathcare operasional support (Million Rupiah)
X_8 : Ratio of public healthcare facilities number per 1000 Population
X_9 : The ratio of paramedics number per Population
X_{10} : logarithm for odds of infant mortality in first nearest neighbor district
X_{11} : logarithm for odds of infant mortality in second nearest neighbor district

3. RESULTS

Model selection using stepwise consists of *five significant explanatory variables*, X_2, X_4, X_5, X_8, X_{10}. From this model, the park dan white tests were carried out. The test results showed homoscedasticity in error variance but these tests were applied to individual data, not for grouped data.

Residual scatter plot provides a visualization of homoscedasticity. Figure 1 (a) shows a funnel pattern but it is not clear. If we plot the residuals with the number of children in each district, a clear pattern as in Figure 1 (b) is apparent, indicating the existence of heteroscedasticity problems.



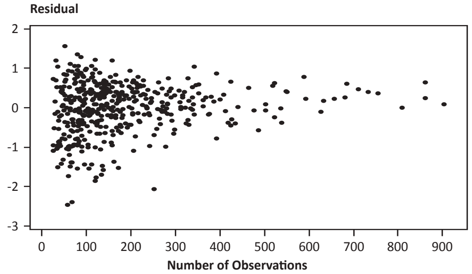


Figure 1. Scatter Plots of Residuals

To fix the heteroscedasticity problem, districts are classified into 19 classes based on number of children who ever born in each district. Characterictics of each class are shown in Table 1. Figure 2 (a) provides confidence intervals for all classes without weighting. The P-value of Bartllet test is 0.000, showing evidence of heteroscedasticity. Figure 2 (b) provides confidence intervals for all classes after weighteding is applied. The weighted model successfully fixed the heteroscedasticity problem.

Weighted logistic regression model fit the data with R squared 54.80 percent. The regression equation is as follow:

$$\widehat{L_j^*} = -2.1880 + 0.01376X_{2j}^* + 0.01034\, X_{4j}^* - 0.00154\, X_{5j}^* - 0.01750X_{8j}^* + \; 0.5986\, X_{10j}^*$$

$(p$=0.000) $(p$=0.002) $(p$=0.068) $(p$=0.457) $(p$=0.000) $(p$=0.000)

Table 1. Classification of District Based on Number of Children

Class	Range Number of Children	Number of district	Variance	Standard Deviation
1	26-37	17	0.71	0.84
2	38-50	21	0.45	0.67
3	51-62	25	0.89	0.94
4	63-75	27	0.76	0.87
5	76-86	33	0.47	0.68
6	87-100	39	0.39	0.62
7	101-113	27	0.46	0.68
8	114-125	23	0.57	0.76
9	126-137	24	0.40	0.63
10	138-150	26	0.29	0.54
11	151-175	39	0.35	0.59
12	176-200	23	0.20	0.44
13	201-225	26	0.21	0.46
14	226-250	16	0.15	0.39
15	251-275	17	0.37	0.61
16	276-300	16	0.21	0.46
17	301-400	32	0.15	0.39
18	401-500	12	0.16	0.40
19	> 500	23	0.09	0.30

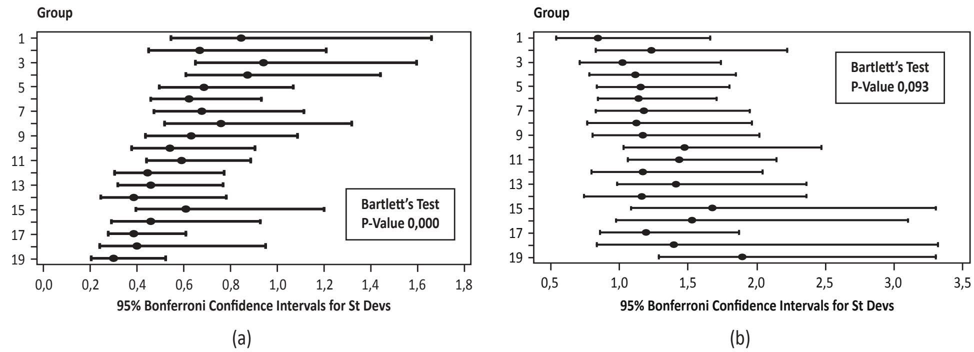


Figure 9. Confidence Interval of Each Class Before and After Weighted

4. CONCLUSION

Grouped data is usually suffered by heteroscedasticity problem and a proper weight is required to overcome the problem. We showed that in case of infant mortality data in Indonesia the appropriate weight isrelated to the number of children who ever born in each district. The resulting weighted logistic regression model fit the data well.

REFERENCES

Afri LE. 2012. *Geographically weighted Negative Binomial Regression for Data of Infant Mortality (Case Study 38 Regency/City in East Java)* [Tesis]. Bogor (ID): Institut Pertanian Bogor.

Anselin L. 1988. *Spatial Econometrics: Methods and Models (Studies in Operational Regional Science)*. New York (USA): Springer Publishing.

Ashani TA. 2010. Kematian Bayi Menurut Karakteristik Demografi dan Sosial Ekonomi Rumah Tangga di Provinsi Jawa Barat: Analisis Data SDKI 2007. [Internet]. *Jurnal Bumi Indonesia*. 1(3). [Diakses 10 Juli 2015 dari URL www.lib.geo.ugm.ac.id]

Astuti ET. 2013. Statistical Modelling for Mortality Data Using Local Generalized Poisson Regression Model. *International Journal of Applied Mathematics and Statistic (IJAMAS)*. 33(3):92-101.

[BPS] Badan Pusat Statistik. 2013. *Survey Demografi dan Kesehatan Indonesia Tahun 2012*. Jakarta (ID): BPS

Hosmer DW, Lemeshow JS. 2000. *Applied Logistic Regression.* New York (USA): John Wiley and Sons. Inc.

Gujarati D. 2004. *Basic Econometrics*. New York (USA): McGraw-Hill.

Gujarati D. 2015. *Econometrics by Examples*. London (UK): Palgrave.

Lesage JP. 1999. *The Theory and Practice of Spatial Econometrics*. Ohio (AS): University of Toledo

Pramono MS, Aditie NB, Sutikno. 2012. Regresi Spasial Durbin Model Untuk Mengidentifikasi Faktor yang Berpengaruh Pada Angka Kematian Bayi di Jawa Timur. *Buletin Penelitian Kesehatan.* 40:190 – 200.

Poerwanto S, Stevenson M, Klerk N. 2003. Infant Mortality and Familiy Welfare: Policy Implications for Indonesia. *Epidemol Community Health.* 57:493-498.

Sastri R, Notodiputro KA, Indahwati. 2015. Autologistics Models for Modelling and Mapping Infant Mortality in Indonesia. Proceeding of International Conference on Research, Implementation and Education of Mathematics and Sciences, Yogyakarta State University

Theil H. 1970. On the Estimation of Relationships Involving Qualitative Variables. *The American Journal of Sociology.* 76(1):103-154

Titaley CR. 2011. *Neonatal Mortality in Indonesia : the Protective Role of Antenatal. Delivery and Postnatal Care Services* [Tesis]. Sidney (AU): University of Sidney

Winarno D. 2009. *Analisis Angka Kematian Bayi di Jawa Timur dengan Pendekatan Model Regresi Spasial* [Tesis]. Surabaya (ID): Institut Teknologi Sepuluh November.

Trends of Age-Specific Mortality Rates for Female in 54 of the World's Most Populous Countries

Nirmal Gautam*
Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand – gnirmal655@gmail.com
Apiradee Lim
Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand– api_45@hotmail.com
Attachai Ueranantasun
Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand– attachai@gmail.com
Metta Kuning
Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand– metta.kuning@gmail.com

Abstract

If net migration is negligible, the structure of the population of the country is regulated by the births and deaths rate. Moreover, if age-specific mortality and fertility rates are known, the population for each sex can be forecasted precisely, using the simple fact that a person at the age x who survives mortality will be aged x+t in t years. In our study, we adopted these ideas and applied the departure rate method to investigate the trends of female mortality with the population data, which are freely available from the website of the US Census Bureau. The aim was to estimate the trends of age-specific mortality rates for female populations in 54 of the world's most populous countries and groups them according to their patterns. The method involved first estimating departure rates for these countries, and then applying these rates to mortality of female. The results indicated that the trends of mortality rates were very asymmetric and dramatic across the Europe, USA, Canada, Australia, few Asian, and some other African countries with the immigrant population. While these countries had a strong immigration, they showed a strong negative departure rate.

Key words: Population forecast, Departure rate, Mortality comparison
JEL classification: J10

1. INTRODUCTION

A distorting pattern of morbidity and death by place, time and cause are defined as mortality (WHO, 2015). It is a key indicator for the country to assess the health of the population, and it also depends on a mixture of the socioeconomic condition of the countries and reflects the condition of the longevity (Gardner et al., 2004; Health at a Glance 2013). Recent studies showed that most of the African and some Asian countries were a major contributor to the high number of female deaths due to the poor socioeconomic status (Johri et al., 2014). Therefore, an accurate and sensible scenario of the mortality and magnitude of the people's life is directly linked to the socioeconomic status of the country (Gardner et al., 2004)

All over the world, females live longer than males due to the biological and environment factors (Stevens et al., 2013). These factors bring differences of living of standards and inequalities in the health of both genders, which affect the improvements in survival rate and longevity of both female and male population (Currie et al., 2015; Jonas et al., 2014). In 2015, females averagely lived 73.5 years and were accounted for 54% of population at the age of 60 years and 62% of population at age of 80+ years. Furthermore, most of these elderly women live in less developed regions (United Nations 2015). However, it was different in the past scenario as in 1884s, when females had a short life span than males with the life expectancy of 40.8 and 39.5 years, respectively (Wrigley et al., 1997). It was not until the early twentieth century when this trend reversed to 67.1 to 72.3 years for female lives and 62.5 to 67.8 years for those of males, with the sex ratio of 107 males per 100 females from 1990 to 2015.

The sex ratio stayed the same until the age of 50 years. After that age, it was reversed with a huge margin (45 males per 100 females) for population aged from 80 to 84 years (Berin et al., 1989; United Nations., 2013; World's women 2015). In developed countries, unhealthy and risk behaviors contribute for higher chances of mortality in males than females (Kipping et al., 2012; Sabo et al., 1995), while, for developing countries, environmental factors contribute to a higher level of male mortalities than in those of females. Furthermore, the high excess of female mortality in the developing countries is due to the maternal complications, iron deficiency and sexual abuse (Population Reference Bureau., 2007; Global Health Risk report. 2009).

Determination the trends of the mortality, especially in females, has been conducted by many researchers and the international organizations with different statistical methods (Alkema et al., 2016). Previously, world mortality trends and patterns in older females were identified by using the life table method based on the vital registration data for older females in both developed and developing countries (Stevens et al., 2013). However, vital registration data, especially mortality registration, are not available in many countries due to the lack of a proper recording system. Therefore, with only population data being accessible, we used the simple method to identify the trends of overall female mortality in this study. This method was based on the freely provided population data of 54 countries which help to provide the complete information about the female departure as the age specific mortality rate.

2. MATERIALS AND METHODS

The data used in this study, from 1990 to 2015 of female population in 54 countries, were retrieved from US Census Bureau's website (United States Census Bureau). The selection was based on the world's most populous countries, and these 54 countries were equally selected from Asia, Africa, and Western countries with 18 countries being chosen for each region.

In our study, we applied the departure rate method based on the concept that a person at the age x in any country will be aged x+t in t years if that person is still staying or surviving in that country. To determine the trends of age-specific departure rates for female populations, we used the following equation

$$\textit{Departure rate} = 1 - \frac{N(x+1,t+1)}{N(x,t)}$$

Where $N(x, t)$ is the number of persons aged x in year t and $N(x+1,t+1)$ is the number of persons

In Coale & Demeny (1966), the life table models were basically grouped according to geographical regions of countries. The countries within the same region usually shared the same life table model. However, there could be an exception that a country from another region was included in the life table model out of its region, for example, Taiwan and Japan in North model ((Ansley et al., 1966).). We adopted the idea of grouping population trends for countries from their locations and similarities in this study. Firstly, the method began with the calculation of departure rate for each age year, starting from zero to 84, while ages over 84 was not considered in this study. The resulted departure rates for all countries were grouped into patterns, based on a level of fluctuation in the plots. To create a baseline, countries with the least fluctuating patterns were selected and then divided into a sub-group, depending on the similarity of the departure patterns. The countries with slightly deviating departure rates were considered having a small amount of migration. In each sub-group, the median of departure rates could provide a filtering mechanism to diminish the migration, and therefore represented the baseline for the departure rates in the sub-group approximately. This baseline was considered the mortality rate for each sub-group because of the cancellation of the migration parts from the departure. These medians or the mortality rates were then used to compare with the departure rates from other levels of fluctuations to form other sub-groups in each level. If the departure rate for a country was following the pattern of any particular mortality rate, that country would be added in to a corresponding sub-group. In each further sub-group, the difference between the departure rates and the mortality rate indicated migration, either immigration or emigration for each age level. If the departure rates for all 54 countries were classified into groups according to the medians or the mortality rates, it could be proven that the patterns of mortality rates could systematically identified the trends of mortality rates for each group.

3. RESULTS

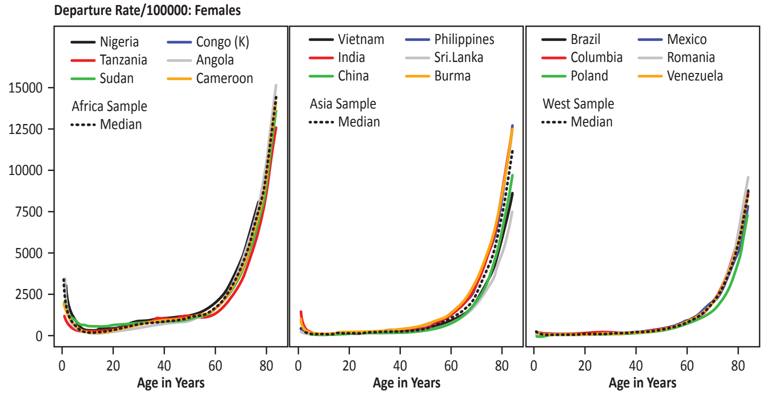


Figure 1. Trend of Female Departure Rate with Mild Fluctuation Pattern

After calculating the departure rate for 54 countries and plotting the graphs of all departure rates, we found that the departure rates could be separated into three groups, namely, mild, moderate and severe fluctuations. In the mild fluctuation group, the countries showed the smoothest patterns of female departure rates. Therefore, these countries were considered the group of the lowest activity of migration. The countries in this group were further divided into sub-groups based on the patterns of the departure rate plots. The result showed that there were three possible subgroups as shown in Figure 1. The first sub-group, with the highest departure rates at lower ages, consisted of Angola, Cameroon, Congo-Kinshasa, Nigeria, Sudan and Tanzania, while the second group comprised Burma, China, India, The Philippines, Sri Lanka and Vietnam. The group with the smallest departure rates at lower ages was the third group including Poland, Romania, Brazil, Columbia, Mexico and Venezuela. Afterwards, we further found the medians which represented the baseline for each sub-group. These medians were considered the mortality rates for each particular sub-group because of the omission of the migration parts, and hence named Africa model, Asia model and West model, respectively, according to the locations of the countries for each sub-group.

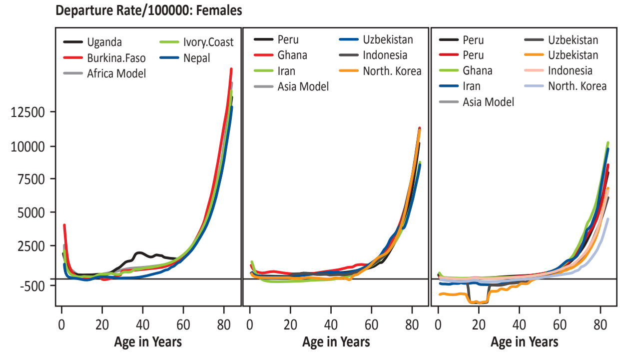


Figure 2. Trend of Female Departure Rate with Moderate Fluctuation Pattern

For the moderate fluctuation in Figure 2, there were 18 countries. The three models, achieved from the mild fluctuation group, were used to distinguish these countries by matching with the departure rates. The result revealed that all 18 countries could be separated by the three models and thus from other three sub-groups. For Africa model, there were four countries consisting of Burkina Faso, Ivory Coast, Uganda and Nepal. In this sub-group, Uganda showed the hump patterns of departure rate of female at age 20 to 60 years. Asia model consisted of Indonesia, Iran, North- Korea, Uzbekistan, Ghana and Peru, whereas there were eight countries, grouped by West model, comprising Germany, France, Russia, United Kingdom (UK), Algeria, Argentina, Egypt and Thailand. Compared to other countries in the sub-group. Within their corresponding sub-group, North Korea had some fluctuation during the productive age of females, while UK and Germany had more fluctuations of departure rates from zero to 65 years of ages.

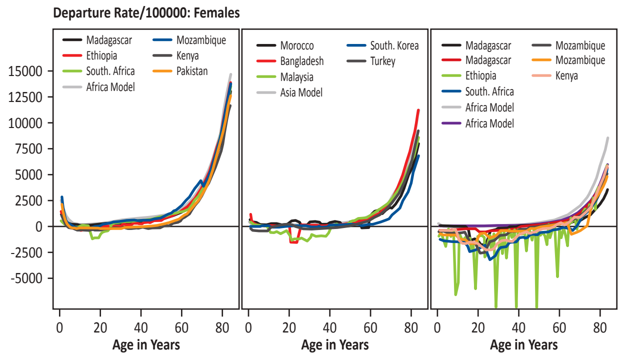


Figure 3. Trend of Female Departure Rate with Sever Fluctuation Pattern

In Figure 3, there were 18 countries with severe fluctuation of departure rates. Following the previous process, we further classified these countries into sub-groups. Subsequently, we found there were three sub-groups as all three models could match all these countries as well. For Africa model, it consisted of six countries including Ethiopia, Kenya, Madagascar, Mozambique, South Africa and Pakistan. Asia model matched five countries, consisting of Bangladesh, Malaysia, South-Korea, Turkey and Morocco, while West model fitted with seven countries - Canada, Italy, Spain, United State of America (USA), Australia, Japan and Taiwan, respectively. For this group, all countries showed erratic fluctuations of female departure rates with negative deviation mostly. In the sub-group of Africa model, South Africa revealed the "U" shape of female departure rate at 15 to around 50 years of female, whilst South Korea and Morocco displayed the steep up and down trends of departure rates in Asia model. For West model, USA had huge fluctuation of female departure rates from zero to 70 years, while Japan had the lowest fluctuation. For all of these results, it showed that the departure rates for all 54 countries were systematically classified into groups using the patterns of the median or the mortality rates. Therefore the patterns for each model could identify the trends of the mortality rate.

4. DISCUSSIONS

This study carried out the age-specific female mortality trends for the 54 countries using the straightforward method. This method outlined a new innovative and simple approach to investigating the trends of female departure rates. It was found that the mortality rates of the highest populous countries could be divided into three categories - Africa model, Asia mode and West model. Even though, the names originated from the locations of most of the countries in the groups that created the models, it was not exclusive for only countries within the same region as the name. For example, Nepal and Thailand were included in Africa model and West model, respectively. The systematic classification principally depended on the patterns of the models. The patterns of the models, when being proven their systematical repetition with all countries, could be considered the trends of the mortality rate of the models.

All the results showed the consistency findings. Africa model showed the highest mortality in infant, adult and older female rates than those in Asia and West models. In Asia model, the trends of mortality rate were quite lower than the Africa, but higher than West model. This is possibly due to the various factors playing the important roles for having higher number of mortality. For example, in most of Asian countries, there is relatively lower in a quality of fundamental human right services and a level of economical factor than those of western countries. In contrast, Africa countries are facing the low accessible of health and other human right services, accounted for a higher number of deaths, which are showed in our findings. Therefore, it can be implied that the trends of female mortality particularly depend on the country's level of socioeconomic development and status. Furthermore, most of the countries in the study have been reducing their mortality rates, obviously starting from the last three decades due to the substantial investments of the medical services and public health awareness and intervention programs. For African countries and the likes of Nepal and Pakistan, the female mortality rates were higher completely from infant to old ages in comparison to Asian and Western continents' females. This is possibly due to a long distance from health facilities, a high prevalence of HIV/AIDS, a nutritional deficiency, an act of violence and a quality of cares services. These findings are strongly supported by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division, where infant, maternal and old age mortalities of females were more in the African Sub-Sharan countries, due to the major epidemic of HIV and AIDS and other associated factors have been initiated for the age-specific patterns of mortality (Global Report 2013; Maternal mortality (2014). Nevertheless, in western countries, the trends of mortality have been incredible declining in all ages of female population. This is possibly due to the considerable improvement in public health policy and economic development system (Alkema et al 2015; world population prospects; maternal mortality 2015; Global report2013). However, in the Asian countries, the trends of female mortality is quite different from the African and western countries. In Asia, a higher number of female deaths are estimated in the old age of the female, while infant and child mortality trends are relatively low, also the trends of an adult age of female mortality are quite higher. This result is also relevant to the Regional World Health study, which shows that in Asia there is a low infant and child mortality and a higher mortality is noted in old age (Health in Asia and Pacific 2008).

On the other hand, this study reveals that the departure rate can reflect the immigration, when the trends of deviation of departure are negative. It can be observed that this mostly happens at younger ages of females in high and even some middle-income countries due to more people migrate into these countries, but this trend rather stops at the older ages of female population. However, in low and most middle-income countries, the deviation of departure is found to be low. It is reasonable to assume that these countries have a negligible immigration. If the immigration is negligible, then it is justified to indicate that the departure rate vastly depends on then morality rates and possibly emigration in some cases. This is supported by the findings by the World Bank, United Nation for Population found (UNFPA), and United Nation that western countries and some African and Asian countries including, Canada, Germany, Italy, Russia, Spain, Turkey, UK, USA, Australia, Algeria, Ethiopia, South-Africa, Bangladesh, Iran, Japan, Malaysia and South-Korea have hosted a large number of immigration as the legal and illegal entrances and refugee (International migration 2015; United Nations, 2013). Finally, it can be seen that this method provide the trends of female departure as the mortality rates for developing and developed countries from the departure rates. However, in high income countries where the immigration is fairly high, the approach in this study cannot explain all the departure rates as the mortality rates except for the old ages of females.

5. CONCLUSION

The trends of mortality improvement are noted in developed country, but the countries, those have been crippled for their basic fundamental needs, still face the higher female mortality rates in all ages, especially in most of the African and some Asian countries. Furthermore, the findings suggest that the trends of mortality can be varies across the countries due to the different investment of basic human fundamental services.

REFERENCES

Ansley J. Coale and Paul Demeny (1966). Regioml Model Life Table and Stable Population Princeton University Press. Demography Model, New Jersey, USA

Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A. B., Gemmill, A, & Say, L. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. The Lancet, 387(10017), 462-474.

Berin, B. N., Stolnitz, G. J., & Tenenbein, A. (1989). Mortality trends of males and females over the ages. Transactions of Society of Actuaries. 41, 1-19.

Currie, J., & Rossin-Slater, M. (2015). Early-life origins of life-cycle well-being: Research and policy implications. Journal of Policy Analysis and Management. 34(1), 208-242.

Gardner, J., & Oswald, A. (2004). How is mortality affected by money, marriage, and stress? Journal of Health Economics. 23(6), 1181-1207.

Global Health Risk 2009. World Health Organization: Mortality and burden of disease attributable to selected major risks, http://www.who.int/healthinfo/ accessed 30 December 2016.

Global report (2013). Geneva: Joint United Nations Programme on HIV/AIDS (UNAIDS); http://www.unaids.org/en/, accessed 30 December 2016.

International Migration 2015. Department of Economic and Social Affairs, United Nations, New York, http://www.un.org,en, accessed 22 December 2016.

Johri, M., Ridde, V., Heinmüller, R., & Haddad, S. (2014). Estimation of maternal and child mortality one year after user-fee elimination: an impact evaluation and modelling study in Burkina Faso. Bulletin of the World Health Organization. 92(10), 706-715.

Jonas, S., Grieco, G., Norman, R., Grumet, S., & Kedan, I. (2014). Mortality, ethnicity, and education in an occupational cohort. Ethnicity and Inequalities in Health and Social Care, 7(3), 137-145.

Kipping, R. R., Campbell, R. M., MacArthur, G. J., Gunnell, D. J., and Hickman, M. 2012. Multiple risk behaviour in adolescence. Journal of Public Health, 34 (suppl. 1), i1-i2.

Maternal mortality (2014). UNICEF Data: Monitoring the situation of children & women, http://www.mics.unicef.org. Accessed 20 December 2016.

Health in Asia & Pacific (2008). South-East Asia Region Western Pacific Region. Health Status Indicators, Demography statistics trends, Cost of Illness, Mortality, Communicable Disease Control, Non communicable Diseases, and Delivery of Health Care. www.wpro.who.int/health information evidence//Health in Asia Pacific.pdf. Accessed 29 December 2016.

South-East Asia. 11. World Health Organization, Regional Office for the Western Pacific. Health at a Glance (2013). Organization for Economic Co-operation & Development (OECD) Asia/Pacific Publishing.

Population Research Bureau (2007): Gender disparities in health and mortality http://www.prb.org/. Accessed 22 December 2016.

Stevens, G. A., Mathers, C. D., & Beard, J. R. (2013). Global mortality trends and patterns in older women. Bulletin of the World Health Organization. 91(9), pp. 630-639.

United Nations, (2013). World Population Ageing 2013. ST/ESA/SER.A/348. New York: United Nations, Department of Economic and Social Affairs, Population Division. pp. 1-3.

United Nations Children's Fund (2015). Maternal and Newborn Health, (https://www.unicef.org. Accessed 31 December 2016).

United States Census Bureau. International Data Base (IDB) Information Gateway Search form for locating data on Countries and Regions from 1950 to 2050. https://www.census.gov. Accessed 24 December 2016.

World population prospects revision (2015): United Nations, Department of Economic and Social Affairs, Population Division; http://.www.esa.un.org/ unpd/wpp/. Accessed 28 December 2016.

World Population Prospects; revision (2013). Population Division Department of Economic and Social Affairs United Nations New York, NY 10017, www.un.org. Accessed 30 November 2016

World Health Organization (2015). Global Health Observatory data on maternal health and maternal mortality, www.who.int/gho/en. Accessed 24 November 2016.

World's women trends & statistics (2015). United Nation Trend and Statistics, http://www.unstats.un.org. Accessed 24 December 2016.

Wrigley, E. A. (1997). English population history from family reconstitution 1580-1837 (No. 32). Cambridge University Press. Doi: 10.1017/CBO9780511660344. 601-610.

Statistical Modeling of Mortality and Morbidity among the Victims of Bhopal Gas Disaster

Akanksha S. Kashikar*
Savitribai Phule Pune University, Pune, India - akanksha.kashikar@gmail.com

Abstract

Bhopal Gas disaster has been considered as the world's worst industrial disaster. It occurred on the night of $2^{nd}/3^{rd}$ December 1984 at the Union Carbide India Limited pesticide plant in Bhopal, Madhya Pradesh. Over 500,000 people were exposed to methyl isocyanate gas and other chemicals. There have been many debates regarding the number of casualties that can be attributed to this gas leak. The numbers vary from 3787 to around 16000. The main reason behind this disparity in the estimates is the lack of well planned study. Indian Council of Medical Research undertook a properly designed study to give more reliable estimates of the number of victims. However, since the duration of the study was quite long, there was a significant attrition in the cohort at the intermediate stages. As a result, the accurate counts of the affected individuals are not available. Many statistical models, especially from survival analysis and actuarial statistics can be used for modeling mortality and morbidity, in presence of such censored observations. In this paper, we try to apply some of these parametric as well as non-parametric statistical modeling techniques to arrive at a more reliable estimate of the number of victims of Bhopal Gas Disaster.

Keywords: Censored data; Mortality rates; Survival analysis.

1. INTRODUCTION

The Bhopal disaster, also referred to as the Bhopal gas tragedy, was a gas leak incident in India, considered as the world's worst industrial disaster. On the night of 2nd/3rd December 1984, an industrial accident took place at the pesticide plant (owned by an American Multinational, the Union Carbide Corporation) in Bhopal in the state of Madhya Pradesh of India. Approximately, 40 tons of highly toxic liquid Methyl Isocyanate (MIC) stored in a tank suddenly escaped in a gaseous form into the atmosphere. This occurred around midnight when the ambient temperature was below 10° celsius, since December is a winter season in India. Most people were at home when the tank burst out emitting a thick cloud of deadly fumes into the night. The gases entered into their houses without warning. They woke up choked and coughing, unable to breathe; rubbing their eyes, unable to see because of acute swelling and burning. In a state of utter confusion and panic they rushed out of their houses, sometimes only to be engulfed by more gas waiting there. Approximately, half a million individuals out of a total population of over 0.8 million were reported/estimated to have been affected. A large number of deaths occurred in the exposed population. The official immediate death toll was 2,259. The government of Madhya Pradesh confirmed a total of 3,787 deaths related to the gas release. However, many unofficial estimates state the total number of casualties to be significantly higher than 10000. The main reason behind this disparity in the estimates is the lack of a well planned study. Indian Council of Medical Research undertook a properly designed study to give more reliable estimates of the number of victims. However, since the duration of the study was quite long, there was a significant attrition in the cohort at the intermediate stages. As a result, the accurate counts of the affected individuals are not available. In this paper, we use the data available in the ICMR report by Dwivedi et al (2010) and use statistical techniques to get estimates of the number of casualties over the period 1984-1994. Section 2 gives some details about the data that have been used, Section 3 discusses some preliminary analysis related to the mortality and morbidity rates among the victims, Section 4 discusses the estimation of number of deaths that can be attributed to this tragedy and Section 5 gives some concluding remarks. All the calculations have been performed using R language for statistical computing (R Core Team, 2016).

2. DATA DESCRIPTION

These details about the data have been obtained from Dwivedi et al (2010, p. 40-41). It is known that the eye and upper as well as lower respiratory passages are extremely sensitive to even as low as 2-4 ppm of MIC. This symptomatology was used as a dividing line between the exposed/affected and the unexposed/ unaffected pop- ulation, for the purpose of epidemiological studies. It is also a safe assumption that higher the concentration of the toxic gases inhaled, the more severe would be the mortality and greater the morbidity.

On the basis of clinical symptomatology after toxic gas inhalation, the entire population of Bhopal (832904) in 56 municipal wards was divided into the exposed and affected population (521262) in 36 municipal wards; and the unaffected population (311642) in 20 municipal wards. The exposed and affected population was further categorized on the basis of death rates during 3rd December to 6th December, 1984, into severely exposed/affected area (average death rate 21.98/1000 - population 32476); moderately affected/exposed area (average death rate 1.33/1000 - population 71917); and mildly affected/exposed area (average death rate 0.2/1000 - population 416869). The entire area with unaffected population of 311642, was treated as 'Control Area', for the sake of epidemiological studies, to draw comparisons with the affected areas. Expressed as percent of the total population of Bhopal, 3.9% were severely affected, 8.6% were moderately affected, 50.1% were mildly affected, while 37.4% were unaffected. From the affected/exposed areas the 'Registered Cohort' consisted of overall 80021 persons with 26382 from severely affected, 34964 from moderately affected and 18675 from mildly affected, for long term epidemiological studies. Another Cohort of 15931 persons was registered from the 'Control Area' for comparison. A questionnaire from a prestructured proforma was administered to the population of the Registered Cohort. A number of doctors and scientists at national level were actively involved in the planning and execution of epidemiological studies following the Disaster.

3. PRELIMINARY ANALYSIS

In the present section, we present some preliminary findings related to the mortality and the morbidity rates among the victims. At the beginning, we need to establish that the mortality rates differ across the four groups representing severe to no exposure. For this purpose, we use the data from Tables 2 and 3 in Dwivedi et al (2010). The data has been collected for ten years. The tables present annual mortality rates (per 1000) among males as well as females at each of the four exposure levels. We consider the *Year* and the *Exposure Level* as the factors affecting the mortality rates and *Male Mortality Rates* and *Female Mortality Rates* as the response variables. Hence, we have bivariate data being affected by two factors. To examine the significance of these factors we use the Permutational Multivariate Analysis of Variance (MANOVA) proposed by Anderson (2001). To allow for sequential addition of terms, we use the algorithm proposed by McArdle and Anderson (2001) for the permutational MANOVA. The calculations have been performed using R package vegan (Oksanen et al, 2017). The two factors *Year* and *Group* turn out to be significant with the p-values of 0.002 and 0.004 respectively. The *Year-Group* interaction is also marginally significant with a p-value of 0.055, indicating that the temporal change in the mortality pattern is not constant across the four exposure levels. The same can be seen in Figures 1(a) and 1(b). Table 42 in Dwivedi et al (2010) gives the data on abortions in pregnant women in each group in the cohort for the years 1984-1989. We compare the proportion of abortion in each group to the corresponding value from the control group. Using the usual estimate for variance of the estimate of the proportion and the normal approximation, we can construct a 95% confidence interval for each of these proportions. These estimates and the confidence intervals (dotted lines) are shown in Figure 2. It can be seen that in the severly exposed group and the moderately exposed group, the proportion of abortions is significantly higher than that in the control group for almost all the years under study.

To examine the temporal and group-wise variation in the morbidity pattern, we use the data available in Tables 22-29 of Dwivedi et al (2010). These data give percentage age-specific morbidity for each exposure group separately for males and females. The population is divided into five age groups, viz. 0 – 4, 5 – 14, 15 – 44, 45 – 64, 65+. The data have been collected through 17 visits spanning over a period of nine years from 1986 to 1994. The morbidity rates in the five age groups have been considered as response variables and the *Exposure Level, Visit Number* and *Gender* have been considered as factors affecting the morbidity.

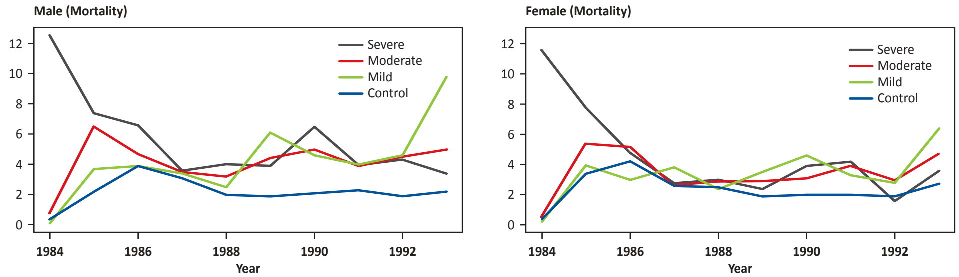


Figure 1. Annual Mortality Rates (per 1000)

The permutational MANOVA discussed above is appplied to the dataset. The results from this analysis are presented in Table 1. It can be seen that the factors *Exposure Level, Visit* and the interactions *Group-Visit* and *Gender-Visit* have turned out to be significant. This shows that the morbidity rates as well as the temporal pattern in the morbidity rates differ across the exposure levels.

Table 1. Permutational MANOVA for Age-Specific Morbidity Rates

Source of Variation	Degrees of Freedom	Sum of Squares	F statistic	P Value
Group	3	1.7069	9.1811	**0.0010**
Gender	1	0.1733	2.7971	**0.0590**
Visit	1	0.6689	10.7937	**0.0010**
Group: Gender	3	0.0628	0.3379	0.9600
Group: Visit	3	0.8117	4.3663	**0.0010**
Gender: Visit	1	0.3308	5.3378	**0.0050**
Group: Gender: Visit	3	0.2851	1.5333	0.1670
Residual	120	7.4365		

4. MORTALITY RATES

In this section, we again use the data from Tables 2 and 3 of Dwivedi et al (2010). We also use the sizes of the affected and unaffected populations and the estimated number of people at each exposure level reported in Section 2 of the current paper. Since the gender-wise distribution of the affected population sizes is not available, we combine the data for both the genders and use the combined data for further analysis.

As discussed in the introduction, since the study continued for almost ten years, there was a noticeable

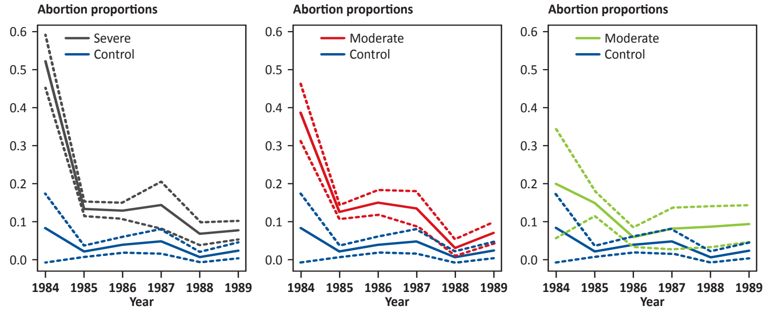


Figure 2. Estimates of Proportion of Abortion and Confidence Intervals for the Proportions

attrition in the initial cohort. Thus, we have censored data on some of the individuals. Hence, in this section, we use the non-parametric method of estimation of survival function in presence of censored data. We focus on the Actuarial method of estimation. The observations which are lost to follow-up are considered as censored observations. Further, we assume that, on the average, those individuals who are lost to follow-up during a particular interval were at risk for half of the interval. This assumption is used to modify the sample size to take care of the censored observations.

We consider each year (from 1984-1993) as one interval. Further, let

n_i = Number of people alive at the beginning of i^{th} year

d_i = Number of deaths during i^{th} year

l_i = Number of people lost to follow-up during i^{th} year

P_i = Conditional probability of surviving through i^{th} year given that the person was alive at the end of $(i-1)^{th}$ year

$Q_i = 1 - P_i$

Because of our assumption about the censored data, we get the modified sample size as

$$n'_i = n_i - \frac{l_i}{2}$$

and hence,

$$\hat{Q}_i = \frac{d_i}{n'_i} \text{ and } \hat{P}_i = 1 - \hat{Q}_i$$

Thus, the actuarial estimate of the survival function at the end of k^{th} interval is given by

$$\hat{S}_k = \prod_{i=1}^{k} \hat{P}_i$$

Further details about the method can be found in Deshpande and Purohit (2005, Section 5.3).

By using the above formula, we estimate the survival function for each exposure level separately. Further, to examine if the difference between the survival functions is significant or not, we construct the 95% confidence interval for each survival function using Greenwood's formula (Greenwood, 1926). The Greenwood's formula for estimate of variance of the survival function is given by,

$$\widehat{Var}(\hat{S}_k) = \hat{S}_k^2 \sum_{i=1}^{k} \frac{d_i}{n'_i(n'_i - d_i)}.$$

Thus, using this variance and the percentiles of the standard normal distribution, we get the confidence intervals for the estimates of the survival function. The estimates and the corresponding confidence bounds (dotted lines) for each exposure level are shown in Figure 3.

It can be seen that the survival functions of the severely exposed group and the moderately exposed group differ significantly from that of the control group for the entire duration. For the mildly exposed group, initially the survival function is similar to that of control group. However, after 1989, there is a significant decrease in the survival function for that group as well.

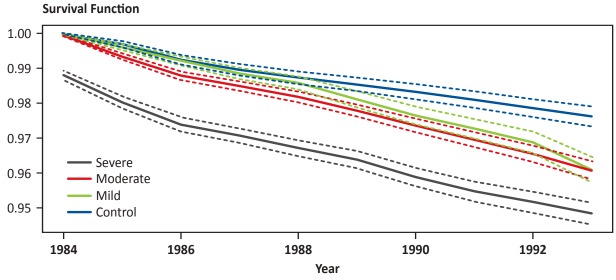


Figure 3. Estimates of Survival Functions and Confidence Intervals for the Survival Functions

These point estimates and the confidence bounds can be used to get the estimate of number of deaths during 1984-1993 due to Bhopal gas disaster. The minimum number of deaths that can be attributed to the tragedy is computed by considering the difference between the highest possible number of deaths (lower limit of 95% confidence interval for survival function) in control group and the lowest possible number of deaths in that particular exposure group. Similarly, the upper bound on the number of deaths is obtained by computing the difference between the lowest possible number of deaths (upper limit of 95% confidence interval for survival function) in control group and the highest possible number of deaths in that particular exposure group. These figures are reported in Table 2. This indicates that during the period 1984-1993, the number of deaths due to the Bhopal gas disaster is in the range (5118,11710), with the estimated count of deaths as 8489.

Table 2. Excess Deaths as Compared to Control Group

Exposure	Estimated Number of Deaths	Range for the Number of Deaths
Severe	906	(710, 1102)
Moderate	1144	(741, 1546)
Mild	6439	(3667, 9062)
Total	8489	(5118, 11710)

5. CONCLUDING REMARKS

From the above analyses, it is clear that even after ten years, the effect of the exposure is evident in the mortality and the morbidity patterns in the affected populations. The discussion in Section 3 shows that the morbidity and mortality pattern is significantly different across the different exposure levels. Further, the significant interaction between the exposure level and the time factor indicates that the temporal pattern is also affected by the exposure level. The discussion in Section 4 concludes that the official count of casualties is clearly an underestimate. Even during the first ten years, the estimate of the number of casualties is more than double the official count. If we consider the data on abortions, the estimate of number of deaths will move further in the upward direction. From the graph of survival function, it is indicated that even after ten years, the mortality patterns in the four groups have not become similar. This implies that if we consider the data for the next time periods, we may get even higher estimate of number of deaths that can be attributed to this tragedy. Hence, in order to bring justice to the victims, it is necessary to expand this study further.

REFERENCES

Anderson M. J. (2001). A new method for non-parametric multivariate analysis of variance, *Australian Ecology*, 26: 32-46.

Deshpande J. V. and Purohit S. G. (2005). *Life Time data: Statistical Models and Methods*, Series on Quality, Reliability and Engineering Statistics, Vol. 11, World Scientific Publishing, Singapore.

Dwivedi M. P., Prabhakar A. K. and Jain S. K. (2010). *Health Effects of the Toxic Gas Leak from the Union Carbide Methyl Isocyanate Plant in Bhopal*, Technical Report on Population Based Long Term Epidemiological Studies (1985 - 1994), Indian Council of Medical Research, New Delhi-110029.

Greenwood M. (1926). The natural duration of cancer, *Reports on Public Health and Medical Subjects* 33,

126. Her Majestys Stationery Office, London.

McArdle B. H. and Anderson M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82: 290-297.

Oksanen, J., Blanchet F. G., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin P. R., O'Hara R. B., Simpson G. L., Solymos P., Stevens M. H. H., Szoecs E. and Wagner H. (2017). *vegan: Community Ecology Package*. R package version 2.4-2, URL: https://CRAN.R-project.org/package=vegan

R Core Team (2016). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/.

Prevention of Demographic Disaster with Awareness of National Transfer Account

Lilia Endriana
Statistics Indonesia, Jakarta, Indonesia – lilia@bps.go.id
Yulia Mardani*)
Statistics Indonesia, Jakarta, Indonesia – yulia_tri@bps.go.id
Anisa Nuraini
Statistics Indonesia, Jakarta, Indonesia – anisa@bps.go.id

Abstract

Since 2015 secondary distribution of income and allocation of primary income account and use of disposable income have been constructed by BPS. By the year 2012 these accounts are in possible detail by certain age group such as productive age and dependency ratio age. Next year of 2017 these accounts are to become a national transfer account (NTA).

The country is entering a demographic transition. It is hoped that this transition will not turn into unwanted demographic disaster; Canning, et al. (2015). One suggested way to have awaited demographic dividend instead of unwanted demographic disaster is by securing a transfer. Evidence of transfer from productive age is reflected by NTA. Allocation by age group will also be based on Flow of Fund and Balance of Payment and Social Accounting Matrix (SAM).

On one hand BPS has finished a round of full sequence of accounts. On the other hand BPS has Household Expenditure Survey (SuSENas) to assist more detail accounts by productive age group. An early indicative signal of can be as evident as accelerated increase in value added.

Besides SuSENas other data sources are Input – Output table, and Household Saving and Investment Survey (SKTIR).

Data for 2010 indicates that life cycle deficit is approximately a thousand (NTA Training Workshops on Increasing Technical Capacity, Selangor Darul Ehsan, Malaysia, 23-27 May, 2016). It is hoped that productive age group of people be aware to start reducing this deficit. Hopefully the awareness may turn the deficit approaching into surplus toward the year 2035, in other words to have awaited demographic dividend instead of unwanted demographic disaster.

Keywords: Income account, NTA, Productive age

UNFPA stated that demographic dividend will reach the peak in 2028-2035. While Kompas has estimated the span will increase from 2020-2040. No one can estimate precisely the peak of the demographic dividend. This could be windows of opportunity of economic growth when all productive age have decent living to meet their consumption needs and prepare for future investment.

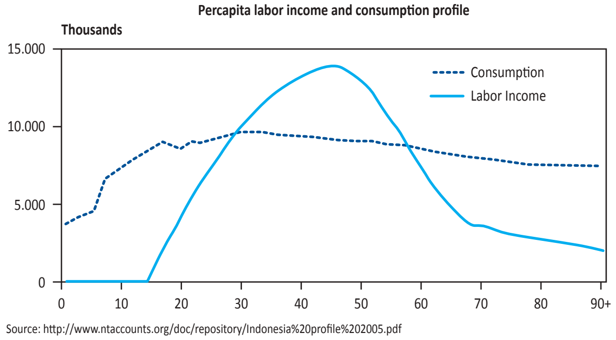
Fifth account is use of disposable income account covering household final consumption expenditure. Balancing item for fifth account is saving. Data for 2010 indicates that life cycle deficit is approximately a thousand (NTA Training Workshops on Increasing Technical Capacity, Selangor Darul Ehsan, Malaysia, 23-27 May, 2016). It is hoped that productive age group of people be aware to start reducing this deficit. Hopefully the awareness may turn the deficit approaching into surplus toward the year 2035. Reducing this deficit requires investment and dominant prerequisite of investment is saving. Productive age group of people to our thought is primary target of investment and saving awareness. During demographic transition size of productive age group of people is increasing to the maximum. This opportunity may not come for the second time. Later in demographic transition the size of productive age group of people is decreasing and may never increase. By the time size of productive age group of people is flat and may never increase the population is perhaps relying on return on investment. In other words time span of increasing size of productive age group of people is the only chance to increase investment and saving.

Fifth account is use of disposable income covering household final consumption expenditure. Balancing item for fifth account is saving. The other side of the fifth account is a chance to reduce household final consumption expenditure. Particular to this reduction is temporary restraint of luxury goods and service. Further it is desirable to reduce consumerism temporarily during demographic transition for the sake of investment and saving.

Secondary distribution of income account is the fourth account covers current transfers including social contribution. Balancing item for fourth account is national income. Temporary restraint of luxury goods and service can be used to prevent productivity decline. Further temporary restraint of luxury goods and service can be used to increase productivity. Social contribution is beneficial for certain productive age group of people having unsaturated productivity. The challenge is to identify productive age group of people having unsaturated productivity which is common in the country. Another challenge is to identify who is willing to participate in social contribution with government as facilitator. This far willingness participate in social contribution with government as facilitator has been around for some time in physical sport particularly football and badminton. Awareness of social contribution with government as facilitator has also been around for pupil and student. The author and co-authors would like to expand social contribution with government as facilitator to include employee productivity or per capita productivity. Certain exported goods and services are produced in other countries so multinational productivity comparison is possible. If possible priority is given to export-oriented per capita productivity improvement.

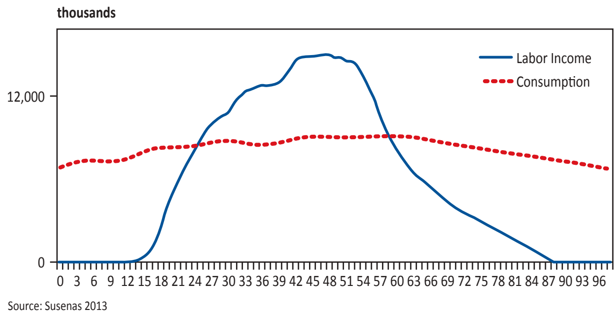
The old population are able to provide benefits when they remain productive in their old age. This is because of high quality of human capital investment on health and education and also asset that they have. Long preparation is needed to achieve the goal. The issue of nutrition quality, sanitation, and support of healthy environment and several priorities must be realized, as well as quality preparation of productive age population (http://www.kemenkeu.go.id/Artikel/penduduk-lansia-dan-bonus-demografi-kedua). Besides, the availability of services, facilities, and health insurance is absolutely necessary in order to maintain health. (Harian Kompas 3 Desember 2014-http://cpps.ugm.ac.id/penduduk-lanjut-usia-peluang-bonus-demografi-kedua/).

Third account namely allocation of primary income account covers taxes on production and imports and property income. Assuming that government for now on acknowledges the importance of asset and less importance on taxes on product. Next step can be the importance of asset based reallocations.



Graph 1. Economic Life Cycle 2005

Compare to 2005, study of 2013 data of economic life cycle by single age shows that entry market shifts to the left means that productive age earn more income. On the other side, old age earn less.



Graph 2. Economic Life Cycle 2013

Economic life cycle account measures how individual on each single age fulfilled their need with his job. Economic life cycle account consists of consumption, labor income, and the difference, which is next called as life cycle deficit/surplus. Components in the National Transfer Accounts are Life Cycle Deficit (LCD) and age reallocation (transfers and asset based reallocation - public and private). LCD shows whether a country have a surplus or deficit. To construct LCD, calculation of consumption using use of disposable income account. Calculation of labor income is using allocation of primary income account. Some adjustments are made to convert those accounts to NTA. Assumption is used to obtained labor income from allocation of primary income account. Labor income is 2/3 of gross mixed income. 1/3 goes to gross capital income.

CONCLUSION

Demographic transition leads to a demographic dividend that would only benefit at certain period. When productive age is not preparing themselves to enter the workforce, then there will be an unemployment explosion. Additionally, in the future it will pose a burden for productive age which number is less than old age.

The views here are that of author and co-authors and not necessarily reflect official views. The author and co-authors has received anonymous input.

REFERENCES

Canning, D., Raja, S., & Yazbeck, A. S. (Eds.). (2015). Africa's demographic transition: Dividend or disaster?. World Bank Publications

Harian Kompas 3 Desember 2014-http://cpps.ugm.ac.id/penduduk-lanjut-usia-peluang-bonus-demografi-kedua/http://www.ntaccounts.org. http://www.kemenkeu.go.id/Artikel/penduduk-lansia-dan-bonus-demografi-kedua.

Kinugasa, T. (2004). *Life expectancy, labor force, and saving* (Doctoral dissertation, UNIVERSITY OF HAWAI'I).

Lee, R. D., & Mason, A. (Eds.). (2011). *Population aging and the generational economy: A global perspective*. Edward Elgar Publishing.

NTA Training Workshops on Increasing Technical Capacity, Selangor Darul Ehsan, Malaysia, 23-27 May, 2016

Salim, Emil, SM. Adioetomo, EN. Arifin, Nizam, A. Pratama, (2015). Population Dynamics and Sustainable Development in Indonesia. UNFPA Indonesia. Jakarta

United Nations. National Transfer Accounts Manual: Measuring and Analysing the Generational Economy. (2013). United Nations Publications

This page intentionally left blank

IPS19:
Statistical Modelling, Risk Analysis and Risk Assessment

Port Value-At-Risk Estimation Through Generalized Means

M. Ivette Gomes*
CEAUL and DEIO, FCUL, Universidade de Lisboa, Portugal – ivette.gomes@fc.ul.pt
Fernanda Figueiredo
Faculdade de Economia, Universidade do Porto and CEAUL, Lisboa, Portugal – otilia@fep.up.pt
Lígia Henriques-Rodrigues
Universidade de São Paulo, IME, São Paulo, Brasil and CEAUL, Lisboa, Portugal – ligiahr@ime.usp.br

Abstract

In many areas of application, like environment, finance, insurance and statistical quality control, and on the basis of a sample of either independent, identically distributed or possibly weakly dependent and stationary random variables from an unknown model F, it is a common practice to estimate the *value-at-risk* (VaR) at a small level q, i.e. a high quantile of probability 1-q, i.e. a high enough value so that the chance of an exceedance of that value is equal to q, often smaller than $1/n$, where n is the size of the available sample. The semi-parametric estimation of these high quantiles depends heavily on a reliable estimation of the *extreme value index* (EVI), one of the primary parameters of extreme events, related to the heaviness of the right tail of F. It happens that most semi-parametric VaR-estimators available in the literature do not enjoy the adequate behaviour, i.e. they do not suffer the appropriate linear shift in the presence of linear transformations of the data, as does any theoretical quantile. For heavy tails, i.e. for a positive EVI, new VaR-estimators were introduced with such behaviour, the so-called PORT VaR-estimators, with PORT standing for *peaks over a random threshold*. Regarding EVI-estimation, new classes of PORT EVI-estimators, based on powerful generalizations of the Hill EVI-estimator were recently introduced. Now, also for heavy tails, we discuss the use of new classes of VaR-estimators with the aforementioned behaviour, using classes of EVI-estimators based on adequate generalized means related to the Hill EVI-estimators.

Keywords: Extreme value theory; Heavy tails; Monte-Carlo simulation; Semi-parametric estimation.

1. INTRODUCTION

In many areas of application, it is a common practice to estimate the *value at risk at a level q* (VaR_q), a value, high enough, so that the chance of an exceedance of that value is equal to q, small, often smaller than $1/n$, with n the available sample size. Such a sample is denoted by $\underline{X}_n=(X_1,\cdots,X_n)$ and their members are assumed to be either independent, identically distributed or stationary weakly dependent *random variables* (RVs) from a *cumulative distribution function* (CDF) F. Let us denote by $X_{1:n}\le\cdots\le X_{n:n}$ the associated ascending order statistics (OSs) and assume that there exist sequences of real constants $\{a_n>0\}$ and $\{b_n\in\mathbb{R}\}$ such that the maximum, linearly normalized, i.e. $(X_{n:n}-b_n)/a_n$ converges in distribution to a non-degenerate RV. Then (Gnedenko, 1943), the limit CDF is necessarily of the type of the general *extreme value* (EV) CDF, given by

$$G_\xi(x)=\begin{cases}\exp\left(-(1+\xi x)^{-1/\xi}\right),\ 1+\xi x>0, & \text{if } \xi\neq 0,\\ \exp\left(-\exp(-x)\right),\ x\in\mathbb{R}, & \text{if } \xi=0.\end{cases}\qquad(1.1)$$

The CDF F is said to belong to the max-domain of attraction of G_ξ and we write $F\in D_M(G_\xi)$. The parameter ξ is the *extreme value index* (EVI), the primary parameter of extreme events. This index measures the heaviness of the *right-tail function* $\bar{F}(x) := 1-F(x)$, and the heavier the right-tail, the larger ξ is. We here consider *heavy-tailed* models, i.e. *Pareto-type* underlying CDFs, with a positive EVI, working thus in $D_M^+ := D_M(G_{\xi>0})$. These *heavy-tailed* models are quite common in many areas of application, like biostatistics, computer science, finance, insurance, statistical quality control and telecommunications, among others. For *heavy-tailed* models, the classical EVI-estimators are the Hill (H) estimators (Hill, 1975), which are the average of the log-excesses, $V_{ik} := \ln X_{n-i+1:n}-\ln X_{n-k:n}$), $1\le i\le k<n$, i.e.

In this article, dealing only with heavy tails, we suggest ways to improve the performance of the existent VaR-estimators. High quantiles depend on the EVI, ξ in (1.1), and recently, new classes of reliable EVI-estimators based on adequate *generalized means* (GMs) have appeared in the literature, and will be introduced in Section 2. But these GM EVI-estimators are NOT location-invariant, contrarily to the PORT-GM EVI-estimators, which depend on an extra *tuning* parameter s, and where PORT stands for *peaks over a random threshold*. The use of the GM EVI-estimators enables us to introduce interesting classes of GM VaR-estimators, as can be seen in Section 3. But again, the associated GM VaR-estimators do not enjoy the adequate behaviour, i.e. they do not suffer the appropriate linear shift in the presence of linear transformations of the data, as does any theoretical quantile. We now discuss, also in Section 3, the use of new classes of VaR-estimators with the aforementioned behaviour, the so-called PORT-GM VaR-estimators, using classes of EVI-estimators based on adequate GMs related to the H EVI-estimators, in (1.2). In Section 4 we provide information on the possibly normal asymptotic behaviour of the aforementioned classes of EVI and VaR-estimators and on the adaptive choice of the most reliable VaR-estimate, jointly with some overall comments.

2. CLASSES OF GM EVI-ESTIMATORS

Hölder's *mean-of-order-p* (MO_p) EVI-*estimators.* First note that we can write

$$H(k) = \sum_{i=1}^{k} \ln \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k} = \ln \left(\prod_{i=1}^{k} \frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k}.$$

The H EVI-estimators are thus the logarithm of the *geometric mean* (or *mean-of-order*-0) of the statistics $U_{ik} := X_{n-i+1:n}/X_{n-k:n}$, $1 \le i \le k < n$. Brilhante *et al.* (2013), and almost simultaneously Paulauskas and iulis (2013), and Beran *et al.* (2014), considered as basic statistics, the MO_p of U_{ik}, $1 \le i \le k < n$, for $p \ge 0$. More generally, Gomes and Caeiro (2014), and also Caeiro *et al.* (2016), considered those same statistics for any $p \in \mathbb{R}$ and the associated class of EVI-estimators:

$$\mathrm{H}_p(k) = \mathrm{H}_p\left(k;\underline{X}_n\right) := \begin{cases} \left(1-\left(\frac{1}{k}\sum_{i=1}^{k} U_{ik}^{\,p}\right)^{-1}\right)/p, & \text{if } p < 1/\xi,\ p \neq 0, \\ \frac{1}{k}\sum_{i=1}^{k} \ln U_{ik} = \mathrm{H}(k), & \text{if } p = 0. \end{cases} \tag{2.1}$$

Lehmer's mean-of-order-p (L_p) EVI-*estimators.* Beyond the average, the p-moments of log-excesses, i.e. $M_{k,n}^{(p)} := \frac{1}{k}\sum_{i=1}^{k} \{ln\, X_{n-i+1:n} - ln\, X_{n-k:n}\}^p))$, $p \ge 1$ $[M_{k,n}^{(1)} = \mathrm{H(k)}]$, introduced in Dekkers *et al.* (1989), have also played a relevant role in the EVI-estimation, and can more generally be parameterized in $p \in \mathbb{R}\backslash\{0\}$. And another simple generalization of the average is Lehmer's mean-of-order-p: Given a set of positive numbers $\underline{a} = (a_1,\cdots,a_k)$ such a mean generalizes both the arithmetic mean (p=1) and the harmonic mean (p=0). Lehmer's mean-of-order-p is defined as

$$L_p\,(\underline{a}) := \sum_{i=1}^{k} a_i^{\,p} \Big/ \sum_{i=1}^{k} a_i^{\,p-1},\ p \in \mathbb{R}$$

The H EVI-estimators can thus be considered as the Lehmer's mean-of-order-p of the k log-excesses $\underline{\mathbf{V}} := (V_{ik}, 1 \le i \le k < n)$, for Following Penalva *et al.* (2016) (see also, Gomes *et al.*, 2016c), note now that $V_{ik} = \xi E_{k-i+1:k}(1 + o_p(1))$ with denoting a standard exponential RV and the $o_p(1)$-term uniform in i, $1 \le i \le k$. Since $\mathbb{E}(E^p)=\Gamma(p+1), \forall p>-1$, with $\Gamma(.)$ denoting the Gamma function, the law of large numbers enables us to say that, as $n \to \infty$, $\frac{1}{k}\sum_{i=1}^{k} V_{ik}^{\,p}$ converges in probability to $\Gamma(p+1)\,\xi^p$. Hence the reason for the class of L_p EVI-estimators, consistent for all $\xi \ge 0$, $p > 0$, and given by

$$\mathrm{L}_p\,(k) \equiv \mathrm{L}_p\left(k;\,\underline{X}_n\right) := \frac{L_p\,(\underline{V})}{p} = \frac{1}{p}\,\frac{\sum_{i=1}^{k} V_{ik}^{\,p}}{\sum_{i=1}^{k} V_{ik}^{\,p-1}} = \frac{M_{k,n}^{(p)}}{pM_{k,n}^{(p-1)}} \qquad [\mathrm{L}_1 \equiv \mathrm{H(k)}]. \tag{2.2}$$

Classes of PORT-GM EVI-*estimators*. The classes of GM EVI-estimators, in (2.1) and (2.2), depend on this *tuning parameter* $p\in\mathbb{R}$, are highly flexible, but, as often desired, they are not location-invariant, depending strongly on possible shifts in the model underlying the data, contrarily to what happens to the EVI, which is independent of shifts in the data. It is thus sensible to suggest the use of the classes of PORT-GM EVI-estimators. They are similar in spirit to the PORT-H EVI-estimators, studied in Araújo Santos *et al.* (2006), and further considered in Gomes *et al.* (2008). Classes of PORT estimators are based on a *sample of excesses* over a random threshold $X_{n_s:n}$, with $n_s := \lfloor ns \rfloor + 1$, $0 \le s < 1$,

$$\underline{X}_n^{(s)} := \left(X_{n:n} - X_{\lfloor ns\rfloor+1:n}, \cdots, X_{\lfloor ns\rfloor+2:n} - X_{\lfloor ns\rfloor+1:n}\right). \tag{2.3}$$

For $0 \le s < 1$ and $k < n - n_s$, the PORT-GM class of EVI-estimators has the same functional form of the GM class of EVI-estimators, but with $\underline{X}_n = (X_1, \cdots, X_n)$ replaced by the sample of excesses $X_n^{(s)}$, in (2.3). With GM denoting either H or L, respectively defined in (2.1) and (2.2), they are thus given by

$$\mathrm{GM}_p^{(s)}(k) \equiv \mathrm{GM}_p\left(k; \underline{X}_n^{(s)}\right).$$

These estimators are now invariant for both changes of scale and location in the data, and depend on the extra *tuning parameter* s, which provides a highly flexible class of EVI-estimators. Indeed, as shown in Gomes *et al.* (2016b), for the MOp EVI-estimation, these estimators may compare favorably with the PORT versions of the second-order *minimum-variance reduced-bias* (MVRB) EVI-estimators in Caeiro *et al.* (2005), provided that we adequately choose (p,s).

3. VAR-ESTIMATION

Just as we did before for the EVI-estimation, we are going to base inference on the largest k top OSs. Let us denote $U(t) := F^{\leftarrow}(1-1/t) = \inf\{x: F(x) \ge 1-1/t\}$. Using the notation $a(t) \sim b(t)$ if and only if $\lim_{t\to\infty} a(t)/b(t) = 1$, most heavy-tailed parents are such that $U(t) \sim Ct^{\xi}$ as $t\to\infty$. The, and since $\chi_{1-q} \equiv \mathrm{VaR}_q$ is such that $1-F(\mathrm{VaR}q) = q$,

$$\mathrm{VaR}_q = U(1/q) \sim C_q^{-\xi}, \text{ as } q\to 0,$$

and an obvious estimator of VaR_q is $\widehat{\mathrm{VaR}}_q = \hat{C}_q^{-\hat{\xi}}$, with $\hat{C}$ and $\hat{\xi}$ any consistent estimators of C and ξ, respectively. Denoting Y an RV from a standard Pareto model, with CDF $F_Y(y) = 1-1/y, y \ge 1$, $X_{n-k:n} \overset{d}{=} U(Y_{n-k:n}) \underset{\sim}{\overset{p}{}} CY^{\xi}_{n-k:n} \underset{\sim}{\overset{p}{}} C(n/k)^{\xi}$, as $n\to\infty$. An obvious estimator of C is thus $\hat{C} = (k/n)^{\hat{\xi}}\ X_{n-k:n}$, and the obvious VaR_q-estimator was introduced in Weissman (1978), being given by

$$Q_{\hat{\xi}}^{(q)}(k) := X_{n-k:n}\,(k/(nq))^{\hat{\xi}}. \tag{3.1}$$

For heavy-tailed models, the 'classical' EVI-estimators, usually the ones which are plugged in the previous formula, are the H EVI-estimators, already defined in (1.2), the average of the log-excesses. We thus get the so-called 'classical' VaR-estimators, based on the H EVI-estimators, with the obvious notation, $Q_H^{(q)}(k)$.

GM VaR-*estimation*. The high asymptotic bias of the H EVI-estimators, for small up to moderate k-values, has recently led researchers to consider the possibility of dealing with the bias term in an appropriate way, building new estimators, $\hat{\xi}_R(k)$ say, the so-called *second-order reduced-bias* (SORB) estimators (see Gomes and Guillou, 2015, for an overview of the topic). Caeiro *et al.* (2005), considered *corrected-Hill* (CH) MVRB EVI-estimators,

$$\mathrm{CH}_{\hat{\beta},\hat{\rho}}(k) := H(k)\left(1 - \hat{\beta}\,(n/k)^{\hat{\rho}}/(1-\hat{\rho})\right), \tag{3.2}$$

with $(\hat{\beta},\hat{\rho})$ adequate consistent estimator of (β,ρ), a vector of the second-order parameters, so that the asymptotic variance is kept at the same level of the variance of the H EVI-estimators. Gomes and Pestana (2007) considered then, as a possible alternative to the classical VaR-estimator, $Q_{\mathrm{H}}^{(q)}(k)$, the estimator in (3.1) based upon the EVI-estimators in (3.2), i.e. $Q_{\mathrm{CH}}^{(q)}(k)$, a reference for the VaR-estimation.

With GM_p denoting either H_p or L_p, respectively given in (2.1) and (2.2), we now think sensible to work with the new VaR_q-estimators $Q^{(q)}_{GM_p}(k)$, with the obvious functional form

$$Q^{(q)}_{GM_p}(k) := X_{n-k:n}\,(k/(nq))^{GM_p(k)}.$$

The Monte-Carlo simulations in Gomes *et al.* (2015) show the potentiality of the MO_p VaR_q semi-parametric estimators, being still under development the study of the L_p VaR_q-estimators.

PORT-GM VaR-*estimation*. Most of the semi-parametric VaR-estimators in the literature (see the functional equation in (3.1), Beirlant *et al.*, 2004, and de Haan and Ferreira, 2006), do not enjoy the adequate behaviour in the presence of linear transformations of the data, a behaviour related to the fact that for any high-quantile,

$$VaR_q(\lambda + \delta X) = \lambda + \delta VaR_q(X), \tag{3.3}$$

for any model X, real λ and positive δ. Recently, and for $\xi>0$, Araújo Santos *et al.* (2006) provided VaR-estimators with the linear property in (3.3), based on a *sample of excesses* over the random threshold $\underline{X}_n^{(s)}$, $n_s := \lfloor ns \rfloor + 1, 0 \le s < 1$, given in (2.3), being s possibly null only when the underlying parent has a finite left end point (see Gomes *et al.*, 2008b, for further details on this subject). They were named PORT VaR-estimators, and were based on the PORT-H, $H(k;\underline{X}_n^{(s)})$, $k < n - n_s$, with $H(k;\underline{X}_n)$ provided in (1.2).

Now, we further suggest for an adequate VaR-estimation, the use of the PORT-GM_p EVI-estimators,

$$GM_p(k;s) := GM_p(k;\underline{X}_n^{(s)}),\ k < n - n_s \qquad [GM=H \text{ and } GM=L], \tag{3.4}$$

with H_p, L_p and $\underline{X}_n^{(s)}$ respectively given in (2.1), (2.2) and (2.3). Such PORT-GM_p VaR-estimators are given by

$$\widehat{VaR}_q(k;p,s) := (X_{n-k:n} - X_{n_s:n})\,(k/(nq))^{GM_p(k,s)} + X_{n_s:n}.$$

4. ASYMPTOTIC BEHAVIOUR OF ESTIMATORS, ADAPTIVE CHOICE OF THE TUNING PARAMETERS AND OVERALL COMMENTS

First and second-order frameworks for heavy tails. Let R_a denote the class of regularly varying functions with an index of regular variation equal to $a \in \mathbb{R}$, i.e. measurable function $g(.)$ such that $\forall\ x>0, g(tx)/(g(t) \to x^a,)$ as $t\to\infty$. A model F is said to be heavy-tailed if $\xi > 0$, in (1.1), and we have the first-order condition,

$$F \in D_M^+ \Leftrightarrow U \in R_\xi \Leftrightarrow 1\text{-}F \in R_{-1/\xi}.$$

To obtain information on the non-degenerate *normal* behaviour of the estimators, it is usual to assume the following second-order condition,

$$\lim_{t\to\infty} (\ln U(tx) - \ln U(t) - \xi \ln x)/A(t) = \begin{cases} (x^\rho - 1)/\rho, & \text{if } \rho < 0, \\ \ln x, & \text{if } \rho = 0, \end{cases} \tag{4.1}$$

valid for all $x > 0$, where $\rho \le 0$ is a second-order parameter. Slightly more restrictively, and essentially for VaR-estimation, we shall assume to be working in Hall-Welsh class of models (Hall and Welsh, 1985), where $\exists \xi > 0, \rho < 0, C > 0$ and $\beta \neq 0$ such that

$$U(t) = Ct^\xi\,(1 + \xi\beta\, t^\rho/\rho + o(t^\rho)), \text{ as } t\to\infty \tag{4.2}$$

Asymptotic behaviour of EVI-*estimators*. To have consistency of the aforementioned EVI-estimators in all D_M^+, we need to work with *intermediate* values of k, i.e. a sequence of positive integers $k=k_n, 1\le k<n$, such that $k=k_n\to\infty$ and $k_n=o(n)$, as $n\to\infty$. Under the aforementioned second-order framework in (4.1), the asymptotic behaviour of the H EVI-estimator was derived in de Haan and Peng (1998). More generally (Brilhante *et al.*, 2013, and Gomes and Caeiro, 2014, for the H_p EVI-estimators, and Penalva *et al.*, 2016, for the L_p EVI-estimators), and again using the notation GM_p for both H_p and L_p, the asymptotic distributional representation

$$\mathrm{GM}_p\,(k) \stackrel{d}{=} \xi + \sigma_{\mathrm{GM}_p}(\xi)\, Z_k^{(\mathrm{GM}_p)}/\sqrt{k} + b_{\mathrm{GM}_p}(\xi,\rho)\, A(n/k) + o_p\,(A(n/k))$$

holds with $Z_k^{(\mathrm{GM}_p)}$ asymptotically standard normal RVs. Then, when $\sqrt{k}\, A(n/k)\to\lambda$, finite, as $n\to\infty$, $\sqrt{k}(\mathrm{GM}_p\,(k)-\xi)$ converges in distribution to a $N(\lambda b_{\mathrm{GM}_p}, \sigma^2_{\mathrm{GM}_p})$.

Remark 1. *At optimal levels, in the sense of minimal root mean squared error* (RMSE), *the optimal* MO_p (OMO_p) *class,* $\mathrm{H}^*\,(k)=\mathrm{H}_{p_M}\,(k)$, *outperforms the* H EVI-*estimator in the whole* (ξ,ρ)-*plane.*

And, again at optimal levels, the optimal Lehmer EVI-*estimator, say* L*, *beats on its turn* H*, *also in the whole* (ξ,ρ)-*plane.*

To derive the asymptotic properties of the PORT-GM EVI-estimators, it is worth noting that since $X_{[ns]+1:n} - U(1/(1-s)) = O_p\,(1/\sqrt{n})$), the EVI-estimator $\mathrm{GM}p\ (k,s)$, in (3.4), has the same asymptotic behaviour of $\widetilde{\mathrm{GM}}_p\,(k,s)$, defined as $\mathrm{GM}_p\,(k,s)$, but with $X_{n-i+1:n}$ replaced everywhere by $X_{n-i+1:n}-U(1/(1-s)), 1\le i\le n$.

The PORT-MO_p EVI-estimators were studied in Gomes *et al.* (2016b). Similar results, still under development, are expected for the PORT-L_p EVI-estimators. The PORT methodology leads to no change in the asymptotic variance. There is only a change in the asymptotic bias, no longer ruled by $A(t)$, but ruled by

$$B(t) = \begin{cases} \xi\,\chi_s/U_0\,(t), & \text{if } \xi+\rho_0<0 \wedge \chi_s\neq 0,\\ A_0\,(t)+\xi\,\chi_s/U_0\,(t), & \text{if } \xi+\rho_0=0 \wedge \chi_s\neq 0,\\ A_0\,(t), & \text{otherwise,}\end{cases}$$

where $\chi_s=F^{\leftarrow}(s)$, and (A_0,U_0) are the functions (A,U) associated with a shift $s=0$.

Asymptotic behaviour of the VaR-*estimators.* For all the aforementioned classes of VaR-estimators, generally denoted $Q_\xi^{(q)}\,(k)$, and with $r_n=k/(nq)$, we can write

$$Q_\xi^{(q)}\,(k)-\mathrm{VaR}_q \stackrel{d}{=} \ln r_n\, \mathrm{VaR}_q\,(\hat{\xi}-\xi)(1+o_p\,(1)),$$

whenever working in Hall-Welsh class of models, in (4.2). For intermediate k, and whenever $q=q_n\to 0, \ln(nq_n)=o(\sqrt{k})$, and $nq_n=o(\sqrt{k})$, a similar normal behaviour appears for the EVI and associated VaR-estimators, but with a rate of convergence which is no longer $1/\sqrt{k}$ but $1/(\sqrt{k}\, \ln r_n\, \mathrm{VaR}_q)$. For a PORT VaR-estimation, see Henriques-Rodrigues and Gomes (2009) and Figueiredo *et al.* (2016).

Adaptive choice. The adaptive choice of the tuning parameters can be done through heuristic sample-path stability algorithms, like the ones in Gomes *et al.* (2013) and Neves *et al.* (2015). Alternatively, it is also sensible to use a bootstrap algorithm of the type of the ones in Gomes *et al.* (2011; 2012), Brilhante *et al.* (2013), Caeiro and Gomes (2015) and Gomes *et al.* (2016a), where R-scripts are provided.

Overall comments. For all k, there is a clear reduction in RMSE, as well as in bias, with the attainment of estimates closer to the target value ξ. At optimal levels, even the PORT-H* beats the MVRB estimators. Indeed, the PORT-H_p, considered as a function of p, can even beat the PORT-MVRB EVI-estimators. The patterns of the estimates are always of the same type, in the sense that, for all k, the MVRB clearly beat the Hill, the H* moderately beat the MVRB, regarding minimal MSE, and adequate MO_p and PORT-MO_p strongly beat the MVRB EVI-estimators.

For recent overviews on statistics of univariate extremes, see Beirlant *et al.* (2012), Scarrot and McDonald (2012) and Gomes and Guillou (2015).

Acknowledgement. Research partially supported by National Funds through **FCT** — Fundação para a Ciência e a Tecnologia, project UID/MAT/00006/2013 (CEAUL), and by COST Action IC1408 — CroNos.

REFERENCES

Araújo Santos, P., Fraga Alves, M.I., & Gomes, M.I. (2006). Peaks over random threshold methodology for tail index and quantile estimation. *Revstat—Statistical J.* **4**:3, 227-247.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, England.

Beirlant, J., Caeiro, F., & Gomes, M.I. (2012). An overview and open research topics in statistics of univariate extremes. *Revstat—Statistical J.* **10**:1, 1-31.

Beran, J., Schell, D., & Stehlik, M. (2014). The harmonic moment tail index estimator: asymptotic distribution and robustness. *Ann. Inst. Statist. Math*. **66**, 193-220.

Brilhante, M.F., Gomes, M.I., & Pestana, D. (2013). A simple generalization of the Hill estimator. *Computational Statistics and Data Analysis* **57**:1, 518-535.

Caeiro, F., & Gomes, M.I. (2015). Threshold selection in extreme value analysis. In Dipak Dey and Jun Yan (eds.), *Extreme Value Modeling and Risk Analysis: Methods and Applications*, Chapman-Hall/CRC, Chapter 4, 69-87.

Caeiro, F., Gomes, M.I., & Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator. *Revstat—Statistical J.* **3**:2, 113-136.

Caeiro, F., Gomes, M.I., Beirlant, J., & Wet, T. de (2016). Mean-of-order p reduced-bias extreme value index estimation under a third-order framework. *Extremes* **19**:4, 561-589.

Dekkers, A., Einmahl, J., & Haan, L. de (1989). A moment estimator for the index of an extreme-value distribution. *Ann. of Statist.* **17**, 1833-1855.

Figueiredo, F., Gomes, M.I., & Henriques-Rodrigues, L. (2016). Value-at-risk estimation and the PORT mean-of-order-p methodology. *Revstat—Statistical J.* Accepted.

Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics* **44**:6, 423-453.

Gomes, M.I., & Caeiro, F. (2014). Efficiency of partially reduced-bias mean-of-order-p versus minimum-variance reduced-bias extreme value index estimation. In M. Gilli *et al.* (eds.). *Proceedings of* COMPSTAT 2014, The International Statistical Institute/International Association for Statistical Computing, 289-298.

Gomes, M.I., & Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: A Review. *International Statistical Review* **83**:2, 263-292.

Gomes, M.I., & Pestana, D. (2007). A sturdy reduced bias extreme quantile (VaR) estimator. *J. American Statistical Association* **102**:477, 280-292.

Gomes, M.I., Fraga Alves, M.I., & Araújo Santos, P. (2008). PORT Hill and moment estimators for heavy-tailed models. *Commun. in Statist.: Simul. and Comput.* **37**, 1281-1306.

Gomes, M.I., Mendonça, S., & Pestana, D. (2011). Adaptive reduced-bias tail index and VaR estimation via the bootstrap methodology. *Comm. in Statistics: Th. and Meth*. **40**:16, 2946-2968.

Gomes, M.I., Figueiredo, F., & Neves, M.M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action. *Extremes* **15**, 463-489.

Gomes, M.I., Henriques-Rodrigues, L., Fraga Alves, M.I., & Manjunath, B.G. (2013). Adaptive PORT-MVRB estimation: an empirical comparison of two heuristic algorithms. *J. Statist. Comput. and Simul*. **83**:6, 1129-1144.

Gomes, M.I., Brilhante, F., & Pestana, D. (2015). A mean-of-order-p class of value-at-risk estimators. In Kitsos, C. *et al.* (eds.), *Theory and Practice of Risk Assessment*, Springer Proceedings in Mathematics and Statistics 136, Springer International Publishing, Switzerland, 305-320.

Gomes, M.I., Caeiro, F., Henriques-Rodrigues, L., & Manjunath, B.G. (2016a). Bootstrap methods in statistics of extremes. In Longin, F. (ed.), *Handbook of Extreme Value Theory and Its Applications to Finance and Insurance*. Handbook Series in Financial Engineering and Econometrics (Ruey Tsay Adv.Ed.). John Wiley and Sons, Chapter 6, 117-138.

Gomes, M.I., Henriques-Rodrigues, L., & Manjunath, B.G. (2016b). Mean-of-order-p location-invariant extreme value index estimation. *Revstat—Statistical J.* **14**: 3, 273-296.

Gomes, M.I., Penalva, H., Caeiro, F., & Neves, M.M. (2016c). Non-reduced versus reduced-bias estimators of the extreme value index: efficiency and robustness. In A. Colubi *et al.* (eds.) COMPSTAT 2016—22*nd International Conference on Computational Statistics*, The International Statistical Institute/International Association for Statistical Computing, 279-290.

Haan, L. de, & Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer.

Haan, L. de, & Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* **52**, 60-70.

Hall, P., & Welsh, A.H. (1985). Adaptative estimates of parameters of regular variation. *Annals of Statistics* **13**, 331-341.

Henriques-Rodrigues, L., & Gomes, M.I. (2009). High quantile estimation and the PORT methodology. *Revstat—Statistical J.* **7**:3, 245-264.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**:5, 1163-1174.

Neves, M.M., Gomes, M.I., Figueiredo, F., & Prata-Gomes, D. (2015). Modeling extreme events: sample fraction adaptive choice in parameter estimation. *J. Statistical Theory and Practice* **9**:1, 184-199.

Paulauskas, V., & Vaičiulis, M. (2013). On the improvement of Hill and some others estimators. *Lithuanian Mathematical J.* **53**, 336-355.

Penalva, H., Caeiro, F., Gomes, M.I., & Neves, M.M. (2016). *An Efficient Naive Generalisation of the Hill Estimator: Discrepancy between Asymptotic and Finite Sample Behaviour*. Notas e Comunicações CEAUL 02/2016.

Scarrot, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Revstat—Statistical J.* **10**:1, 33-60.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *J. American Statistical Association* **73**, 812-815.

Models and Applications of BIB Designs in Life and Health Sciences

Teresa A. Oliveira[1,2]
[1]Universidade Aberta, Palácio Ceia, Rua da Escola Politécnica, Lisboa, Portugal
2Center of Statistics and Applications, University of Lisbon
– Teresa.Oliveira@uab.pt
Amilcar Oliveira
Universidade Aberta and CEAUL, Lisbon, Portugal – amilcar.oliveira@uab.pt
Carla Francisco[1,3]
3Faculté de Foresterie, de Géographie et de Géomatique, Québec, Canada
– carla.francisco.1@ulaval.ca

Abstract

Balanced Incomplete Block Designs (BIBD) play a key role on Experimental Design and many applications are emerging now-a-days in areas connected with Health Sciences and Medicine. In this work we will review some of these applications and explore some special models and structures of BIBD allowing block repetition (BIBDR) which reveal to be a very important tool on treatments comparison. The main advantage of using BIBDR is that these special designs keep the optimality properties of a BIBD and provide some block contrasts with minimum variance. Examples of different structures for a same particular BIBDR will be presented, as well as illustrations of such designs usefulness on solving real problems in Life and Health Sciences. The advantages of using software R for computations in problems connected with BIBD and BIBDR will also be discussed.

Keyword : Balanced Incomplete Block Designs, Health Science.

The Presence of Distortions in the Extended Skew-normal Distribution

Antonio Seijas-Macias*
Universidade da Coruña and Universidad Nacional de Educación a Distancia (UNED), A Coruña, Spain – antonio.smacias@udc.gal
Amílcar Oliveira
Universidade Aberta and Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal. – Amilcar.Oliveira@uab.pt
Teresa Oliveira
Universidade Aberta and Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal – Teresa.Oliveira@uab.pt

Abstract

In the last years, a very interesting topic has arisen and became the research focus not only for many mathematicians and statisticians, as well as for all those interested in modeling issues: The Skew normal distributions' family that represents a generalization of normal distribution. The first generalization was developed by Azzalini in 1985, which produces the skew-normal distribution, and introduces the existence of skewness into the normal distribution. Later on, the extended skew-normal distribution is defined as a generalization of skew-normal distribution. These distributions are potentially useful for the data that presenting high values of skewness and kurtosis. Applications of this type of distributions are very common in model of economic data, especially when asymmetric models are underlying the data. Definition of this type of distribution is based in four parameters: location, scale, shape and truncation. In this paper, we analyze the evolution of skewness and kurtosis of extended skew-normal distribution as a function of two parameters (shape and truncation). We focus in the value of kurtosis and skewness and the existence of a range of values where tiny modification of the parameters produces large oscillations in the values. The analysis shows that skewness and kurtosis present an instability development for greater values of truncation. Moreover, some values of kurtosis could be erroneous. Packages implemented in software R confirm the existence of a range where value of kurtosis presents a random evolution.

Keywords: Skewness; Kurtosis; Skew normal; Computational Cumulants.

JEL Classification: C15, C63

1. INTRODUCTION

The skew-normal probability distribution (Azzalini, 1985) was introduced by Azzalini in 1985 which including the normal distribution as a special case. The skew-normal is a generalization of the normal distribution where the more important characteristic is the presence of different level of skewness. In the motivation of this new approach to the normal distribution there are two basic: first motivation, lies in the essence of the mechanism which starts with a continuous symmetric density function that then is modified to generate a variety of alternative forms (Azzalini & Capitanio, 2014, page 1). The second one is the need to approximate the real data with the theoretical distributions using the fewer unrealistic assumptions for the model.

In the seminal paper (Azzalini, 1985), author considers a generalization of the skew-normal distribution that is named as extended skew-normal distribution, which skew-normal one is an instance. Generalization to the bivariate case and multivariate one are analyzed in Arnold et al. (1993) and Capitanio et al. (2003).

The skew-normal is derived from the standard normal density function ϕ and the standard normal distribution function Φ. In the univariate case the skew-normal a random variable X is distributed as a skew-normal distribution with density function

$$f_X(x)=\frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right).$$

where we use three parameters: location $\xi\in\mathbb{R}$, scale $\omega\in\mathbb{R}^+$ and shape $\alpha\in\mathbb{R}$. When $\xi=0$ and $\omega=1$ then we have the normalized skew-normal distribution, and the only free parameter α determines the skewness level. As a special case for $\alpha=0$ then we have the standard normal distribution N (0,1) with skewness equals zero and kurtosis equals 3. Obviously, the parameter α could control the skewness of the distribution. Unfortunately, kurtosis is not considered.

The introduction of a new parameter $\tau\in\mathbb{R}$, named truncation parameter produces the extended skew-normal distribution. For the univariate case, the density function for the Extended Skew-Normal distribution (ESN) is

$$f_X(x)=\frac{1}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\frac{\Phi\left(\tau\sqrt{\alpha^2+1}+\alpha\frac{x-\xi}{\omega}\right)}{\Phi(\tau)}$$

For the case, which $\alpha=0$ then we have the skew-normal distribution and $\omega=0$ the normal one. In Canale (2011) some properties of the density function are presented and the name of truncation for the new parameter τ is explained. The effect of the new parameter τ has consequences on the skewness and the kurtosis of the distribution of the random variable. The effect of the τ is not independent of the parameter α. For the smaller value of α variations on τ it produces less effect on the shape of the distribution. Basically, the two parameters determine values of skewness and kurtosis of the distribution.

During last years, the use of Skew-normal distribution and ESN is increasing, to represent different types of data with skewness and where the test of normality (Shapiro Test) is not satisfied (Canale, 2011). Some authors (Figueiredo & Gomes, 2013; Chenglong et al. 2017) consider different types of hypothesis contrast using the Skew-normal and ESN distributions.

For R software (2015) the package **sn**, developed by Azzalini, provides facilities to define and manipulate probability distributions of the skew-normal family and some related ones. The first version of the package was written in 1997, and in 2014 the version 1.0-0 was uploaded to CRAN. In this paper, we have used the implemented functions available in the version 1.4-0 of the package.

2. SKEWNESS AND KURTOSIS OF ESN

The moment generating function of an ESN distribution is computed from the expression $Y=\xi+\omega Z$ where $Z\sim SN(0,1,\alpha,\tau)$, as in

$$M(t)=\exp\left(\xi t+\frac{1}{2}\omega^2 t^2\right)\frac{\Phi(\tau+\delta\omega t)}{\Phi(\tau)}$$

where $\delta=\dfrac{\alpha}{\sqrt{1+\alpha^2}}$

Expressions for the statistics (mean, variance, skewness and kurtosis) could be calculated as a function of the derivatives of the moment-generating function. Skewness and kurtosis are independent of the values of the two first parameters of extended skew-normal distribution; only α and τ are relevant. The value of mean and variance is a function of the four parameters.

Another approach is using the cumulant generating function given by:

$$\kappa(t)=\log M(t)=\xi t+\frac{1}{2}\omega^2 t^2+\zeta_0(\tau+\delta\omega t)-\zeta_0(\tau)$$

where $\zeta_0(x)=\log 2\Phi(x)$ and $\zeta_r(x)=\frac{d^r\zeta_0(x)}{dx^r}$, with $r=1,2,\ldots$. The derivatives of $\kappa(t)$ are κ_i, the cumulant of order i. Thus, the derivatives of ζ are:

1. $\zeta_1(x)=\frac{d\zeta_0(x)}{dx}=\exp(\log(\phi(x))-\log(\Phi(x)))$
2. $\zeta_2(x)=-\zeta_1(x)(\zeta_1(x)+x)$
3. $\zeta_3(x)=-\zeta_2(x)(\zeta_1(x)+x)-2\zeta_2(x)(1+\zeta_2(x))=\zeta_1(x)(\zeta_1(x)+x)^2-\zeta_1(x)(1-\zeta_1(x)(\zeta_1(x)+x))$
4. $\zeta_4(x)=-\zeta_3(x)(2\zeta_1(x)+x)-2\zeta_2 x(1+\zeta_2(x))=-(\zeta_1(x)(\zeta_1(x)+x)^2-\zeta_1(x)(1-\zeta_1(x)(\zeta_1(x)+x)))^*(x+2\zeta_1(x))+2\zeta_1(x)(\zeta_1(x)+x)(1-\zeta_1(x)(\zeta_1(x)+x))$

The values of mean, variance, skewness and kurtosis are calculated from the cumulants:

1. Cumulant of first order is the mean.
2. Cumulant of second order is the variance.
3. Skewness is calculated as $\frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\zeta_3(\tau)\delta^3}{(1+\zeta_2(\tau)\delta^2)^{(3/2)}}$.
4. Kurtosis is calculated[1] as $\frac{\kappa_4}{(\kappa_2^2)} + 3 = \frac{\zeta_4(\tau)\delta^4}{(1+\zeta_2(\tau)\delta^2)^2} + 3$.

Value of cumulants is obtained by function **sn.cumulants** (ψ,ω,α,τ,n) where n is the number of cumulants (default=4, max=6). This procedure is part of the package **sn** (Azzalini, 2004).

The value of mean and variance are directly, but for calculation of the skewness and kurtosis could be obtained using the cumulants of order 3 and 4, respectively and the cumulant of order 2. For example: Considering the extended skew normal with parameters ξ=0,ω=1,α=-50,τ=-6, skewness is -1.825792 and kurtosis is 4.874607. Mean and variance of ESN(0,1,-50,-6) is given by the two first elements of the result of procedure sn.cumulants, then, mean equals -6.157251 and variance is 0.0243778.

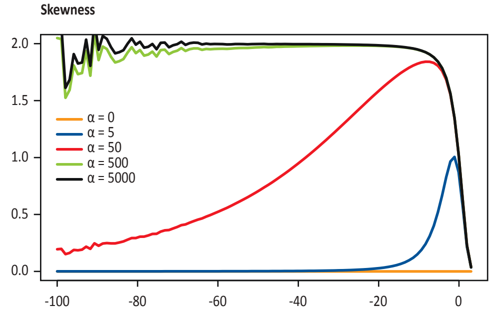


Figure 1. Evolution of Skewness ESN Distribution for Several Values of α

The sign of α determines the sign of skewness, when α<0 then skewness is negative and α>0 is associated to positive skewness. For α=0 then skewness is 0. The value of τ determines the value of skewness that tends to zero when τ→±∞, although is more sensible for the positive branch. Figure 1 shows the evolution of skewness for several values of α as a function of τ. Range of skewness is in (-2,2).

For positive values of τ skewness tend to zero when τ is increasing, for smaller values; but in the negative branch skewness presents a wider range of variation, although, when τ decrease, skewness decreases to zero (for positive values of α) and skewness increases to zero (for negative values of α).

In Canale (2011), the maximum value of skewness is cited as 1.995, but the observation of the figure 1 shows, for values of τ negative and very large the presence of oscillations that produce values of kurtosis greater than 2. We might consider that situations as errors for the precision machine limits of computers. As consequence, the level of skewness is around 2 and no decreasing movement to zero is observed. The approach for lower values of τ reduce the range for skewness to interval (-2,2), but using the **sn.cumulants** function some strange values for skewness are possible, when α is greater (more than 1000) and τ<-500, in this case the range of values of skewness is increasing when τ is decreasing.

Kurtosis presents the same strange oscillations for greater values of α and when τ is large and negative (see figure 2). But now, the oscillations appear for lower values (t≤-40).

1 Some authors consider the excess kurtosis that is the quotient between cumulant of order 4 and the square of cumulant of order 2. Kurtosis could not be negative, but excess kurtosis reaches negative values.

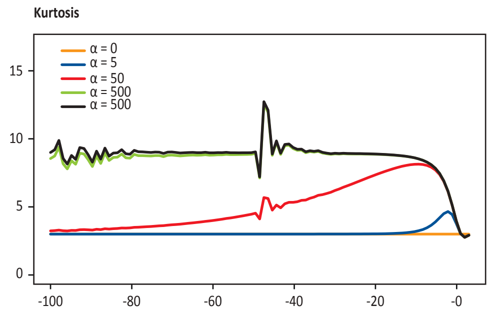


Figure 2. Evolution of Kurtosis ESN Distribution for Several Values of α.

The range of kurtosis of the ESN distribution could be (2,10). The large value of α the large value of kurtosis (although a limit could be exists rounded 10). And when $|\tau|$ is large, kurtosis tends to zero. This tendency is more sensitive in the positive side. For very large values of α (>50) kurtosis tends to 10 but oscillations produce values very large >15.

In Azzalini & Capitanio (2014), Fig. 2.5 shows graphically the range of skewness and kurtosis (excess kurtosis) with the Extended Skew-normal distribution with $\alpha>0$ and $\tau \in(-10,2)$. The range of both statistics is wider than in the Skew-normal distribution, but the whole range is limited. However, there is no reference to the oscillations of skewness and kurtosis that appears when values for $\tau<-50$ and $\alpha>500$ are considered.

3. OSCILLATIONS IN SKEWNESS AND KURTOSIS OF THE ESN DISTRIBUTION.

A graphical analysis of the evolution of the values of both statistics for $\alpha>500$ and $\tau<-50$ showed the presence of instabilities and oscillations on the range of the values. In this section, we analyses the reasons for that will be studied.

Shape of the density function (PDF) of the ESN distribution might be affected by the oscillations into skewness and kurtosis. But we can prove, using an example, that the shape of function doesn't reflect the variations of those statistics, see Figure 3.

The evolution of the shape of the density function is very smooth. In Figure 3, it can be observed the pdf of several ESN distributions for $\xi=-1,\omega=2.25,\alpha=5000$ and $\tau=[-110,-105]$. The same effect is observed for large values of the parameter τ (Figures 4), in this case the shape of the distribution is the same for values $\tau=\{-360,-361,-362,-363,-364,-365\}$. One can conclude that oscillations in skewness and kurtosis are not reflected at the shape of the distribution which presents a regular evolution without deep variations.

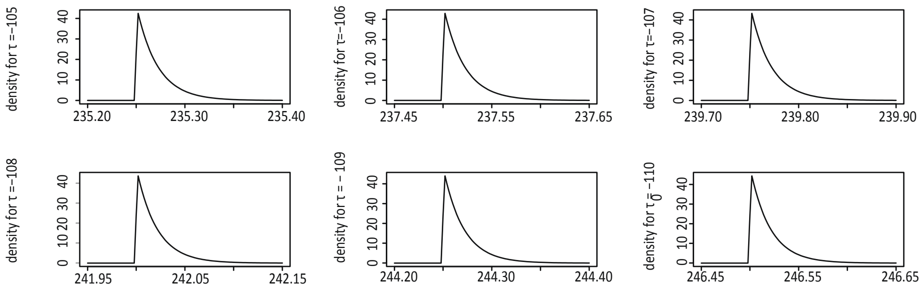


Figure 3. PDF of ESN(-1,2.25,5000,τ)τ∈[-110,-105]

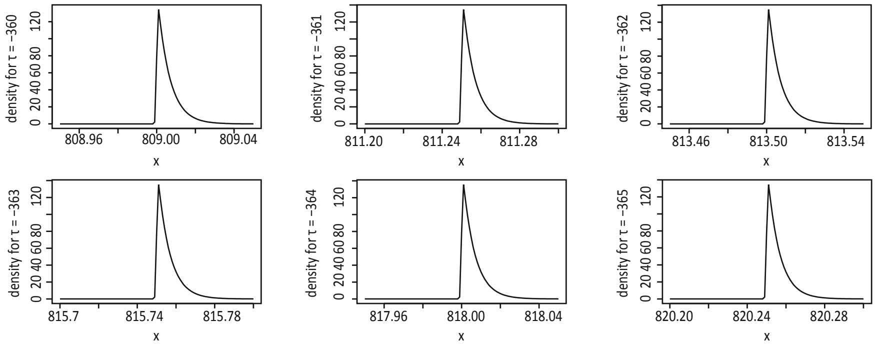


Figure 4. PDF of ESN(-1,2.25,5000,τ) τ∈[-365,-360]

The statistics are in Table 1. Skewness and kurtosis are increasing as τ is increasing, but the variation is decreasing, in a concave manner; however, when the value of τ is large and negative, the evolution of both statistics is strange and it presents oscillations.

When considering the more negative values for τ and large values for α, the skewness and kurtosis has strange oscillations for small variations. Theses variations could not be reflected at the shape of the pdf of the Extended Skew-normal distribution.

Table 1. Statistics for ESN(-1,2.25,5000, τ)

τ	Mean	Variance	Skewness	Kurtosis
-1	2.43155	1.007932	1.316228	4.997356
-2	4.33973	0.578538	1.536431	6.018594
-3	6.38697	0.357206	1.679095	6.80238
-4	8.50761	0.236281	1.770906	7.364103
-5	10.66963	0.165525	1.831079	7.759171
-105	235.2714	0.000459	1.998394	6.118636
-106	237.5212	0.000455	1.998172	8.395796
-107	239.7710	0.000442	1.998079	9.265955
-108	242.0208	0.000434	1.99806	9.308503
-109	244.2706	0.000426	1.998015	9.650362
-110	246.5204	0.000418	1.997757	12.57062
-360	809.0062	0.000039	1.779087	26562.73
-361	811.2562	0.000039	1.759785	29207.98
-362	813.5062	0.000038	2.275366	-38031.15
-363	815.7562	0.000038	2.23714	-33228.33
-364	818.0062	0.000038	1.61084	49343.2
-365	820.2561	0.000038	2.196916	-28272.63

For studying, the reasons for oscillations into the values of kurtosis and skewness, we must analyses the cumulants. Skewness is a function of δ that depends only of the parameter α, the range δ∈(-1,1), thus when $\alpha\to\infty$ then $\delta\to 1$ and $\alpha\to-\infty$ then $\delta\to-1$ and δ=0 when α=0. The other elements in skewness are $\zeta_2(\tau)$ and $\zeta_3(\tau)$, both of them are functions of ζ_1 that is the first derivative of $\zeta_0(x)=\log 2\Phi(x)$ without analytical expression for $\tau\in(-\infty,\infty)$. For values of τ<-50 an expression using a series expansion.

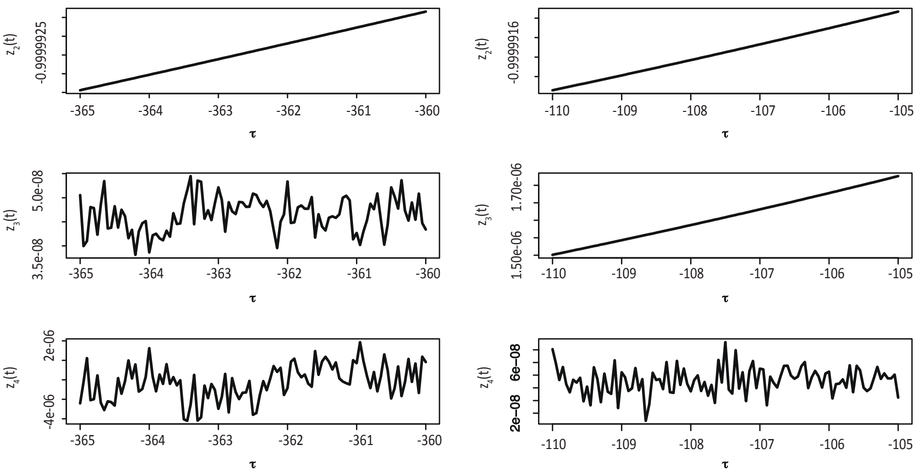


Figure 5. Evolution of $\zeta_2(\tau),\zeta_3(\tau),\zeta_4(\tau)$

For large negative values of τ, $\zeta_1(\tau)\rightarrow-\tau$ and factor $\tau+\zeta_1(\tau)\in(0,0.02)$ then range $\zeta_2(\tau)\in(-1,-0.9996)$, both elements are increasing function of τ. The shape is smooth, but when analyzing the factor ζ_3 one can observed that for values of $\tau<-100$ there is oscillations in the value in the range $(-1^{*}10^{-5},1^{*}10^{5})$, that is probably cause of oscillation in skewness, and are more evident for large negative values of τ. Oscillations exist in ζ_2 for values of $\tau<-1000$, but are not observable in ζ_1 (see Figure 5). In the graphic one can observed evolution of the components of the skewness (ζ_2,ζ_3) in the range (-110,-106) and (-365,-360), in the first case there is not oscillations and the value of skewness is smooth increasing but, in the second case, the oscillations of ζ_3 are the explanation of the movements of the values of skewness, without a defined trend.

Kurtosis is a function of ζ_2 and ζ_4, and oscillations are more evident because the presence of ζ_3 in the expression of ζ_4 produce a wider movement. Oscillations are evident for smaller negative values of τ (see Figure 5). Oscillations into ζ_2, ζ_3 and ζ_4 are a consequence of the definition of ζ_1 that is a nonlinear function of τ. The difference between two consecutive elements tends to 1 when τ tends to $-\infty$, as the range of variation is very small and the shape of ζ_1 is quasi-linear. However, this effect is augmented in calculation of ζ_2, ζ_3 and ζ_4. Variations in ζ_1 and ζ_2 preserves the sign, but for large negative values of τ preservation of sign do not hold for ζ_3 and ζ_4, and that is the reason for the existence of jumps in skewness and kurtosis.

4. CONCLUSIONS

The Extended Skew-Normal distribution (Azzalini & Capitano, 2014) is developed as a generalization of the normal distribution, considering the existence of skewness. The normal distribution would be a special case of the extended skew-normal with skewness equals zero and kurtosis equals 3 (excess kurtosis null). For the software R, a package for the extended skew-normal distribution is available, named sn. The development of the package is referred in the webpage http://azzalini.stat.unipd.it/SN/.

The Extended Skew-Normal distribution is a function of four parameters, where $\xi,\alpha,\tau\in\mathbb{R}$ and $\omega>0$. Nevertheless, the range of variation of the other parameters is not limited.

In this paper, we have observed that for values of $\tau<-8$ the statistics (skewness and kurtosis) of the distribution present oscillations and, for large negative values of τ and large values of α kurtosis presents no reasonable values.

Effects are different for skewness and kurtosis. Skewness presents a trend to established around 2 but with big oscillations when small variations of τ are considered. Kurtosis presents larger variations with values very large, and, in same cases, the sign is not according the theory of distribution.

Although the **sn** package is valid for the study of extended skew-normal distribution we might be careful for τ negative values. For $\tau<-8$, the results could not be trustable because the presence of errors in the statistics and distribution could not defined in a well-manner.

ACKNOWLEDGEMENT

This work was partially supported by Fundaçao para a Ciência e a Tecnologia, Portugal - FCT [UID/MAT/00006/2013].

REFERENCES

Azzalini, A. (1985). "A class of distributions which includes the normal ones" in Scandinavian Journal of Statistics, 12, 171-178.

Azzalini, A. & Capitanio. A. (2014). The Skew-Normal and Related Families. Institute of Mathematical Statistics Monographs (IMS). Cambridge University Press.

Canale, A. (2011). "Statistical aspects of the scalar extended skew-normal distributions" in Metron, LXIX (3), 279-295.

Capitanio, A., Azzalini, A. & Stanghellini, E. (2003). "Graphical models for skew-normal variates" in Scandinavian Journal of Statistics, 30, 129-144.

Chenglong, L., Mukherjee, A., Qin, S. & Min, X. (2017) "Some monitoring procedures related to assymetry parameter of azzalini's skew-normal model" Accepted for publication in Revstat – Statistical Journal.

Figueiredo, F., Gomes, M. (2013). "The skew-normal distribution in spc" in Revstat – Statistical Journal, 11(1), 83-104.

Team, R.C. (2015). A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

Condition Based Process Monitoring for Minimizing Fuzziness Surrounding Statistical Inference – a Philosophical Perspective with some Examples and Illustrations

Amitava Mukherjee
XLRI –Xavier School of Management, Jamshedpur, India – amitmukh2@yahoo.co.in

Abstract

With significant development in the area of statistical process monitoring (SPM) over the last nine decades, more and more complexities creep both in the theory and methods related to the SPM procedures. While we have seen the development of a plethora of interesting tools and techniques for SPM, those, in turn, often infuse many fuzzy areas within the domain of SPM and unfortunately, make a decision making process rather complicated. Starting from identification of appropriate of Phase-I distribution to choice of the best available SPM technique in Phase-II, there are certain degrees of fuzziness, in various areas. We provide some classic examples of illusions which lead to certain practical challenges both in the realms of parametric and nonparametric SPM. Certain conditions based techniques are discussed that can reduce the impact of fuzziness in some areas of SPM and ensure ease of decision making. Some philosophical perspectives are discussed rather than complex technicalities. Certain future research problems are identified.

Keywords: Fuzziness; Process Monitoring; Quality Control; Type-III and IV Errors.

This page intentionally left blank

IPS23:
Enhancement in Monetary and Financial Statistics in The Post Great Financial Crisis

Enhancing Global Financial Statistics after the Crisis – What is the Focus?

Bruno Tissot*
Bank for International Settlements, Basel, Switzerland – Bruno.Tissot@bis.org

Abstract

The BIS has significantly enhanced its statistical offering to support monetary and financial stability analyses in the aftermath of the 2007-09 crisis. A substantial part of this work has been undertaken in the context of the Data Gaps Initiative endorsed by the G20, in close coordination with central banks and international organisations. The main steps taken aim at (i) enhancing the provision of data covering financial markets; (ii) measuring the impact of international groups, by going beyond traditional residency-based statistics; (iii) providing adequate series to monitor financial cycles and systemic risk; and (iv) collecting data to support evidence-based financial regulation.

Keywords: Monetary analysis; Financial stability; G20 Data Gaps Initiative; Financial cycles.
JEL classification: E00, E32, F30, G01, G15

1. INTRODUCTION: THE GREAT FINANCIAL CRISIS OF 2007–09, A WAKE-UP CALL FOR STATISTICIANS

Many observers will agree that data issues were not the sole, nor even a major cause of the Great Financial Crisis (GFC) of 2007–09. Yet this crisis highlighted important statistical shortcomings (Borio (2013)). Four areas were particularly obvious. The first data issue related to *leverage*. The degree to which agents had borrowed was not correctly appreciated before the crisis; this was particularly the case for banks but also for a wide range of economic agents, especially households. The second area was *financial intermediation*. Countries' statistical apparatus was not sufficient to capture the activities of all the players intervening in financial markets, as highlighted by the increased importance of 'shadow banks' that provide leverage-based maturity and liquidity transformation (FSB (2015)). The third area related to *cross-sector and -border linkages*. Such interconnections proved particularly important in propagating financial stress that originated in a specific sector and in a specific country (ie, the US mortgage industry during the GFC) to the entire globe. A fourth key data gap was related to the monitoring of the situation of individual *firms that proved 'too big to fail'* in times of financial stress.

2. POST-CRISIS POLICY RESPONSE TO IMPROVE GLOBAL FINANCIAL STATISTICS

A key policy response after the GFC was to enhance the availability of global financial statistics. In 2009, the International Monetary Fund (IMF) and the Financial Stability Board (FSB) issued *The Financial Crisis and Information Gaps* report to explore information gaps and provide proposals for strengthening data collection – see IMF and FSB (2009). This initial Data Gaps Initiative (DGI-I) endorsed by the G-20 comprised 20 recommendations focussing on three key statistical domains, ie the build-up of risks in the financial sector, international network connections, and vulnerabilities to shocks. This initial phase highlighted the limited availability of reliable and timely statistical data in various domains. To address these challenges, the international community decided to launch in 2016 the second phase of the DGI (DGI-II) to implement "*the regular collection and dissemination of comparable, timely, integrated, high quality, and standardized statistics for policy use*" (IMF and FSB (2015)). Three main areas deserved attention: the monitoring of risks in the financial sector; the assessment of interlinkages; and the adequate communication of official statistics (cf Table 1).

A key support for these global statistical initiatives was, in addition to its endorsement by the G20 economies, the underlying cooperation among various international organisations. To ensure that, the

Inter-Agency Group on Economic and Financial Statistics (IAG) was tasked to coordinate and monitor the implementation of the initiative.[1] This means that improving global financial statistics has become a key objective for all the main international bodies involved in financial stability work.

Table 1. The Recommendations in the Two Phases of the Data Gaps Initiative

First Phase (2000-2005)	Second Phase (2016-20)
I.1: Mandate of the DGI	II.1: Mandate of the DGI
Build-up of risk in the financial sector	**Monitoring risks in the financial sector**
I.2: Financial Soundness Indicators (FSI)	II.2: FSI
I.3: Tail risk	
	II.3: FSI Concentration and Distribution Measures
I.4: Aggregate leverage and maturity mismatches	
	II.4: Data for Global Systemically Important Financial Institutions (GSIFIs)
	II.5: Shadow Banking
I.5: Credit default Swaps (CDS)	II.6: Derivatives
I.6: Structured products	
I.7: Securities data	II.7: Securities Statistics
Cross-border financial linkages	**Vulnerabilities, Interconnections, and Spillovers**
I.8&9: Data for Global Systemically Important	
Financial Institutions (GSIFIs)	
I.12: International Investment Position (IIP)	II.10: IIP
I.10&11 Coordinated Portfolio Investment Survey	II.11: IBS
(CPIS) & International Banking Statistics (IBS): Participation / Enhancement	II.12: CPIS
I.13&14: Financial and non-financial corporations' cross border exposures	II.14: Cross border exposures of non-bank corporations
Vulnerability of domestic economies to shocks	II.13: Coordinated Direct Investment Survey (CDIS)
I.15: Sectoral accounts	II.8: Sectoral accounts
I.16: Distributional Information	II.9: Household Distributional Information
I.17: Government Finance Statistics (GFS)	II.15: GFS
I.18: Public Sector Debt Database (PSDS)	II.16: PSDS
I.19: Real estate prices	II.17: Residential Property Prices
	II.18: Commercial Property Prices
Communication of Official Statistics	**Communication of Official Statistics**
I.20: Principal Global indicators	
	II.19: International Data Cooperation and Communication
	II.20: Promotion of Data Sharing

3. LARGE EXPANSION IN BIS'S INTERNATIONAL FINANCIAL STATISTICS AFTER THE GFC

BIS's statistics, in particular those on international banking, securities markets and derivatives activity, have been a unique source of information about the global financial system for many decades. They are compiled in cooperation with central banks and other national authorities and are designed to inform analysis of financial stability, international monetary spillovers and global liquidity. After the GFC, the BIS undertook a marked expansion of its statistical offering, by publishing additional data, revamping how these data are disseminated and strengthening their policy orientation (BIS (2015)).

Primus inter pares are the **international banking statistics** (IBS). In particular, the "locational" IBS dataset complements "traditional" monetary and credit aggregates by providing information on banks' cross-border and foreign currency positions. This balance sheets information captures the positions of banking offices located in a given country, following the same residency principles as the System of National Accounts (SNA; cf European Commission et al (2009)). Internationally active banks located in almost 50 countries, including many offshore financial centres, report their outstanding claims and

1 The IAG, established in 2008 to coordinate statistical issues and data gaps highlighted by the GFC and to strengthen data collection, comprises the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the IMF (Chair), the Organisation for Economic Co-operation and Development (OECD), the United Nations and the World Bank.

liabilities against counterparties residing in more than 200 countries. Numerous breakdowns have been added after the GFC, so that the new data allow to identify precisely the residence and nationality of the reporting banks, the residence and sector of their counterparties, and the instruments and currencies in which transactions take place. These statistics help to analyse the global geography of the capital flows intermediated by banks and how financial stress can propagate across sectors and borders.

Another key BIS dataset relates to **debt securities** statistics. It captures borrowing in money and bond markets, with a distinction between international and domestic issuance. International debt securities (IDS) are defined as those issued in a market other than the local market of the country where the borrower resides (Gruić and Wooldridge (2012)), and domestic debt securities (DDS) are those issued by residents in their local market, regardless of the currency in which the securities are denominated. Total debt securities (TDS) sum IDS and DDS together as all debt securities issued by residents – cf BIS et al (2015)). This information helps to analyse borrowers' exposures to, for instance, foreign exchange and rollover risks. It also provides light on the evolution of international funding sources, especially the relative contribution of bank and debt financing. In particular, the progressive shift from bank credit towards debt securities financing experienced by major emerging market economies has been described as the "second phase of global liquidity" (Shin (2013)).

A third important statistical area relates to the **derivatives** markets. They are the primary sources for assessing the size and structure of global derivatives markets and provide internationally consistent information. In particular, they shed light on who is transferring risks and on the aggregate amount of risk transferred. They also help to monitor the progress of efforts by policymakers after the GFC to reduce systemic risks in derivatives markets by shifting the clearing and trading of OTC instruments to central counterparties (CCPs) and organised exchanges.

The BIS compiles several sources on derivatives (Tissot (2015)). First, the BIS conducts regular surveys of "amounts outstanding" in over-the-counter (OTC) derivatives markets' segments – eg commodity, equity, foreign exchange, interest rate and, since the GFC, credit default swap (CDS) – and instruments. Various indicators are considered. A first is the *national amounts outstanding*, which is the nominal value of all the deals concluded and not yet settled. But this indicator is influenced by changing structural factors and exaggerates the real amounts at risk. A second indicator is the *gross market value*, defined as the sum of the absolute values of all open contracts: this represents the maximum loss that market participants would incur if all counterparties failed to meet their contractual payments and the contracts were replaced at market prices. A third indicator is the *gross credit exposures*, calculated as gross market values minus amounts netted with the same counterparty, and across all risk categories. It provides a sense of aggregated dealers' exposures to counterparty credit risk. The BIS OTC amounts statistics are in fact split into two closely-linked datasets. One is the *Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity (amounts outstanding part)*, which is the most comprehensive source. It covers around 1,300 banks and other dealers from more than 50 jurisdictions. The second dataset is the *Semiannual survey of OTC derivatives*, whose coverage was expanded after the GFC and is reported by large dealers in 13 jurisdictions.

Second, the BIS also conduct the *Triennial Survey/turnover part* to measure turnover in derivatives transactions. Turnover is defined as the gross value of all transactions presented as daily averages. So it is a "flow" indicator, in contrast to the "positions data" provided at a point in time by the surveys on outstanding amounts. This turnover survey covers interest rate derivatives as well as all FX transactions in general – including spot transactions, and not just FX derivatives instruments.

Third, the BIS compiles statistics covering derivatives traded on organised exchanges, complementing OTC information. These statistics, derived from various market sources, cover both the turnover and the amount outstanding of derivatives instruments traded on organised exchanges. Their coverage and details have been greatly enhanced since the GFC. The data are rapidly updated every quarter so they can be particularly useful for monitoring the international derivatives market activity. However the coverage is limited to organised exchanges and does not comprise OTC information.

4. COMPLEMENTING RESIDENCY-BASED INDICATORS WITH GROUP-LEVEL INFORMATION

After the GFC, increased policy focus on the global financial system led observers to realise that a different type of information was needed, compared to the aggregated, country-based statistics that are usually available. Indeed, worldwide information has to be considered for properly assessing firms' group-level exposures and identifying sources of financial stress – in particular as today's business model of global corporates relies to an important extent on the establishment/acquisition of entities located outside the domestic area. But operations through their foreign affiliates can only be captured by the respective residency-based statistics of the 'host' countries, and not by those of the 'home' country. From this perspective, there must be a paradigm shift in the statistical community (Tissot (2016)): statisticians need to complement their analysis with nationality-based, consolidated group data. This is necessary to understand, for instance, who makes underlying economic decisions, who takes on the final risk and who needs to hold sufficient buffers to cover global potential losses.

The BIS has been at the forefront of these efforts to develop alternative statistics to capture such issues. First in line are the **IBS**, which provide information on cross-border banking positions not only on a 'locational' but on a 'consolidated' basis (BIS (2013)). The consolidated IBS capture the worldwide consolidated positions of internationally active banking groups headquartered in reporting countries, ie excluding intragroup positions and in line with to the consolidation approach followed by supervisors. A bank's foreign claims vis-à-vis a specific country comprise its "pure" cross-border claims on that country's residents but also the claims that are booked locally by the affiliates of this bank – in contrast, such locally-booked claims would not be treated as non-residents' assets in the SNA framework, although they can represent a very sizeable part of consolidated foreign banks' claims. The data are compiled in two different ways, on an immediate counterparty basis and on an ultimate risk basis – that is, taking into account the transfer of the bank's credit exposure on its immediate counterparty to another one through the use of credit risk mitigants such as derivatives and guarantees.

As regards the **IDS**, the fact that they are compiled from a granular, security-by-security database enables unique identification of all bonds issued by the specific residency of the issuer and by its nationality defined as the residency of the parent company controlling it. One can thus compare debt issuance activity from both a residency- and a nationality-based perspective. Data by nationality are useful for identifying links between borrowers in different countries and sectors – for instance, when assessing the international issuance of emerging market borrowers in advanced markets and / or offshore centres, either directly or through their controlled affiliates (Gruić and Wooldridge (2015)).

Turning to BIS **derivatives** statistics, amounts outstanding indicators are reported on a consolidated basis by all the worldwide affiliates (ie the branches and majority-owned subsidiaries) of the head offices located in the reporting countries. Operations between affiliates of the same institution are excluded from the reporting: for instance hedging operations conducted by a local branch with its parent entity, which merely reflects intra-group risk management practices, are excluded. The main exception are the OTC turnover data, reported every three years on an unconsolidated basis by the sales desks of reporting dealers so as to provide a sense of the geographical distribution of market activity.

5. MONITORING THE FINANCIAL CYCLE AND SYSTEMIC RISK

From a financial stability perspective, a key element to consider is the role played by the financial cycle.[2] It is defined as a succession of long-lasting episodes of financial booms and busts, and characterised by a much wider amplitude and length compared to "traditional" business cycles. Both its upward and downward phases are highly influenced by developments in credit, asset prices and financial conditions more generally. These forces can be self-amplifying, with a feedback loop between overly optimistic perceptions of risk and value, on the one hand, and weak financing constraints, on the other hand – as a result, the financial system is displaying strong "procyclicality" (Crockett, 2000). One key factor is the role played by international finance: cross-border bank credit, especially in foreign currency, is often a major driver of domestic credit booms and provides the marginal source of financing in the

2 See Debt and the financial cycle: domestic and global, in BIS (2014), Chapter IV.

run-up to financial crises (see eg Avdjiev, McCauley and McGuire (2012)). The increasing influence of global factors on domestic financial cycles highlights the powerful role played by "global liquidity". This concept refers to the property of the system as a whole, resulting from the interaction of private investors, financial institutions and monetary authorities, and which can facilitate the build-up of vulnerabilities in the form of asset price inflation, leverage, or maturity or funding mismatches (Caruana (2012)). A key element supporting these mechanisms is the role played by international funding currencies, which are increasingly used outside the issuing country's borders (McCauley et al, 2015).

A significant and increasing number of BIS indicators have been constructed to shed light on these various mechanisms. They draw on national data but incorporate estimations by BIS economists and statisticians to construct specific measures – eg aggregate credit, global liquidity indicators, debt service ratios, asset prices, effective exchange rates. In addition, the focus is on long-term series, which are better suited for the analysis of the long financial cycles. Since such information is not easily available across countries, the BIS has therefore expanded its publication of long series, drawing on data provided by its member central banks and applying ad hoc statistical techniques and assumptions.

To identify financial cycles, a first area is to focus on medium-term fluctuations in property prices and credit. As regards **property prices**, they play an important role in influencing trends in aggregate demand and financial positions, with implications for both macroeconomic and financial stability. The BIS has developed since the GFC a detailed residential property price data set covering a large number of countries in the world. More recently, the BIS has also started under the aegis of the DGI to disseminate commercial property prices. As regards **credit**, the BIS publishes quarterly statistics on the borrowing of the government sector and the private non-financial sector in more than 40 economies. "Total credit" comprises financing from all sources, including domestic banks, other domestic financial corporations, non-financial corporations and non-residents. The financial instruments cover (i) currency and deposits, (ii) loans and (iii) debt securities – the sum of these three instruments being defined as "core debt", which generally represents the bulk of total debt (Dembiermont et al (2015)).

Trends in these credit series, relative to GDP, can facilitate the monitoring of financial developments and the detection of episodes of financial cycles. From a policy perspective, in particular, an excessive level of credit has been found to be a reliable early warning indicator of episodes of stress. Of course, quantifying what "excessive credit" means is not easy. One approach followed by the BIS is to construct **credit-to-GDP gap** series. They help the identification of the rapid build-up in the provision of credit to domestic agents. They also play a prominent role within the Basel III framework, by providing guidance to banking supervisors on the use of the countercyclical capital buffer introduced in 2010 to strengthen banks' defences against the build-up of systemic vulnerabilities.[3]

Other useful indicators in this context are **debt service ratios** (DSRs) compiled for the private non-financial sector. The DSR reflects the share of income used to service debt, given interest rates, principal repayments and loan maturities. By providing a comprehensive assessment of credit burdens, it helps to understand the interactions between financial conditions and demand and has proven to be an informative early warning indicator for systemic banking crises (Drehmann et al (2015)).

Credit is also a key component of the **Global Liquidity Indicators** (GLIs) of the BIS. The GLIs are constructed by mixing various statistics, in particular on bank and debt financing. Of particular interest is the international component of credit, ie cross-border lending to non-residents or lending to residents in foreign currency. Although this international component is often small relative to total credit, its swings can amplify domestic trends and appear correlated with financial booms and busts.

Another important element when gauging the impact of financial conditions relate to the exchange rates. The BIS has a long tradition of computing **effective exchange rate** indices for a large number of countries. These indices are expressed on a trade-weighted basis both in nominal and in real terms (ie adjusted for changes in relative prices). More recently, the BIS has complemented this approach by looking at the impact of exchange rates movements vis-à-vis funding currencies. The BIS **"debt-weighted" exchange rates** (DWERs) take into consideration foreign currency-denominated total debt

3 BIS credit-to-GDP gaps may differ from those considered by national authorities as part of their countercyclical capital buffer decisions (these will also apply judgment in the setting of the supervisory buffers and may use different data series).

irrespective the residence of lenders (ie excluding external debt denominated in the local currency of the borrowing country but including local debt expressed in foreign currencies). The indices are constructed by considering the bilateral exchange rates against each of the major global funding currencies weighted by their shares in the country's foreign currency debt. These calculations draw heavily on the BIS banking and securities statistics. An important conclusion is that exchange rate movements can have an impact through both a trade and financial channels, often with opposite effects: a depreciation will typically support exports and domestic output, but in contrast it can lead to a tightening in domestic financial conditions when domestic agents have foreign currency liabilities. Using trade-weighted exchange rates and the newly constructed DWERs can help to distinguish these effects (Berger (2016)).

6. DATA TO SUPPORT EVIDENCE-BASED FINANCIAL REGULATION POLICIES

A major lesson of the GFC was the importance to collect institution-level data from a financial stability perspective. Certainly, such data had been already collected for many years before, as they are indispensable for the supervisory task of micro prudential authorities. The GFC highlighted three key data needs from this perspective.

The first has been the development of policy tools that need to be applied at a granular level. An obvious example is the growing variety of macroprudential measures adopted, focussing on specific instruments, creditor sectors and borrowers. Another relates to monetary policy: the assessment of credit risk is instrumental in determining the quality and conditions of assets used as collateral in monetary operations, and which have been in increasing demand with the new quantitative easing policies.

Second, the GFC highlighted the need for supervising large financial firms at the international level. In particular, an ambitious collection and sharing exercise related to global systemic institutions has been promoted by the FSB and is being conducted with the operational support of the International Data Hub (IDH) set by the BIS (see FSB (2011)). Hub data are shared among national supervisors and macro-prudential authorities of the jurisdictions participating in the exercise, and a number of IFIs are also progressively receiving special information derived from it. Actual data have started to be collected for a subset of the global systemically important banks (G-SIBs) that have been characterised as of "systemic importance" (BCCS (2013)). Data collected encompass a variety of micro indicators – based on banks' assets (exposures), liabilities (funding) and off-balance figures (contingent positions) – aiming at assessing interlinkages among the institutions surveyed as well as with their key counterparties ("network effects") and the concentration of these institutions in specific sectors and markets ("size effects"). The set-up of the Hub was organised along three phases. Phase I, started in 2013, involved the collection of simple I-I ("Institution-to-Institution") bilateral data to measure the G-SIBs' exposures to their major counterparts. It also comprised I-A ("Institution-to-Aggregate") data to assess the concentration of G-SIBs to specific sectors and markets. These latter I-A data are in fact the institution-level data underlying the consolidated IBS of the BIS. The data collected by the IDH have thus progressively become more detailed in parallel with the post-GFC enhancements to the IBS. Phase II, launched in 2014, focused on I-I liabilities, ie information on the largest funding providers of a G-SIB, as well as on its funding structure. After the start of the implementation of Phase III in 2015, additional I-A information will be provided as from 2017 for the consolidated balance sheet of each G-SIBs, with detailed breakdowns by counterparty country, sector, instrument, currency and maturity.

A third area relates to the activities of international standard-setting bodies, especially those hosted by the BIS. Almost all new regulatory initiatives are now supported by some kind of granular data collections. Quantitative impact studies (QIS) have now become a central element of these new indicator-based frameworks developed to, among other tasks, draw the lessons of previous policies, assess the ex-ante impact of new measures, identify additional areas of weakness, and clarify the functioning of regulation by measuring feedback effects, behavioural responses and unintended consequences. Moreover, they facilitate the assessment of the cross-impact of the various regulatory requirements introduced in parallel. The BCBS has been leading ahead in this evolution, and has developed in recent years a large number of regular monitoring reports on various items such as capital regulation, liquidity rules, the selection and measurement of G-SIBs (BCBS (2017)). Other Basel-based groups are also increasingly participating in this new way of steering and implementing policy.

REFERENCES

Avdjiev, S, R McCauley and P McGuire (2012): "Rapid credit growth and international credit: challenges for Asia", *BIS Working Papers*, no 377, April.

Bank for International Settlements (BIS) (2013): Guidelines for reporting the BIS international banking statistics – version incorporating Stage 1 and Stage 2 enhancements recommended by the CGFS.

——— (2014): *84th Annual Report*, June.

——— (2015): Introduction to BIS statistics, *BIS Quarterly Review*, September.

——— (2016): Recent enhancements to the BIS statistics, *BIS Quarterly Review*, September.

BIS, ECB and IMF (2015): *Handbook on Securities Statistics*.

Basel Committee on Banking Supervision (BCBS) (2013): Global systemically important banks: updated assessment methodology and the higher loss absorbency requirement, July.

——— (2017): *Basel III Monitoring Report*, February.

Berger, B (2016): Debt-weighted exchange rate indices, *BIS Quarterly Review*, December, Box A, pp 100-102.

Borio, C (2013): The Great Financial Crisis: setting priorities for new statistics, *Journal of Banking Regulation*, vol 14, July, pp 306–17. Also available as *BIS Working Papers*, no 408, April.

Caruana, J (2012): Assessing global liquidity from a financial stability perspective, speech at the 48th SEACEN Governors' Conference and High-Level Seminar, Ulaanbaatar, November.

Crockett A (2000): Marrying the micro- and macroprudential dimensions of financial stability, *BIS Speeches*, 21 September.

Dembiermont, C, M Scatigna, R Szemere and B Tissot (2015): A new database on general government debt, *BIS Quarterly Review*, September.

Drehmann, M, A Illes, M Juselius and M Santos (2015): How much income is used for debt payments? A new database for debt service ratios, *BIS Quarterly Review*, September.

European Commission, International Monetary Fund, Organisation for Economic Cooperation and Development, United Nations and World Bank (2009): System of National Accounts 2008.

Financial Stability Board (2011): Understanding financial linkages: a common data template for global systemically important banks, Consultation paper, October.

——— (2015): *Global Shadow Banking Monitoring Report*.

Gruić, B and P Wooldridge (2012): Enhancements to the BIS debt securities statistics, *BIS Quarterly Review*, December.

——— (2015): BIS debt securities statistics: a comparison of nationality data with external debt statistics, *IFC Bulletin*, no 39, April.

International Monetary Fund and Financial Stability Board (2009): 'The Financial Crisis and Information Gaps', Report to the G20 Finance Ministers and Central Bank Governors, October.

——— (2015): 'The Financial Crisis and Information Gaps – Sixth Implementation Progress Report of the G20 Data Gaps Initiative', September.

McCauley, R, P McGuire and V Sushko (2015): Dollar credit to emerging market economies, *BIS Quarterly Review*, December.

Shin, H (2013): 'The Second Phase of Global Liquidity and Its Impact on Emerging Economies', Keynote address at the Federal Reserve Bank of San Francisco Asia Economic Policy Conference, November 3–5 2013, Princeton University.

Tissot, B (2015): Derivatives statistics: the BIS contribution, *Invited Paper Session 089*, International Statistical Institute World Statistics Congress 2015.

——— (2016): Globalisation and financial stability risks: is the residency-based approach of the national accounts old-fashioned?, *BIS Working Papers*, no 587, October

Ana Credit – Overview and Implementation from an NCB`s Point of View

Sebastian Grünberg*
Deutsche Bundesbank, Frankfurt am Main, Germany – sebastian.gruenberg@bundesbank.de

Abstract

On May 18th 2016, the Governing Council of the European Central Bank (ECB) approved the Regulation (EU) 2016/867 of the ECB on the collection of granular credit and credit risk data ("Analytical Credit Datasets"; short: "Ana Credit"). Ana Credit is a statistical project to establish a common granular credit database comprising detailed information on individual bank loans in the euro area based on harmonised ECB statistical reporting requirements. Since data is collected at borrower-by-borrower and loan-by-loan level, this granular database enables a high flexibility with regards to analysis. Moreover, in line with the statistical legal framework of the European System of Central Banks (ESCB), this approach allows the collected credit data to be used for multiple purposes, both across institutions and across different user groups within institutions. Ana Credit will, therefore, support the ESCB in a number of central bank tasks, as for example monetary policy analysis and operations and financial stability surveillance. In addition, Ana Credit shall also serve external institutions such as the European Commission, the Bank for International Settlements or the International Monetary Fund. This brief paper aims at introducing Ana Credit and presenting an overview of its scope. Furthermore, it provides insight into the implementation of Ana Credit from a National Central Bank`s (NCB`s) point of view.

Keywords: Credit Data; Granularity; Granular Credit Data.

1. INTRODUCTION

Demand for statistics on a granular basis has grown as a result of recent financial and economic crises. This can be traced back to the beginnings of the last global financial meltdown in 2008 revealing that aggregate statistics were not sufficient for an adequate and detailed overview of the developments that had been taking place inside the global economy and the financial system.

However, these developments have to be understood and monitored going forward in order to react in an appropriate way to economic shocks or, at best, predict and maybe even prevent crises in advance.

Especially the knowledge about the distribution of data has shown to be necessary for a proper understanding of the causes and economic interdependencies in the financial system. As a consequence, this need can be regarded as a trigger for a paradigm shift from aggregate to granular micro data in the world of statistics.

In this context, the granular credit database "Ana Credit", which is currently set up within the Eurosystem, provides a good example for a new database of granular micro data and will, therefore, be an important component in realising this paradigm shift. Improving statistical information on credit and credit risk, current data gaps could be closed and multiple objectives could be met at once.

Ana Credit shall be introduced in the following paragraphs in detail. Thereafter, it is shed light upon some aspects of its implementation in Germany.

2. ORIGIN AND GENERAL CONCEPT OF ANA CREDIT

Ana Credit stands for "Analytical Credit data sets" and is a statistical project initiated to establish a granular harmonised credit database comprising detailed information on individual bank loans in the euro area. The European Central Bank (ECB), euro area and some non-euro area National Central Banks (NCBs) initially launched the project in 2011. A fundamental decision on the planned introduction of

Ana Credit had been taken by the Governing Council of the ECB in 2014[1] defining the general scope of the project and, thus, starting the work on a draft regulation.

The potential scope and complexity of Ana Credit sparked interest of media and discussions among the public and banking industry during 2015. As a consequence, the ECB decided to publish a draft regulation of Ana Credit on 4th of December 2015 shortly after the Governing Council had in principle agreed on this draft version. Following this, the public was given the opportunity to comment the draft regulation until end of January 2016.

The comments received – mainly from banking associations - were assessed and accordingly some adjustments to the draft regulation were made. As a result, the Governing Council of the ECB finally approved the Regulation (EU) 2016/867 on the collection of granular credit and credit risk data on May 18th 2016 with the data collection scheduled to start in September 2018.

The Regulation foresees that loan-by-loan credit data is collected by the NCBs of the euro area and forwarded to the ECB. NCBs of non-euro area member states of the European Union (EU) can voluntarily participate in Ana Credit.

3. USE OF ANA CREDIT[2]

Ana Credit shall serve multiple purposes and close current data gaps. This can be achieved as loan-by- loan data in such granularity, as Ana Credit will provide, enables a high flexibility with regards to analysis. Depending on the specific purpose customized data requests and analyses can be run. Thus, a full re-use of micro data can be accomplished.

As data is planned to be used for multiple purposes, the statistical legal framework of the ESCB[3] allows institutions marked in red in Table 1 to access highly granular information on a need-to-know- basis. The remaining institutions listed in Table 1 may access Ana Credit data, but not those which are confidential.

Table 1. Examplary User Institutions of AnaCredit Data

- European Central Bank (ECB)
- NCBs of European System of Central Banks (ESCB)
- European Commission (Eurostat)
- European Systemic Risk Board (ESRB)
- Bank for International Settlements (BIS)
- International Monetary Fund (IMF)
- World Bank
- Financial Stability Board (FSB)

Above all, the ECB and NCBs of the ESCB will use and benefit from Ana Credit data. For instance, Ana Credit will support the performance of the following central bank tasks:

Monetary Policy Analysis and Operations: Although small and medium-sized enterprises (SMEs) are regarded as the backbone of the economy, there is only scarce information on loans granted to that sector. Ana Credit will close this data gap and help to better understand the monetary policy transmission mechanism also in this regard. In order to capture loans to SMEs a relatively low reporting threshold of 25,000 € was chosen for Ana Credit.

Financial Stability Surveillance: Risks to financial stability can arise when large institutions have negative external effects ("Too big to fail?"), when institutions are strongly interrelated ("Too connected to fail?"), but even if many small institutions are exposed to similar risks ("Too many to fail?"). Ana Credit will enable to recognize these risks and their distribution to individual financial institutions, sectors or entire European countries at an early stage.

1 Decision ECB/2014/6 and Recommendation ECB/2014/7 of the European Central Bank of 24 February 2014 on the organization of preparatory measures for the collection of granular credit data by the European System of Central Banks

2 Please see also Damia, V., &Israël, J-M. (2014). Standardised granular credit and credit risk data. Paper presented at 7th IFC conference on "Indicators to support Monetary and Financial Stability Analysis: Data Sources and Statistical Methodologies".

3 Council Regulation (EC) No 2533/98 of 23 November 1998 concerning the collection of statistical information by the European Central Bank

Risk Management: Ana Credit will contain certain risk attributes (for example *probability of default*) on a borrower-by-borrower basis which allow assessing the creditworthiness of debtors.

This will be especially helpful for the Bundesbank. Currently, credit institutions may use credit claims from creditworthy debtors as monetary policy collateral. So far, the creditworthiness of debtors in Germany is assessed on the basis of data received via the German Central Credit Register (CCR)[4]. However this data basis will not be sufficient anymore as new ECB requirements for the NCB`s Inhouse Credit Assessment System (ICAS) apply as of March 2018. From then on, the German CCR can no longer deliver the required data since its one Million € threshold is too high and only quarterly reporting is available. Ana Credit data, however, will also fulfil the new ECB standards.

Once implemented, Ana Credit provides the possibility to combine its credit data with micro data of other statistics as for example the Centralised Securities Database (CSDB); this way the full indebtedness of a borrower could be captured.

Beyond that, Ana Credit has the potential to allow the ESCB to (partially) consolidate existing statistical requirements and, to a certain extent, even replace them in the medium term. As a result, both central banks and reporting agents could potentially benefit from a reduced data collection effort.

4. SCOPE OF ANA CREDIT

The Regulation (EU) 2016/867 determines that for the time being only credit institutions[5] and resident foreign branches of credit institutions have to report information on loans and the counterparties involved into these loans, mainly on a monthly basis.

Information merely has to be reported in case

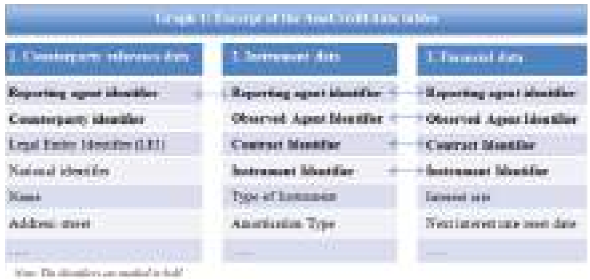
1) A loan is granted to a corporation or another legal entity (a loan solely granted to a natural person is exempted from being reported), and
2) All summarized outstanding loan amounts of the debtor granted by the reporting credit institution are above 25,000 €.

The scope of loans to be reported is outlined in the subsequent list:

- Deposits other than reverse repurchase agreements;
- Overdrafts;
- Credit card debt;
- Revolving credit other than overdrafts and credit card debt;
- Credit lines other than revolving credit;
- Reverse repurchase agreements;
- Trade receivables;
- Financial leases;
- Other loans[6].

With regard to the data structure, Ana Credit consists of ten different data tables that are connected to each other via identifiers. In general, Ana Credit data can be distinguished between counterparty reference data which contain basic information[7] on debtors and creditors and are included in one data table and credit data that are contained in the remaining nine data tables.

4 Groß- und Millionenkreditmeldewesen
5 Credit institutions are defined according to Article 4 (1) (1) of Regulation (EU) No 575/2013 of the European Parliament and of the Council
6 Other loans not included in any of the categories listed above. Loan has the same meaning as defined in paragraphs 5.112, 5.113 and 5.114 of Annex A to Regulation (EU) No 549/2013.
7 For example the address or the institutional sector of a debtor/creditor



Graph 1. Exemplarily Displays the Connection of the Data Tables and Highlights the Role of the Identifiers

Overall, Ana Credit comprises 89 attributes and six identifiers. Table 2 summarizes the main elements of the Ana Credit scope.

Table 2. Main elements of the AnaCredit scope

Type	Requirements
Creditor	• Credit institutions (incl. branches)
Debtor	• Corporations & other legal entities
Threshold	• 25,000 €
Attributes	• 89 attributes • 6 identifier
Reporting frequency	• Mainly monthly
Schedule	• Counterparty reference data: March 2018 • Credit data: September 2018
National discretion	• NCBs may collect data from December 2017 onwards • NCBs may grant derogations to small reporting agents (i.a.)

Besides the aforementioned aspects, Regulation (EU) 2016/867 entitles the NCBs to determine parts of the Ana Credit scope at national discretion. NCBs may grant derogations to small reporting agents[8] for instance.[9] The derogations may cover some or all reporting requirements defined in the Regulation, i.e. an NCB could decide that small reporting agents do not have to report at all in the respective country.[10]

With regard to the time schedule the NCBs are also given leeway.[11] While the Regulation requires credit data to be reported first time by reference date September 2018 and the counterparty reference data by reference date March 2018, NCBs have the right to collect both counterparty reference and credit data already from December 2017 onwards.

Furthermore, it is intended to establish Ana Credit in stages. If the ECB`s Governing Council seeks an extension of the Ana Credit scope, the decision on a subsequent stage has to be taken at least two years prior to its implementation.[12]

5. ANA CREDIT IMPLEMENTATION IN GERMANY

The aforementioned option to derogate small institutions from their reporting obligations, which is given by Regulation (EU) 2016/867, plays a central role for the understanding of the Ana Credit

8 Small reporting agents are defined in Article 16 (1) of Regulation (EU) 2016/867
9 According to Article 16 (1) of Regulation (EU) 2016/867
10 Derogations in Regulation (EU) 2016/867 are not limited to the example presented above, but the Regulation contains more. Due to the scope of this paper, however, not all derogations can be highlighted.
11 According to Article 2 (3) of Regulation (EU) 2016/867
12 Recital (10) of Regulation (EU) 2016/867

implementation in Germany. After the draft regulation was published by the end of the year 2015, it became common knowledge that the Ana Credit Regulation would contain derogations which would be up to the decisions of the respective NCBs.

A full exemption of small credit institutions in Germany could relieve the banks affected of an additional reporting burden. However, it has to be noted that there are approximately 1800 credit institutions in Germany to be considered reporting agents for the purposes of Regulation (EU) 2016/867. Out of the 1800 roughly 740 reporting agents would be regarded as "small"[13], which represents a substantial number. Bearing in mind the aim of Ana Credit to also fill data gaps with respect to SMEs, which are financed by small banks to a large extent, a full reporting exemption would entail not covering loans to that sector in a comprehensive way.

Besides, there was another important reason why a full exemption of small reporting agents would be difficult to implement in Germany:

As presented above, new ECB standards with regard to the NCB`s ICAS apply as of March 2018. While the German CCR could no longer fulfil the new requirements from then on, Ana Credit could close this gap and deliver the required credit data with its monthly frequency and the relatively low threshold of 25,000€.

As ICAS standards count on full coverage, however, it therefore requires reporting from all credit institutions, impeding a full exemption of small reporting agents. In addition, the Bundesbank would have to bring the time schedule forward a few months using its right to collect credit data before September 2018, if it intended to meet the ICAS standards from March 2018 onwards.

Taking all these factors into account, the Bundesbank had to ponder how to decide. On the one hand, the small reporting agents should not be overburdened, on the other hand, data gaps should be closed and compliance with the new ICAS standards be achieved.

The following optimisation challenge had to be solved:

- Derogate the small German credit institutions as far as possible in accordance with Article 16 (1) of Regulation (EU) 2016/867;
- Use implementation of Regulation (EU) 2016/867 to make the Bundesbank ICAS fit for the new ECB requirements;
- Give the necessary information in respect of the transmission mechanism (including loans to SMEs) to the monetary policy users.

With the approval of Regulation (EU) 2016/867 in May 2016 the Bundesbank finally decided as follows:

- Instead of collecting first credit data from September 2018 onwards, the Bundesbank decided in favour of an earlier start date. Counterparty reference data would be reported first time by reference date January 2018 and credit data by March 2018.
- Partial derogations were granted to small reporting agents. Although this did not mean a full exemption, the Bundesbank reduced the reporting burden to an absolute minimum. Small reporting agents would have to report only 17 indispensable data attributes for loans originated prior to 1 September 2018 and nine additional data attributes for loans originated after 1 September 2018. Table 3 displays these attributes and provides an insight into the data collected.

13 Please see footnote 8

Table 3. Attributes to be Reported by Small Reporting Agents in Germany

Counterparty reference data	Credit data
• National identifier • Name • Address: street • Address: city / town / village • Address: postal code • Address: country • Accounting standard	• Type of instrument • Default status of the instrument • Date of the default status of the instrument • Outstanding nominal amount • Accrued interest • Off-balance sheet amount • Counterparty role • Probability of default • Default status of the counterparty • Date of the default status of the counterparty
Additional attributes for loans originated after 1 September 2018	**Additional attributes for loans originated after 1 September 2018**
• Institutional sector • Economic activity • Number of employees • Balance sheet total	• Currency • Interest rate type • Purpose • Interest rate • Accumulated impairment amount

This way compliance with ICAS requirements and closing data gaps also with respect to SMEs could be achieved while at the same time taking into account the principle of proportionality regarding the reporting burden of small credit institutions.

Shortly after the approval of Regulation (EU) 2016/867 the Bundesbank announced the German derogations to the banking industry. To the satisfaction of all parties the solution was regarded as a good compromise.

6. CONCLUSION AND OUTLOOK

As outlined in this paper, Ana Credit will provide granular credit data that can be used in a flexible manner and will greatly support the performance of key central bank tasks of the ECB and NCBs of the ESCB. Be it monetary policy or financial stability surveillance, Ana Credit will enable *on demand* analyses that allow a better understanding of the monetary policy transmission channel and disclose the linkages and relationships inside the financial system. On that improved data basis, better conclusions can be drawn enhancing qualitative decision making and increasing the likelihood of choosing appropriate measures to crises, shocks or unwanted conditions.

The use of Ana Credit data, however, is not only limited to the ECB and NCBs of the ESCB, but the statistical legal framework facilitates that, in principle, also certain other institutions can gain data access.

Furthermore, an extension of the Ana Credit scope is conceivable and in principle intended. Later potential stages of Ana Credit could contain for example housing loans to private households. In addition, Single Supervisory Mechanism (SSM) requirements could be also considered at a later stage. Whether the implementation of further steps will be realized, however, depends on future decisions of the ECB`s Governing Council.

These steps, once taken, could further improve the data base, but even without, Ana Credit already represents another step towards micro data oriented statistics and can be regarded as a milestone with respect to credit data statistics in Europe.

Using The Zoom Lens in Banking Statistics[1]

Jean-Marc Israël and Rodrigo Oliveira-Soares
European Central Bank, Frankfurt am Main – jean-marc.israel@ecb.int and rodrigo.oliveira@ecb.int

Abstract

Since the onset of the financial crisis the ECB has put additional emphasis on the analysis of granular data for developments in bank credit, securities markets, and in the interbank market. This paper reviews some initiatives in the area of monetary and financial statistics to complement aggregated statistics with granular data referring to institutions, contracts and counterparts.

In particular, the ECB is now collecting and sharing across the Eurosystem several granular data, such as individual banks' reports and their holdings and issues of securities, as well as granular daily data to monitor the wholesale money market – protecting data confidentiality, as required. This allows better supporting the monetary policy conduct and its implementation and monitoring risks to financial stability. In addition, a major project (Analytical Credit) aims at collecting from the second half of 2018 and analysing in-depth supply and demand factors in credit developments across countries, counterpart sectors, and according to several relevant dimensions.

Granular data allow understanding economic and financial developments better, for example by running timely economic analyses of changes at different levels of aggregation such as macro-economy or sectors, or by assessing stress factors and their possible effects on financial stability.

Keywords: Euro, European Central Bank, Central Banking, Statistics, Monetary Policy, Transmission Channel, Financial Stability, Credit, Bank Loans, Money Market, Granular Data.

JEL Classification: E42

1. INTRODUCTION

Since the onset of the financial crisis the ECB has put additional emphasis on the analysis of granular data for developments in bank credit, securities markets, and in the interbank market. In particular, granular data has allowed uncovering heterogeneities across the banking sector such as different distributions of developments according to the size or location of banks in the euro area.

In particular, the ECB is now collecting and sharing across the Eurosystem several granular data, such as individual banks' reports covering balance sheet items, interest rates on deposits and loans and holdings and issues of securities. In addition, from July 2016 major players report granular daily data transaction by transaction to monitor the wholesale money market. This allows better supporting the monetary policy conduct and its implementation and monitoring risks to financial stability.

Furthermore, a major project (Analytical Credit) aims at collecting supply and demand factors in developments of banking credit. Those developments will be analysed across countries, counterpart sectors, and according to several relevant dimensions. The first data will be reported in the second half of 2018.

Granular data allow understanding economic and financial developments better, for example by running timely economic analyses of changes at different levels of aggregation such as macro-economy or sectors, or by assessing stress factors and their possible effects on financial stability.

In order to support the collection and analysis of granular data on economic and financial activities and sectors, central banks in European countries jointly operate and maintain reference/master data on statistically-relevant institutions, the Register of Institutions and Affiliates Database.

1 The current version has benefited from useful input and suggestions from Josep Maria Puigvert, Riccardo Bonci, Peter Neudorfer, Asier Cornejo Pérez and Maciej Anacki. The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank.

2. THE REGISTER OF INSTITUTIONS AND AFFILIATES DATABASE

A prerequisite for the collection of various types of data on financial transactions is that each bank in the EU is uniquely identified and described with harmonised reference data in a common register, namely the Register of Institutions and Affiliates Database (RIAD). In this context RIAD also captures changes in the population of banks owing to mergers and acquisitions, new banking licences or closures.

Identifying each bank or other reporting agent in a unique way is of particular relevance when compiling statistics for an entire economic area, like the euro area. Lists created from RIAD allow reporting agents and compilers in each member country to properly identify foreign counterparts and allow classifying them by country of residency and institutional sector. For example, national central banks classify in RIAD each resident investment fund either as a money market fund (MMF), which belongs to the money-issuing sector, or as a non-MMF investment fund, which belongs to the money-holding sector. Thus RIAD allows a harmonised classification across all European countries.

The identifiers in RIAD also allow national central banks and ECB putting together information from several sources, such as banks' balance sheets, interest rates in loans and deposits, securities issues and holdings, money market transactions and in the future loan-by-loan data.

The RIAD reference dataset can be used in different ways for statistical purposes, such as selecting a sample of the population to be surveyed, or combining individual data or entities to group structures, or even conglomerates, building up on various criteria. For example, securities holdings statistics are compiled along the perimeter of pan-European groups of banks that are considered systemically important. All in all, RIAD constitutes a backbone both for the collection of data from banks and other reporting agents, and the later use of those data also for analytical and operational purposes within central banks. An accurate picture on who is who helps preparing policy decisions.

3. BALANCE SHEET AND INTEREST RATE STATISTICS

Data on balance sheets and interest rates for euro area banks have been collected by the ECB on a country-by-country basis since 1999 and 2003 respectively. Traditionally, national central banks would share the data with the Eurosystem based on national aggregates only. While these aggregates remain a key component for the euro area indicators, granular data have also become important to better analyse the monetary policy transmission channel and disruptions caused by fragmentation in the markets. Hence, the ECB has established since September 2012 a regular transmission of individual data, with a view to sharing the dataset across the Eurosystem for monetary policy as well as macro-prudential and financial stability analysis, in line with the legal framework underlying the collection of statistical information by the ECB.[2] Tools and processes needed to be adapted so as to handle microdata. In addition, data confidentiality had to be protected.

Originally, from September 2012, the dataset was limited to the main items of the balance sheet and covered a panel of about 250 banks representing 70% of the euro area MFI sector in terms of total assets. The dataset then was gradually extended both in terms of indicators and of institutions. Since October 2015, the coverage has increased to about 300 banks (representing 80% of the euro area banking sector in terms of total assets) and the granularity of the indicators was extended.[3] For instance, for the balance sheet indicators a total of about 130 indicators on outstanding amounts and over 30 on transactions are made available. In particular, the dataset covers cash, loans, debt securities, money market fund (MMF) shares/units, non-MMF investment fund shares/units, equity, non-financial assets and remaining assets. On the liability side, deposits, debt securities, capital and reserves, and remaining liabilities are presented. Data are split by residency and sector of the counterparty, maturity (where relevant), purpose (for loans to households) and type (for deposits). Regarding interest rates, the dataset covers loans and deposits to euro area households and non-financial corporations (NFCs), and distinguishes between interest rates on outstanding amounts (13 indicators) and interest rates on new business (30 indicators, with the corresponding business volumes). The breakdowns broadly match those available for the balance sheet

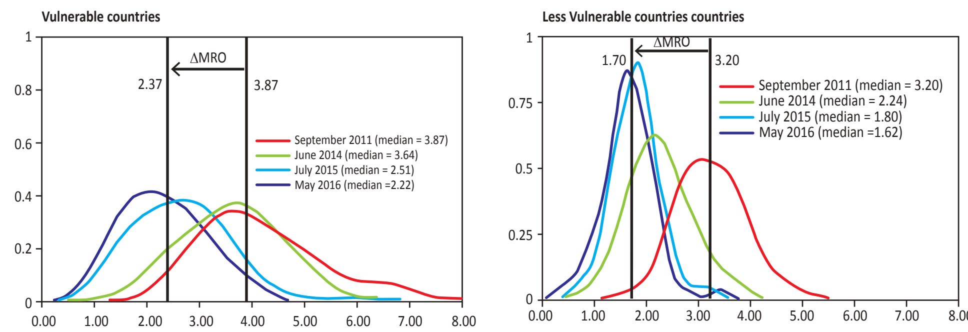
2 Council Regulation (EC) No 2533/98 of 23 November 1998 concerning the collection of statistical information by the European Central Bank, OJ L 318, 27.11.1998, p. 8.

3 The last extension introduced micro-prudential supervision as one of the purposes for which the data can be used.

indicators in terms of maturity (or fixation period for the new business) and instrument type for deposits or purpose for loans to households. Loans to non-financial corporations are additionally made available with a split for the amount of the loan.[4]

Since its establishment, this rich dataset has enabled to enhance monetary analysis with cross-sectional studies, as the responses of individual banks to monetary policy easing and unconventional measures became much affected by their individual characteristics. The standard ECB tool sets for assessing money and credit has thus been complemented with distributional analyses and cross-sectional studies using individual data. Besides the monetary policy transmission mechanism, the microdata can also be used to study the funding conditions of banks, not least as a consequence of prudential or regulatory changes, their profitability (e.g. interest rate margins) or balance sheet structure, allowing for analyses across business models, bank types and countries.[5]

Chart 1 extends the analysis of ECB (2015b) and shows that the pass through of successive cuts in the rate of main refinancing operations to lending rates applied to NFCs was much slower for vulnerable countries than in less vulnerable ones. Comparing the distributions of lending rates of September 2011 (i.e. shortly before the first of a series of cuts in the rate of the ECB main refinancing operations (MRO) starting in November 2011) and June 2014, it appears that vulnerable countries did not experience a significant drop on the median rate (23 basis points) despite the 125 basis point reduction in the MRO rate. In contrast, this reduction was better reflected in the median rate for less vulnerable countries (96 basis points). Since the new asset purchase programmes in June 2014, the reduction in borrowing costs in vulnerable countries was significant, supporting the view that the measures have helped aligning the price of credit with the intended stance of monetary policy.



Source: ECB

Note: The chart shows the density of the lending rate distribution obtained from a sample of 56 MFIs in selected vulnerable countries (Ireland, Spain, Italy and Portugal) and 106 MFIs in less vulnerable countries (Belgium, Germany, France, the Netherlands and Austria) in four different periods (September 2011, June 2014, July 2015 and May 2016). The chart also shows that if the reduction in the MRO rate since September 2011 (i.e. 145 basis points) had been fully passed on to the median lending rates of that period (i.e. 3.87% for vulnerable countries and 3.20% for less vulnerable countries), the lending rate in May 2016 would have been 2.37% for vulnerable countries and 1.70% for less vulnerable countries.

Chart 1. Composite Lending Rates for NFCs: Distribution of Individual MFIs (percentages per annum)

4. HOLDINGS AND ISSUES OF SECURITIES

Securities statistics comprise information that is granular both in terms of institutions (i.e. bank-by-bank) and instruments (i.e. security-by-security). The granular information is collected for each bank's assets -securities holdings- and liabilities -securities issues. The collection of the data follows different processes for assets and liabilities. For assets, each bank reports its securities' holdings, which are then classified according to different criteria. For liabilities, the ECB collects available information from market sources, numbering agencies and national central banks to construct a centralised securities database (CSDB).

4 See Israël *et al* (2016).

5 For instance, see ECB (2014, 2015*b*, and 2016a). Much academic research has also been performed in this context; for example, see Altavilla *et al* (2016) and Holton and Rodriguez d'Acri (2015).

The CSDB contains comprehensive reference data for each security, as well as market prices, issuer and ratings information. It covers both equity and debt securities. Also investment fund shares/units issued by EU residents are registered. All securities are covered whether issued by banks, by investment funds, insurance corporations, non-financial corporations or governments. The focus is on securities likely to be held and traded by EU residents as well all euro-denominated securities, regardless of the residency of the issuer and holders. The reference data reflect different variables such as the residency and institutional sector of the issuer, coupon rates and redemption date of debt securities, corporate events, etc. The CSDB plays a dual role in the compilation of banks' securities statistics, as liabilities can directly be derived from the CSDB, while banks report assets in securities held in their portfolios, which are matched with CSDB data to assess statistical and other classifications, as well as allowing different valuations.

In addition to the monthly data for balance sheet items mentioned in the previous section, banks report their portfolio security-by-security under the framework of the securities holdings statistics. Each security is identified by a single code, which is the International Securities Identification Number (ISIN) whenever it exists. Security-by-security reporting means that the compiler receives raw information on the amount held for each security, and produces all relevant statistical breakdowns and aggregates by combining those amounts with the characteristics of each security in the CSDB. For example, securities holding statistics feed macroeconomic statistics such as external statistics (balance of payments and international investment position) and the financial accounts by sector.

Following the security-by-security approach, NCBs and the ECB classify the securities by instrument, residency and sector of the issuer, etc., in a more reliable classification than anyone made by reporting agents. In addition, different valuation of securities can be performed thanks to the CSDB for different purposes. This ensures consistency in classifying and pricing the securities reported, as well as a symmetrical recording on assets and liabilities. Furthermore, without reporting burden the ECB can compile debt securities portfolios based on, e.g., original or residual maturities.[6]

5. DAILY MONEY MARKET TRANSACTIONS

Monitoring money markets is crucial for the analysis of the monetary policy transmission (as well as for macro- and micro-prudential supervision), especially in a situation of high market fragmentation. In a situation where such fragmentation was expected to remain very high, in 2014 the Eurosystem decided to establish a legal framework to collect statistical data on money market transactions.[7] This dataset provides the Eurosystem with daily, accurate, timely (in the early morning on the working day following the deal) and comprehensive data on transactions concluded by the reporting banks, which will allow an improved monitoring of the transmission of monetary policy decisions in money markets, as well as on market expectations for the future evolution of policy rates.

This new granular dataset covers four segments of the euro money markets, namely unsecured, secured, foreign exchange swaps and overnight index swaps (OIS) transactions denominated in euro. The new collection framework is based on the daily reporting of transaction-by-transaction information on unsecured and secured lending and borrowing transactions in euro with a maturity of up to one year. All foreign exchange swap transactions involving euro and OIS transactions denominated in euro must also be reported. The detailed trade data to be provided include the volume, rate, counterparty type and collateral type, together with the time at which the transaction was conducted.[8]

On 1 July 2016 the ECB started to collect statistical data from the 52 euro area banks with the largest market share in money market segments. With a view to limiting the impact of teething problems and ensuring full automation from 1 July 2016, banks had started to send data on 1 April 2016. This three-month interim period was deemed necessary to fine-tune the reporting process before the legal obligation came into force on 1 July. As shown in Chart 2, the number of transactional records already reached ca. 45,000 per day in early 2017.

6 See European Central Bank (2015a).

7 Regulation (EU) No 1333/2014 of the European Central Bank of 26 November 2014 concerning statistics on the money markets (ECB/2014/48), OJ L 359, 16.12.2014, p. 97.

8 For further information, see ECB 2016b and 2016c.

The data are available at 07:30 in the morning of the next working day, allowing an early assessment to be used as input to the daily monitoring of liquidity by the ECB. The data granularity will also allow more in-depth analysis of market developments. This granularity and timeliness requires that the statistical analysis, e.g. on consistency and plausibility, is run with highly effective and automated processes, part of which will need to be developed and enhanced while gaining experience. In this regard, a full standardisation of the underlying taxonomy and data transmission format based on the ISO 20022 standard has been introduced by the Eurosystem. A set of four reporting messages and a status message containing feedback information have been jointly submitted by the ECB, the Deutsche Bundesbank, the Banco de España and the Banque de France and subsequently been approved by the ISO Registration Authority.

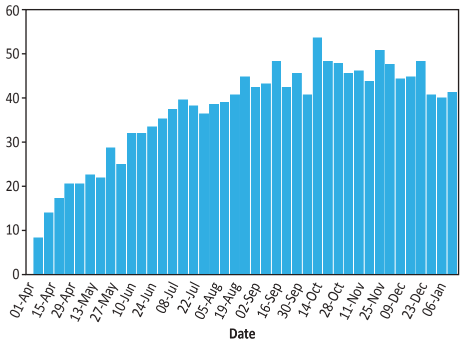


Chart 2. Transactional Records Collected (Weekly Averages, Thousands)

6. THE ANALYTICAL CREDIT DATASETS

The ECB has adopted a strategy to develop and produce new ESCB granular statistics on credit and credit risk with the aim to support the Eurosystem in the performance of its tasks, including those related to monetary policy analysis and operations, risk management, financial stability surveillance and economic analysis and research. The granular data will also be used to produce new indicators of credit intermediation and increase the quality of existing statistical datasets.

To this end, a new statistical regulation was adopted in May 2016,[9] with a view to establishing a common, harmonised granular credit and credit risk dataset (i.e. "AnaCredit") shared between the Eurosystem members and comprising input data from all euro area countries. The AnaCredit dataset will allow to closely monitor the status of the euro area credit market, both as a whole and in specific segments, relying on complete, accurate and timely information on credit and credit risk in the financial system. The new legal framework calls upon a first stage of implementation with a focused scope, namely lending and credit lines by banks to all legal persons, in particular non-financial corporations. The scope may be later on extended to other lenders and other instruments. A phased-in approach makes the overall endeavour more manageable, while the data model, definitions and granularity of data allow any future enrichment to be processed in a forward-looking, non-disruptive manner.

Covering bank loans to non-financial corporations already in the initial phase, the AnaCredit dataset will allow a better understanding of the monetary policy transmission channel, particularly concerning small and medium-sized enterprises (SMEs) – the backbone of the European economy in terms of investment and employment opportunities. AnaCredit will provide high-quality and timely information on creditors, debtors, guarantors, and on the different credits extended (i.e. identity of the lender, of the borrower, type of credit, outstanding debt, number of days past due date, date of origination and contractual maturity, type of interest rate and currency of the credit). Such a credit dataset calls for a

9 Regulation (EU) 2016/867 of the European Central Bank of 18 May 2016 on the collection of granular credit and credit risk data (ECB/2016/13), OJ L 144, 1.6.2016, p. 44.

reference/master dataset with accurate information on all counterparties. While lenders in the financial sector are nearly all covered within the RIAD dataset, the scope of the register will be much expanded to also cover the non-financial entities reported as borrowers or guarantors in AnaCredit.

Based on information available in AnaCredit – combined via RIAD with the Securities Holdings Statistics and the CSDB where applicable – it will be possible to assess the overall indebtedness of a given (e.g.) company and, thus, its creditworthiness. Furthermore, on the lenders' side, information on any risk mitigation measures securing the credits (e.g. credit derivatives, guarantors, financial collateral received) will help estimate the severity of losses in the event of default. Finally, the dataset will support the reliable identification of each individual creditor and debtor (e.g. name and unique ID number, address, type of obligor, firm size, sector of economic activity), as a prerequisite to capture the total indebtedness of debtors accurately, especially of those debtors with cross-border exposures.[10]

7. CONCLUSIONS

The statistics function of the Eurosystem has met new, resource-intensive demands as regards granular data for banks. Value for analysis of the data has further much increased. In particular individual data on balance sheets, interest rates and securities holdings and issues enable economists using the zoom lens to interpret aggregated developments. These demands have translated into significant operational challenges, which in turn also allowed strengthening statistical processes.

Moreover, in the near future the richness of the AnaCredit dataset combined with RIAD, also in liaison and SHSDB/CSDB where appropriate, may allow rethinking the way aggregated statistics are compiled and collected from reporting agents and, going forward, could possibly lead to a decrease in data requirements in the context of reports of balance sheets and interest rates.

An efficient data collection that enables combining the data from different frameworks is facilitated by a common register at the European level. The RIAD dataset is instrumental for creating lists that allow banks to classify their counterparts according to the country of residency and institutional sector, and for the compiling agencies to accurately classify parties in transactions and group banks and firms according to various criteria.

Furthermore, the collection of granular data from banks and other reporting agents has often been developed separately from the collection of data that feed macroeconomic statistics such as the monetary or credit aggregates. A challenge is now to help minimise reporting burden (of e.g. banks) and better servicing users with consistent aggregated and granular data. This will benefit from further integration and streamlining so that in one go banks report data for different purposes, which is the goal of the European Reporting Framework initiative. Already, this is the case of the insurance industry in Europe that reports in most countries only once for supervisory and statistical purposes.[11]

10 For further information, see ECB (2015e).

11 There is a close connection between the insurance corporations' data collected by national central banks (NCBs) for statistical purposes under Regulation ECB/2014/50 and the data collected by national competent authorities (NCAs) for supervisory purposes under the framework established by Directive 2009/138/EC of the European Parliament and of the Council. Given the ECB's general mandate to engage in cooperation with other bodies in the field of statistics, and in order to limit the administrative burden and avoid the duplication of tasks, NCBs may derive the insurance statistics from data collected under Directive 2009/138/EC, including the national law implementing that Directive, having due regard to the terms of any cooperation arrangement between the relevant NCB and the relevant NCA. Article 70 of Directive 2009/138/EC provides that NCAs may transmit information intended for the performance of their tasks under that Directive to NCBs and other bodies with a similar function in their capacity as monetary authorities.

REFERENCES

Altavilla, C., Pagano, M. and Simonelli, S. (2015), "Bank Exposures and Sovereign Stress Transmission," CSEF Working Papers 410, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy, revised 11 Apr 2016.

European Central Bank (2010a), "The Centralised Securities Database in brief," February 2010.

European Central Bank (2010b), "The ECB's response to the financial crisis," Monthly Bulletin, Issue 10/2010.

European Central Bank (2014), "Extension to the models for assessing money and credit," Monthly Bulletin, Issue 2/2014.

European Central Bank (2015a) "Who holds what? New information on securities holdings," Economic Bulletin, Issue 2/2015.

European Central Bank (2015b) "The role of the central bank balance sheet in monetary policy," Economic Bulletin, Issue 4/2015.

European Central Bank (2015c), "New features on monetary and financial statistics," Economic Bulletin, Issue 8 / 2015.

European Central Bank (2015d), "Developments in euro area banks' funding conditions and retail rates," Financial Integration in Europe, April 2015\

European Central Bank (2015e), "Explanatory Note on the Draft ECB Regulation on the Collection of Granular Credit and Credit Risk Data."

European Central Bank (2016a), "Recent developments in the composition and cost of bank funding in the euro area," Economic Bulletin, Issue 1/2016.

European Central Bank (2016b), "Reporting Instructions for the Electronic Transmission of Money Market Statistical Reporting."

European Central Bank (2016c), "The launch of money market statistical reporting," Economic Bulletin, Issue 5/2016.

Holton, S. and Rodriguez d'Acri, C. (2015), "Jagged cliffs and stumbling blocks: interest rate pass-through fragmentation during the Euro area crisis," Working Paper Series 1850, European Central Bank.

Israël, J-M., Colangelo, A. and Oliveira-Soares, R. (2016), "Recent ECB experience of rapidly evolving monetary policy and its statistical implications", paper presented in the 8th Conference of the Irving Fischer Committee.

Upgrading Monetary and Financial Statistics in The Wake of the Financial Crisis – There's Life Beyond Aggregate Data[1]

Luís Teles Dias*
Banco de Portugal, Lisbon, Portugal – ldias@bportugal.pt
António Jorge Silva
Banco de Portugal, Lisbon, Portugal – anjsilva@bportugal.pt

Abstract

The Great Financial Crisis and the ensuing unconventional monetary policy response from central banks have increased the demand for comprehensive, high quality and more detailed statistics. In this context, traditional aggregate statistics have proved insufficient to monitor and interpret the multiple aspects of the monetary transmission mechanism and the evolution of credit to companies and households. This paper presents the recent developments that have occurred in *Banco de Portugal* in the area of monetary and financial statistics as a response to these new challenges. In a nutshell, it was considered important to complement the traditional aggregate monetary and financial statistics with more granular data, so as to increase the flexibility and timeliness of the data, while fomenting the integration of micro and macro level information. This strategy allowed for, *inter alia*, the development of (i) new or upgraded statistics, (ii) easier and faster responses to *ad hoc* requests and (iii) more user-tailored analyses – all features that have proved essential for effectively addressing the challenges posed by the *post* financial crisis.

Keywords: Monetary policy; Central bank's statistics; Monetary and financial institutions' balance sheet; Credit analysis

JEL classification: E42; E52; E58; G21

1. INTRODUCTION

The Great Financial Crisis (GFC) has underscored the gaps in the availability of key information for policy making and for the timely assessment of risks in and across countries. In fact, several aspects of the economic and financial developments, and particularly, the accumulation of risks and imbalances in some sectors and/or financial instruments, were not easily identified by the traditional aggregate statistics.

In 2009, the G-20 Finance Ministers and Central Bank Governors, having recognised the existing limitations, endorsed a set of 20 recommendations to close the data gaps identified in order to support enhanced policy analysis (the so-called "G20 Data Gaps Initiative[2]"). Several of the Recommendations focused on improving the data available on the financial sector and financial institutions given the central role that the monetary and financial sector represent to economies, and its part on the GFC. Monetary and financial statistics (MFS) are particularly relevant in supporting policy making, specifically that of a central bank – they deliver an absolutely critical set of information that provides a comprehensive picture of monetary developments, contributes to the monitoring of the transmission mechanism of monetary policy, helps assessing financing conditions in different sectors of an economy and monitoring financial integration.

In this paper we will present the main developments that have recently occurred in the area of MFS in *Banco de Portugal* (hereinafter referred to as the "Bank"). Some developments were a response to euro area policy initiatives, such as the creation of the Single Supervisory Mechanism (SSM) or the decision

1 The analysis, opinions and findings of this paper represent the views of the authors, which are not necessarily those of the Banco de Portugal or of the Eurosystem

2 For more background information on this initiative please see: http://ec.europa.eu/eurostat/statistics explained/index.php/G20_Data_Gaps_Initiative_(DGI)_%E2%80%93_background

to establish a European analytical credit System (AnaCredit[3]), but, mostly, they are the continuation and deepening of a strategy based on a rich ensemble of granular databases and on the integration of such databases. This strategy allows to complement the traditional aggregate monetary and financial statistics with more detailed data, which increases the flexibility as regards the compilation of new statistics and a more rapid response to *ad hoc* data requirements from the users, while fostering the integration of micro and macro level information.

These efforts have been producing results, as put forward in this paper. In Section 2 we will provide a brief review of the impacts the Single Supervision Mechanism and the pursuit of unconventional monetary policy in the euro area, in the perspective of a national central bank's statistical function, Section 3 delves on the main developments in the Banks' MFS and presents some examples of the results that we have obtained and Section 4 concludes.

2. THE SINGLE SUPERVISORY MECHANISM AND UNCONVENTIONAL MONETARY POLICY – A BRIEF REVIEW

As work progressed on the G-20 Data Gaps Initiative, policy necessities did not wait for the identified data gaps to be filled. On the contrary, partly as a response to the GFC, several policy initiatives have been put forward and the demand for comprehensive, high quality and more detailed statistics has been made even more evident. In the euro area two developments have been particularly relevant in stressing the need to go beyond the aggregates in the domain of MFS: (i) the pursuit of unconventional monetary policies by the European Central Bank (ECB) and (ii) the creation of the SSM.

Unconventional monetary policies

Following the GFC, central banks around the world moved beyond their traditional operating framework and implemented unconventional monetary policies. These policies include zero or negative reference rates, long-term liquidity provision to banks and expanded asset purchase programs.

In order to thoroughly understand how these unconventional monetary policies affect the funding conditions for households and non-financial corporations (NFCs) and the overall economic prospects – *i.e.*, how the monetary transmission mechanism works under these new circumstances – there has been a call from economists and policy makers for unconventional data, which are not aggregate but granular. As highlighted by Ms Sabine Lautenschläger[4], member of the ECB Executive Board, "*conducting (…) unconventional monetary policy is rather difficult when decisions have to be taken on the basis of conventional data, i.e. traditional aggregate statistics*".

Traditional aggregate statistics, although of high quality and internationally harmonised, refer to the average of the distributions and, hence, are not the most adequate to explore the heterogeneity hidden behind the aggregates. In fact, given that in many situations it is the tails of the distribution that provide the most important information, it is clear why these data became crucial in the context of the recent GFC.

For instance, an overall growth in credit to NFCs, which could typically be interpreted as a positive development in the credit markets after a financial crisis, may in fact hide significant differences in the evolution of credit to different types of NFCs. Particularly, the funding may only be directed at large, established firms, while small and medium sized NFCs may have been left out, or, similarly, funding may be flowing to firms with a top-tier credit rating while the rest of firms may be facing constraints to their activities because of a lack of credit. In such cases, looking only at the aggregates or looking into the detailed granular data may lead to very different conclusions by policy makers in terms of policy stance and the risks that may be building up to financial stability.

3 The name AnaCredit stands for "Analytical Credit Datasets"
4 Lautenschläger, S. (2016).

Single Supervisory Mechanism

As part of the answer to the GFC, and with the ultimate goal of building a stronger and sounder banking system in the euro area, the EU decided to establish a single supervisory mechanism in the euro area, involving the national supervisors and the ECB [5]. This Eurosystem's newly entrusted financial supervision demands high quality and harmonized data from all the banks established in the euro area and has benefited significantly from exploring important synergies with the statistical reporting. In fact, integrating both functions – bank supervision reporting and banking statistics – allows reaping large benefits, which positively affect both the data compilers and the reporting entities: the former benefit from the existing infra-structure and the expertise accumulated over time; the latter benefit from a reduced reporting burden through the mitigation of data redundancies and overlapping.

In order to maximize the synergies, new bodies were created at the European level. In particular, the recently created Working Group on Supervisory Statistics is tasked with the collection, production and dissemination of supervisory data harmonised under the European Banking Authority's Implementing Technical Standards and any other additional supervisory data necessary for the SSM. Additionally, the Statistics Committee edorsed the creation of the Task Force on European Reporting Framework (TF ERF), following a recommendation by the "*Groupe de Réflexion*" on the integration of statistical and supervisory data, this recognizing the importance of data requirements harmonization. According to its mandate, the TF ERF shall design integrated reporting schemes, covering a wide range of different statistics, namely credit institutions balance-sheet statistics, money and interest rates, securities holdings and credit statistic, while liaising with SSM structures and other groups to maximize the potential of its work.

The data needs of the SSM can thus be seen as an interesting opportunity to maximize synergies between supervision and statistical activities. Mr Pedro Duarte Neves, Vice-Governor of *Banco de Portugal*, identified three areas in which such synergies may be achieved[6]:

"(i) Concerning data collection and information systems, integrating the reports for both functions will generate large benefits, not only for the data compilers but also for the reporting entities. In this context, highly granular data collection schemes are proving to be fundamental.

(ii) A wide range of analytical studies, which have been crucial for supervision and financial stability, benefit significantly from micro data. These analyses reveal the heterogeneity hidden behind aggregate numbers and allow for a better understanding and monitoring of the financial system, thus providing the supervisor a closer and more comprehensive perspective of the financial sector and of its relations with the other sectors in the economy.

(iii) The core supervisory data, granular credit data and ad-hoc data sets, collected and treated by statistics, will generate value not only for the direct supervision but also for the horizontal functions of the SSM, including sector-wide reviews and for identifying trends and emerging risks."

3. BEYOND THE AGGREGATES – GRANULAR DATA IN *BANCO DE PORTUGAL*

The use of integrated micro-databases for statistical purposes constitutes the cornerstone of the Bank's long-term strategy as regards not only the statistical function, but also other areas within the central banks' competencies – *inter alia* monetary policy, financial stability, supervision and research. The following databases play a vital role in such a strategy:

a. The Central Credit Register (CCR), which contains granular information on credit on a borrower-by-borrower basis (and, in some cases, including details that provide loan-by-loan information) with a virtually full coverage.
b. The Central Balance Sheet Database (CBSD), which holds accounting and financial information covering almost exhaustively the existing resident NFCs.

5 The SSM started in the 4 November 2014. For more information on banking supervision in Europe please see https://www.bankingsupervision.europa.eu/home/html/index.en.html

6 Neves (2014).

c. The Securities Statistics Integrated System (SSIS) database, a security-by-security and investor-by-investor system of both securities holdings and issuances. SSIS complements the CCR data on loans with data on securities and, from a portfolios' perspective, it is a powerful tool to measure the exposure of banks and non-banks to specific issuers; also, putting together the information contained in SSIS and CCR provides a more complete overview of the exposure and indebtedness of the financial system as a whole.

In addition, following a data request in the context of the Economic and Financial Assistance Programme to Portugal[7] and, to better assess current credit conditions of the NFCs sector and monetary policy transmission, the Bank started collecting individual data on new bank loans and their respective interest rates. This new database covers all new operations starting with reference period December 2014 (in its initial stage it was confined to banks with monthly volumes of new loans of €50 million or higher). Combining these individual data with reference data and data available in other databases, we are able to study how interest rates vary according to the characteristics of the firms.

The Portuguese Central Credit Register – a multipurpose tool

Central credit registers are a fundamental tool that facilitates the monitoring of credit risk and its evaluation by banks when granting new credits and allows an overview of credit exposures and the level of indebtedness of both resident and non-resident borrowers *vis-à-vis* national financial intermediaries.

The compilation of comprehensive statistics on credit granted is one of the various goals of the Portuguese CCR. With this in mind, credit instruments and other variables related to the classification of loans are defined in such a manner that they are meaningful for economic analysis. Also, borrowers have to be classified according to proper statistical criteria (*e.g.*, by institutional sector, sector of economic activity, firm size and region of residence). Since the participating institutions only report the borrowers' identifications (*i.e.*, their taxpayer numbers), the statistical classification of the resident borrowers is made in the Bank, mostly by means of a business register.

Statistical information based on the Portuguese CCR data is made available to users on a monthly/quarterly basis. In both cases, the main focus is loans granted by the financial sector to the resident entities classified as NFCs, non-profit institutions serving households and households.

The set of statistical indicators disclosed monthly includes:

a) Outstanding amounts of loans granted and the correspondent annual change of rate;
b) Overdue loans ratios;
c) The percentage of borrowers with overdue loans.

These indicators are compiled for borrowers belonging to the NFCs and households sectors. In the former sector, information is also broken down by firm size. Furthermore, there is a specific set of indicators related to the evolution of credit to exporting companies – this information allows to monitor the access of credit to NFCs which have a significant share of its business with non-resident counterparties and may thus be more insulated from domestic woes. In case of households, a breakdown according to the purpose of the loan is also included. Data using the above-referred metrics are provided for non-profit institutions serving households without additional breakdowns.

More detailed information is disseminated on a quarterly basis, both for the outstanding amounts of regular loans and for loans in default. In the latter case, two indicators are published: overdue loans ratio and percentage of borrowers with overdue loans.

In the case of NFCs, for the referred metrics, data is further broken down by:

a) Region of residence of the company headquarters (according to NUTS[8] classification);
b) Economic activity sector (according to NACE[9] sections);

7 For more information on the Portugal's Economic and Financial Assistance Programme please see: https://www.bportugal.pt/en/page/efap-and-post-programme-surveillance.

8 Nomenclature of Territorial Units for Statistics

9 Statistical Classification of Economic Activities in the European Community

c) Brackets of total amount of loans per borrower. As to households, data are further broken down by:
d) Purpose of the loan;
e) Region of residence (according to NUTS classification and by municipality);
f) Brackets of total amount of loans per borrower.

The Bank has recently enlarged the set of statistical indicators on loans that are compiled on the basis of CCR data[11] , and has published such information starting on the 1st quarter of 2016. Additional breakdowns of the loans granted by the financial sector have been made available:

a) Information related to the main financial products, loans' original and residual maturities and guarantees has been included;
b) Loans granted to non-financial corporations are broken down by corporation size and a distinction is made between public and private corporations;
c) Indicators about the relationship between entities of the financial sector and their credit clients and about the activity of the CCR has been introduced.

The high-quality figures that can be obtained from specific breakdowns of CCR credit data are of great importance for economic analysis and for quality control. In addition, the use of the CCR has made it possible to reduce the reporting requirements in the context of the Bank's MFS, thus alleviating the participants' reporting burden and curtailing data redundancy.

The impact of AnaCredit in the Portuguese CCR

In order to obtain a better overview of the level of indebtedness of the borrowers across European Union Member-States the European System of Central Banks has been exploring, since 2007, the potential statistical use of CCRs. In particular, it sought to understand to which extent their content may be enhanced and adapted to euro area and European Union statistical needs, to minimise the statistical reporting burden and to increase transparency.

Against this background, the ECB launched the so-called AnaCredit project in 2011, together with experts from both the statistical and credit registers' areas of a number of euro area and non-euro area national central banks. Following this avenue, a joint Statistics Committee / Financial Stability Committee Task Force on Analytical Credit Datasets (co-chaired by the Banco de Portugal) was established in 2013. The overarching aim of this task force was the setting up of a long-term framework for the collection of harmonised granular dataset on bank loans in the euro area.

With a view to fulfilling the AnaCredit's requirements, the Portuguese CCR will be redesigned and will adopt a new philosophy: a loan-by-loan basis. Although the first stage of AnaCredit will comprise only loans granted by banks to legal entities, the Portuguese CCR will keep the current coverage both in terms of participating institutions and borrowers[12].

The redesign of the Portuguese CCR is not only due to the need to adapt to the AnaCredit's requirements; rather, there will be a paradigm shift for the Bank in which the CCR will be the single entry point for all credit data that is reported to the Bank, thus creating a multipurpose hub of credit information that will be used by the several functions of the Bank.

In-house credit assessment system

The Bank has recently taken decisive steps towards further exploring the informational potential of the CCR and balance sheet databases by developing an in-house credit assessment system (ICAS)[12]. This system will provide the Bank with its own internal credit risk assessment system, thus reducing its dependence on external sources. Against the background of the recent economic and financial crisis

10 For more information, please see: https://www.bportugal.pt/sites/default/files/anexos/documentos-relacionados/pr_22_2016_crc.pdf.
11 For more details on this project and on the Portuguese CCR as a multipurpose tool please see Matos (2015).
12 https://www.bportugal.pt/en/comunicado/press-release-banco-de-portugal-new-house-credit-assessment-system-icas

and the shortage of assets liable to be used as collateral in monetary policy operations, these systems have recently been gaining importance within the Eurosystem, as can be seen by the increasing number of national central banks that have introduced them. In fact, at the current juncture, a more pressing business case for ICAS stems from monetary policy purposes, for which ICAS will provide an evaluation of debtors' credit notation.

But the benefits of such a system are not exclusive to monetary policy. In fact, there is a broad range of advantages to different business areas, in particular regarding financial supervision and stability. First and foremost, for financial supervision the credit notations derived from ICAS could be used as a benchmark to gauge those provided by institutions with their own internal notation system. Furthermore, the computation of sectoral default probabilities could also be envisaged, providing a useful input for stress-testing. As to financial stability, the monitoring of developments of the non-financial sector (and the potential building-up of imbalances) would benefit from an indicator of NFCs credit risk, which could serve, at least, two purposes: (i) to identify situations of potential financial fragility in a set of companies of a particular economic activity sector; (ii) to contribute to assess other risks stemming from the NFCs sector. Other business areas such as economic analysis and statistical functions would also stand to gain from ICAS's outputs.

4. CONCLUSIONS

The Great Financial Crisis and the ensuing response from central banks have increased the demand for comprehensive, high quality and more detailed statistics. In this context, traditional aggregate statistics have proved insufficient to monitor and interpret the multiple aspects of the monetary transmission mechanism and the evolution of credit to companies and households. The *Banco de Portugal*'s statistical function has been following a strategy based on the integrated management of micro-databases and has recently deepened this strategy in the area of monetary and financial statistics. Complementing aggregate data with more granular data is not only an answer to the need for more flexible and detailed information, it is also a movement towards a more efficient and reliable system. Particularly, when it is possible to substitute the several reports of aggregate data representing different perspectives on a given reality (for instance, credit), with a single report of granular data. In those instances, of which the Portuguese Central Credit Register is an example, central banks can have one multipurpose granular database, consistent and coherent, which can be used by the several functions of the Bank, with each one of those functions analysing the granular data by its own perspective thus increasing not only the flexibility of the data, but also the efficiency of the reporting systems.

REFERENCES

Branco, J. S., Lima, F., Monteiro, O. (2015). What changed in financial intermediation in the aftermath of the crisis? – evidence from Portugal. Supplement 1|2016 to the Statistical Bulletin, Banco de Portugal.

Drumond, I., Lima, F. (2016). How to keep statistics' customers happy? Use micro-databases!. Irving Fisher Committee Bulletin No. 41, Proceedings of the IFC Workshop on "Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges", Warsaw, Poland.

European Central Bank (2015). The role of the central bank balance sheet in monetary policy. Economic Bulletin, Issue 4/2015.

Lautenschläger, S. (2016). Central bank statistics: moving beyond the aggregates. Eighth ECB Statistics Conference.

Lima, F., Mota, S. (2016), Unconventional monetary policy – is there a call for unconventional statistics? Eight IFC Conference.

Matos, J. C. (2015), The Portuguese Central Credit Register: a powerful multi-purpose tool, relevant for many central bank's functions. Irving Fisher Committee Workshop hosted by Narodowy Bank Polski.

Neves, P. D. (2014). Enhancing the synergies between the SSM and statistical reporting. Seventh ECB Statistics Conference.

This page intentionally left blank

IPS24: Inference in Complex Method

Estimating Optimal Dynamic Treatment Regimes with Shared Decision Rules

Bibhas Chakraborty*
Duke-National University of Singapore Medical School, Singapore – bibhas.chakraborty@duke-nus.edu.sg

Abstract

A dynamic treatment regime consists of decision rules that recommend how to individualize treatment to patients based on available treatment and covariate history. In many scientific domains, these decision rules are shared across stages of intervention. Estimating these shared decision rules often amounts to estimating parameters indexing the decision rules that are shared across stages. In this talk, we will present a novel simultaneous estimation procedure for the shared parameters based on Q-learning. Through an extensive simulation study, we will present the merit of the proposed method over reasonable competitors, in terms of the treatment allocation matching of the procedure with an "oracle" procedure, defined as the one that makes treatment recommendations based on the true parameter values as opposed to their estimates. We will also look at bias and mean squared error of the individual parameter-estimates as secondary metrics. We will illustrate the proposed methodology by analyzing data from STAR*D, a multistage randomized clinical trial for treating major depression. (This is based on joint work with Palash Ghosh, Erica E.M. Moodie, and A. John Rush.)

Keyword : Dynamic treatment, Simultaneous estimation, Q-learning.

High-Dimensional Variable Selection for Spatial Regression Models

Taps Maiti *
Michigan State University, East Lansing, MI USA – maiti@stt.msu.edu

Abstract

High-dimensional variable selection techniques are popular for wide range of applications. While the recent developments are mainly suitable for independent data models, we develop methods for dependent data, specifically for spatial regression models. Our methods are applied to variety of applications, such as brain image data analysis and hedonic house price models.

Keywords: First keyword; Second keyword; Third keyword; Fourth keyword.

Modeling Volatility of Daily Returns on Investments Using Spline Functions

Dumaria R Tampubolon[1]
[1]Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung 40132, Indonesia – dumaria@math.itb.ac.id
Don McNeil[2]
[2] Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand – dmcneil@mq.edu.au

Abstract

Financial institutions, such as banks, investment companies and insurance companies, are required to set up enough capital to manage investment risks. One measurement used for such purpose is the Value-at-Risk (VaR) of such investment. In determining the VaR, the volatility of (daily) return of the investment need to be modelled and GARCH(1,1) is one time series model often used to model the volatility. In GARCH(1,1) model, the volatility is dependent on the square of the return. We attempt modelling the volatility using the information on the volatility itself rather than including the square of the return in the model. A GARCH(1,1)-like model for volatility is assumed and generated. Simulated daily return data, based on past daily return data, is smoothed using natural cubic spline functions. It can be seen that the generated volatility data is present in the simulated daily return data.

Keywords: Volatility; Value-at-Risk; Return of Investment

IPS26:
Health Statistics For The Wellbeing Of Nation

Recognizing The Importance of Private Hospitals in Elevating Malaysia's Health Care Industry

Sayeeda Kamaruddin
Department of Statistics Malaysia, Putrajaya, Malaysia – sayeeda@stats.gov.my

Abstract

Malaysia has over the last 50 years moved away from an agricultural and commodity-based to manufacturing, a higher value-added sector and as it moved up the industry value chain, Malaysia has emerged as a services driven economy. Anchored by services sector, Malaysia aims to elevate the country to a high income nation by the year 2020. To become one, Malaysia must rely on high value-added industries which are driven by high technology, knowledge intensive, high-skilled labours and demand - driven as the engine of growth. One of the industries that have such characteristics is the health care industry. Since Malaysia's independence, the country has made enormous progress in health care, creating an internationally acclaimed health system that delivers quality care to the population. The health system in Malaysia consists of both public and private healthcare system where the public health sector is heavily subsidized and widely accessible by the public. On the other hand, the private health sector predominantly concentrated in the urban areas, often *offers* advanced *specialized services* to those who are able to afford out-of-pocket fees or protected by health plans. Although the local industry has been dominated by the public healthcare sector, the market for private sector has grown considerably in accordance with the increasing demand from expanding middle class group resulted from the resilient economic growth and increasing urbanization. This was confirmed by the exhilarating growth of private health sector which has outpaced the overall performance of Gross Domestic Product since 2012. This paper focus on the rapid expansion of the private health sector, particularly the private hospitals that has substantial contribution to health sector.

Keywords: Gross National Income, Public health sector, Public health sector, Gross Domestic Product

Determinants of The Status of Completeness Basic Immunization in Children Age 12-59 Months in Aceh Province 2015 (The Aplication of Ordinal Logistic Regression Analysis)

Ghaida Nasria Azzahra
BPS-Statistics of Indonesia
ghaidaazzahra@yahoo.com
Budyanra*
Institute of Statistics, Jakarta, Indonesia
budy@stis.ac.id

Abstract

Province of Aceh has basic immunization coverage toddler lowest in Indonesia in 2015. Even though, this province has Posyandu and Puskesmas ratio per population of the highest in the western region of Indonesia. This data their concerns regarding immunization coverage has not been handled well in Aceh Province. This papers aims to identify variables that affect the status of complete basic immunization of children aged 12-59 months in Aceh by using ordinal logistic regression analysis. Ordinal logistic regression model used is proportional odds models. Data are obtained from National Socio-Economic Survey (Susenas) 2015 that was held in March 2015 by BPS-Statistic of Indonesia. Based on the results of processing data, known only 37.7% of children aged 12-59 months in the province of Aceh in 2015 which gets fully immunized, the remaining 50.6% receive primary immunization but is not complete, even about 11.7% have not received basic immunization at all. From the proportional odds model results showed that the number of children born to mothers (odds ratio = 0.88), maternal age at delivery (odds ratio = 1.03), the level of maternal education (odds ratio = 1.22), and the educational level of the household (odds ratio = 1,2) have a significant impact on the status of complete basic immunization of children. Future studies are expected to include the element of timeliness and add other variables and also with other models in ordinal logistic regression.

Keywords: Immunization, Ordinal Logistic Regression, Proportional Odds, Susenas

Gaining Insight of Malaysia's Health-Related Industries through Health Satellite Account (HSA)[1]

Khairul Aidah Samah

Department of Statistics Malaysia, Putrajaya, Malaysia aidah.samah@stats.gov.my

Abstract

Healthcare industry is one of Malaysia's 12 National Key Economic Areas (NKEAs) identified as a driver of economic growth with several projects designated to boost this industry. Moreover, current demographic developments especially in urbanisation and aging population, and the increasing health-conscious lifestyles have opened doors to various health-related businesses beyond the usual healthcare services. Therefore, it is imperative for the government to provide comprehensive statistics to keep track of this booming industry. Following the success of the Tourism Satellite Account and Information and Communication Technology Satellite Account, the Department of Statistics Malaysia (DOSM) is now embarking on constructing the Health Satellite Account (HSA) as outlined in DOSM's Transformation Plan 2015-2020. The realisation of HSA will provide prudent method to gauge the size of health-related industries and their contribution to the nation's economy. This paper highlights the importance of HSA in quantifying the overall performance of health-related industries.

Keywords: Health Satellite Account, System of National Accounts, Supply & Use Tables

1. INTRODUCTION

In the international arena, health was one of the agenda under the Sustainable Development Goals (SDGs) which is to ensure healthy lives and promote well-being for all at all ages (The Sustainable Development Goals Report 2016). According to the OECD report (2001), health is an important input to well-being and economic performance. In Malaysia, the healthcare industry is one of Malaysia's 12 National Key Economic Areas (NKEAs) identified as a driver of economic growth with several projects designated to boost this industry. The share of household expenditure on health expanded from 1.4% (2004) to 1.7% in 2014 based on Household Income and Expenditure Survey (HIES) conducted by Department of Statistics Malaysia (DOSM). Besides, government of Malaysia has put a substantial allocation of RM1.5 billion or 3.3% of total federal government development expenditure to the health industry in 2016 (2015: RM1.4 billion; 3.5%). Hence, the growing importance of the health-related industry to the economy has driven the interest of economist and policy makers in measuring its significance to the economy. As such, the satellite account is required to measure the importance of the health-related industries.

The achievement of DOSM in compiling Tourism Satellite Account (TSA) and Information and Communication Technology Satellite Account (ICTSA) allows in expanding the satellite accounts to other industries. Thus, under the Transformation Plan 2015-2020, DOSM has initiated to develop Health Satellite Account which is will be one of the department's new products. The concept and methodology of Health Satellite Account is based on the System of National Accounts (SNA). Satellite accounts use the SNA as the central of reference framework and designed to meet the needs for analysis of specific economic and social sectors. For this purpose, the use of complementary or alternative concepts and classifications is permitted, as are extensions of the central framework to introduce additional dimensions. The main objective of HSA is to analyse the total supply and demand of health products to the economy and contribution of health-related industries to the economy. The HSA also provides the information on the interaction between the health-related industries and the rest of the world. It involves the identification of health products, activities and classifications within the national accounting framework. Therefore this article discusses the insight of Malaysia's health-related industries via the creation of satellite account.

1 The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of Department of Statistics, Malaysia.

Such account would provide a comprehensive picture of health-related industries within an integrated framework in which consistent definitions, measurement tools, and analytics conventions are used.

2. FEASIBLE STUDY IN DEVELOPING HSA BY SELECTED COUNTRIES

There are two prominent health accounting approaches: The Health Satellite Account (HSA) and the System of Health Accounts (SHA). The HSA approach that is discusses in this paper provides a comprehensiveness of the health care goods and services consumed by the country's resident and it permits to examine the production of health care goods and services and the interaction between the health sector and the rest of the world. The primary objective of the HSA is to inform public policies and decision making on programs and projects related to the health sector for macroeconomic growth and development in the economy. This informs health sector planning and broadens policymakers' understanding of the contribution of the health-related industries to the national development.

The first version of HSA manual was produced in 2005, proposed and prepared by a group of international and Pan American Health Organization (PAHO) experts. The document contains methodology for the formulation of health sector accounts, under the System of National Accounts 1993 (SNA 93) analytical framework.

While the SHA is developed by the Organization for Economic Cooperation and Development (OECD), Eurostat, and the World Health Organization (WHO) in 2011. The objective of the SHA is to guide health systems policy making, and to enable allocation of health resources in close alignment with health system strategies and objectives. According to Cogswell (2015), the SHA data only shows the flow of funding, from origin to destination, for the health care goods and services that are consumed by a country's residents. This will inform decision making in budgeting, strategic planning and provide internationally comparable health expenditure data.

The HSA and SHA have both been used by a wide range of countries. The latest version of SHA was developed through a consensual process led by OECD, EUROSTAT, and WHO. In SHA development, health accountants representing all regions in the world agreed on what and how to measure health expenditure of the nation. Meanwhile, the HSA has been used primarily in the Latin America and Caribbean region. To the author knowledge, countries that have officially released health satellite account were Brazil (2013), Chile (2007), Ecuador (2013) and Mexico (2014).

However, interest in HSA has sparked in other parts of the world such as Germany. The compilation of the German HSA for the year 2010 enables them to work on Input-Output analysis for the health industry. Input-Output accounting shows the inter and intra industry in related to the health industry as well as demand and supply analysis. By compiling the supply and use table the linkage within industry could be realised, which ultimately provides a better insight of the health-related industry.

Following the experience in producing the Malaysia's Tourism Satellite Account (TSA) and Information and Communication Technology Satellite Account (ICTSA), it facilitates the realisation of HSA for prudent method to gauge the size of health-related industries and their contribution to the nation's economy.

At this juncture, both satellite accounts produced by the DOSM, provide a detailed framework for depth understanding of tourism and ICT industries. Nevertheless, both accounts have different concept and approach. ICTSA concept is to analyse the detailed of all aspects of demand for goods and services related to ICT, to observe the interactions with the supply of such goods and services of ICT within an economy, as well as with other economic activities within and outside of Malaysia. Hence, ICT satellite account compilation in Malaysia use the framework of supply and use tables. Unlike ICTSA, TSA is a demand-based concept defined not by its output but by its demand. The demand side comprises of inbound and domestic tourism expenditure.

Since the HSA is designed for a comprehensive picture of health-related industries, a detailed of supply and demand of goods and services is essential in order to capture the production of health care goods and services and the interaction between the health sector and the rest of the economy. Thus the HSA's concept and approach are similar to the compilation of ICTSA.

3. FRAMEWORK OF HEALTH SATELLITE ACCOUNT

The basis of Malaysia's health satellite account (HSA) compilation is through supply and use approach. Based on Eurostat Manual of Supply, Use an Input-Output Tables (2008), the supply table shows the supply of goods and services by product and by type of supplier. It is comprised of output of domestic industries and total imports by product. While the use table shows the use of goods and services by product and by type of use, i.e. as intermediate consumption of industries, final consumption, gross capital formation and exports. It also contains the components of value added by industry, i.e. compensation of employees, other taxes less subsidies on production, consumption of fixed capital and net operating surplus.

Hence, a supply and use framework includes the accounts of industries by type of economic activity and the accounts of transactions of goods and services by type of product. These tables provide a detailed analysis of the production process and the supply and use of goods and services, as well as the income generated from such production.

3.1. Classifications of HSA

The classifications provide the basis for design of the accounting framework and compilation of the Health Satellite Account (HSA). In this regards, basic classifications must be established that reflect the characteristics of the health sector in the country.

As recommended in the HSA manual 2005 prepared by the Pan American Health Organization (PAHO), the classification of activities in the HSA is used to identify and detail the productive activities directly or indirectly related to this sector. Two major groups of activities are considered (**Appendix A**):

a. **Characteristic activities** related to the health industry, whose production is of interest in the analysis; and
b. **Related activities** were the activities that produce health-related goods. Related activities do not necessarily require an analysis of their production; it is the supply and use of health-related goods and services that are of interest, rather than the process of producing them.

These activities been identified based on the International Standard Industrial Classification of All Economic Activities (ISIC) which correspondence to Malaysia Standard Industrial Classification (MSIC). Meanwhile, goods and services or products for the HSA divided as follows (**Appendix B**):

a. **Characteristic products** are those typical for health;
b. **Related products** depends on the availability of information, the needs of the level of analysis, and country priorities.**;** and
c. **Other products** are those that are neither characteristic nor related. These can form a single group or can be shown in more detail, considering the product classifications in each country.

Their classification should be based on the Central Product Classification (CPC) with the details from Malaysia Classification of Products by Activity (MCPA).

Currently Malaysia's HSA industry classification consists of 18 categories with 49 industries whereas the HSA products classification has 19 categories and 285 products. The level of disaggregation and classification of Malaysia's HSA are constructed based on the interest of a country and realign with HSA manual. At this moment, the classification is still under development and purview of compilers.

3.2. Data Sources and Tables

The compilation of HSA embraces both public and private spending and the Supply and Use Table which requires a good National Accounts as well as many other routine survey and systems data. As such, the basic requirements for compiling HSA for Malaysia are as follows:

a. Economic census;
b. Annual Surveys of establishments;
c. Household expenditure survey (HES);
d. Labour force survey (LFS);
e. Balance of Payments and External Trade statistics;
f. Gross Domestic Product (GDP);
g. Supply and Use Table (SUT);
h. Input-Output table
i. Secondary data from Government account; and
j. Administrative records from other government agencies.

As aforementioned approach in the ICTSA, the purpose of HSA is to analyse of all aspects of demand and supply of goods and services health industries and the interaction within the economy and the rest of the world. The possible output of the proposed health satellite account tables are:

a. Value added of health-related industries and its contribution to the economy;
b. Total supply-use of health products;
c. Domestics output of health products;
d. Imports and exports of health products;
e. Compensation of employee by health-related industries; and
f. Employment in health-related industries.

3.3. Malaysia's HSA framework

Framework for Malaysia's HSA still under the purview and need to refine specifically on the classification and the availability of data. As mentioned earlier HSA compilation is through the supply and use approach. Therefore, initial phase of Malaysia's HSA have been developed as outlined in the **Appendix C** which was based on the Eurostat Manual of Supply, Use an Input-Output Tables (2008).

4. CONCLUSION

This is the first attempt of developing Health Satellite Account in Malaysia and DOSM has undertaken the role to compile HSA for the nation. The HSA not only requires detailed data, it also entails skills in using the data to describe the system according to the satellite account requirements. As it is a new product, besides looking at availability of data sources, technical expertise and knowledge in health area is also essential for a good analysis.

In order for the HSA to be access by public, in depth study and research needs to done and a technical working group and inter agency committee need to be set up to evaluate the outcome of HSA. We hope that the developing of HSA will improvise our understanding of the health care sector and will be a turning point for policy makers and the public on the growing importance of health related industries in this region.

REFERENCES

Cogswell, Heather, Tesfaye Dereje (2015), Understanding Health Accounts: A Primer for Policymakers. Bethesda, MD: Health Finance & Governance project, Abt Associates

DESA (2016), The Sustainable Development Goals Report 2016, UN. DOI: http://dx.doi.org/10.18356/3405d09f-en

Eurostat (2008), Eurostat Manual of Supply, Use an Input-Output Tables, Luxembourg.

Nakhimovsky, S., Patricia Hernandez-Peña, Cornelius van Mosseveld, Alain Palacios (2014), System of Health Accounts (2011) and Health Satellite Accounts (2005): Comparison of Approaches, Newsletters USAID Health Finance & Governance project. Abt Associates Inc. Bethesda

OECD, Eurostat, WHO (2011) A System of Health Accounts, 2011 Edition. OECD Publishing. Paris.

OECD (2001), The Well-Being of Nations: The Role of Human and Social Capital. OECD Publishing. Paris

PAHO (2005), Pan American Health Organization. Satellite Health Accounts Manual, Version 1, Health Policies and System Unit (HP), Washington D.C.

Schneider, M., Krauss, T., Hofmann, U (2015), National Health Account for Germany, Summary of the research project of the Federal Ministry for Economic Affairs and Energy.

BASYS. Berlin United Nations (2009), System of National Account 2008, New York

APPENDIX A
CLASSIFICATION OF ECONOMIC ACTIVITIES OF THE HEALTH SATELLITE ACCOUNT

Presented below are descriptions of each of the economic activities included in the HSA, based on the International Standard Industrial Classification of All Economic Activities (ISIC).

Characteristic Activities

8412 Regulation of the activities of agencies that provide health care, except for social security services
8430 Compulsory social security activities
8610 Hospital activities
8620 Medical and dental practice activities
8690 Other human health activities

Related activities

2100 Manufacture of pharmaceuticals, medicinal chemical and botanical products
3250 Manufacture of medical and dental instruments and supplies
2651 Manufacture of measuring, testing, navigating and control equipment
2670 Manufacture of optical instruments and photographic equipment
4642 Wholesale of pharmaceutical goods and toiletries
4772 Retail sale of pharmaceutical and medicinal goods, cosmetic and toilet articles
4773 Other retail sale in specialized stores: optical and precision equipment
6511 Life insurance
7210 Research and experimental development in the field of the natural sciences and engineering
7120 Technical tests and analysis
8530 Higher education for health
8550 Adult and other education for health
9601 Delivery of specialized laundry services

APPENDIX B
CLASSIFICATION OF PRODUCT

Below is the HSA classification of products, based on the Central Product Classification (CPC):

Characteristic Products

9112 Administrative services of agencies that provide educational, health care, cultural and other social services, excluding social security services
  91122 Administrative health care services
9131 Administrative services for illness, maternity, or temporary disablement benefit schemes
931 Human health services
9311 Inpatient services
9312 Medical and dental services
  93121 General medical services
  93122 Specialized medical services
  93123 Dental services
9319 Other human health services
  93191 Services provided by midwives, nurses, physiotherapists, and paramedical
  93192 Ambulance services
  93193 Residential health facilities services other than hospital services
  93199 Other human health services n.e.c.

Related Products

342 Basic inorganic chemicals n.e.c.
352 Pharmaceutical products
481 Medical and surgical equipment and orthopedic appliances
  4831 Optical instruments
  53129 Other non-residential buildings
7132 Accident and health insurance services

APPENDIX C

Figure 1. A Framework of Health Supply Table

No.	Supply products	Output of Industries at Basic Prices*				Imports	Total Supply at Basic Prices	Trade and Transport Margin	Taxes less Subsidies on Products	Total Supply at Purchaser's' Prices**
		Industry			Total					
		A	B	...	-1	-2	(3)=(1)+(2)	-4	-5	(6)=(3)+(4)+(5)
1	Manufacture of chemicals and chemical product	Date Source: ECONOMIC CENSUS/ ANNUAL SURVEY/ ADMINISTRATIVE RECORDS FROM HEALTHCARE			SUMMATION OF OUTPUT BY INDUSTRY AT BASIC PRICES	Date Source: TRADE STATISTICS		SUT	Date Source: GOVERNMENT ACCOUNT	
2	Manufacture of basic pharmaceutical products and pharmaceutical preparations									
3	Other manufacturing									
4	Wholesale trade, except of motor vehicles and motorcycle									
5	Retail trade, except of motor vehicles and motorcycles									
6	Insurance/ takaful, reinsurance/ takaful and pension findings, except compulsory social security									
7	Architectural and engineering activities; technical testing and analysis									
8	Public administration and defence; compulsory social security									
9	Human health activities									
10	Residential care activities									
11	Social work activities without accommodation									
	Total Supply (Healthcare products)									

Notes:

(*) Basic prices is the price received by the producer for a unit of good and service produced as output, excluding any tax payable or including any subsidy receivable on the product as a subsequent of its sales or use. It also excludes any delivery charges invoiced separately by the producer.

(**) Purchasers' prices is the price paid by the purchaser to take delivery of a good and service at the time and place required by the purchaser. It includes any transport charges paid separately by the purchaser.

APPENDIX C

Figure 1. A framework of Health Supply Table

No.	Supply products	Intermediate use by Industry				Final Consumption Expenditure	Gross Capital Formation	Exports	Total Use at Purchaser's Prices**
		Industry			Total				
		A	B	...	(1)	(2)	(3)=(1)+(2)	(4)	(5)=(1)+(2) + (3)+(4)
1	Manufacture of chemicals and chemical product	Date Source: ECONOMIC CENSUS/ ANNUAL SURVEY/ ADMINISTRATIVE RECORDS FROM HEALTHCARE			SUMMATION OF INTERMEDIATE USE BY INDUSTRY AT PURCHASERS' PRICES	Date Source: HES/ GOVERNMENT ACCOUNT	Date Source: ECONOMIC CENSUS/ ANNUAL SURVEY/ ADMINISTRA TIVE RECORDS FROM HEALTHCARE	TRADE STATISTICS	
2	Manufacture of basic pharmaceutical products and pharmaceutical preparations								
3	Other manufacturing								
4	Repair and installation of machinery and equipment								
5	Wholesale trade, except of motor vehicles and motorcycle								
6	Retail trade, except of motor vehicles and motorcycles								
7	Insurance/ takaful, reinsurance/ takaful and pension findings, except compulsory social security								
8	Architectural and engineering activities; technical testing and analysis								
9	Public administration and defence; compulsory social security								
10	Human health activities								
11	Residential care activities								
12	Social work activities without accommodation								
	Compensation of Employees	Data source: CONOMIC CENSUS/LABOUR FORCE SURVEY							
	Gross Operating Surplus	VA - COE (derived)							
	Other Taxes less Subsidies on Production	Data Source: GOVERNMENT ACCOUNT							
	Industry Output at Basic Prices								

CPS28:
Sharia Financial and Economic Statistics

Corporate Demand Survey on *Musyarakah* and *Mudarabah* Financing in Malaysia

Hamim Syahrum Ahmad Mohktar
Bank Negara Malaysia, Kuala Lumpur, Malaysia – hamim@bnm.gov.my
Zuraeda Ibrahim
Universiti Teknologi MARA (UiTM), Selangor, Malaysia – zurae229@salam.uitm.edu.my
Zafiruddin Baharum
Universiti Teknologi MARA (UiTM), Selangor, Malaysia – zbaharum@gmail.com
Shariza Abdul Ghani
Bank Negara Malaysia, Kuala Lumpur, Malaysia – shariza@bnm.gov.my
Azren Rizuani Aziz*
Bank Negara Malaysia, Kuala Lumpur, Malaysia – azren@bnm.gov.my

Abstract

One of the key propositions of Islamic finance is derived from its risk-sharing practice, which is embedded in the application of equity-based contracts such as *musyarakah* and *mudarabah*. Such risk-sharing practice strengthens the link between the finance and the real economy through its strong incentives for both parties; the financial provider and the entrepreneur in ensuring success of the underlying business activities. Issuance of the Guidelines on *Musharakah* and *Mudarabah* contracts for Islamic Banking Institutions by the Central Bank of Malaysia in 2007, to a certain extent, has become a catalyst for the actual application of these equity-based contracts in the Islamic banking industry. Clarity in the regulatory requirements that govern the Islamic banking institutions' exposures under *musyarakah* and *mudarabah* contracts has instilled confidence among the industry players and resulted in the double-digit compounded annual growth rate (CAGR) of 45.9% in the total outstanding *musyarakah* and *mudarabah* exposures from RM 4.2 billion in 2010 to RM 40.8 billion in 2016. Understanding on the demand-side of the industry provides customer-driven insights that can be considered in promoting the risk-sharing practice. A survey has been conducted to gauge the potential demand by business entities and corporate bodies in Malaysia for financing product that is structured based on *musyarakah* and *mudarabah* contracts. The questionnaires were also designed to analyse the respondents' behaviour towards these contracts as well as to understand the correlation between the respondents' acceptance of Islamic financial product and their preferences on *musyarakah* and *mudarabah* financing. Respondents to the survey were represented by different economic sectors such as manufacturing, wholesale and communication as well as different types of organisation such as government-linked companies (GLC), public-listed companies, partnership and sole proprietorship. The survey findings conclude that despite the low level of awareness among the respondents on *musyarakah* and *mudarabah* financing, majority indicate preference towards these contracts over the generic loan transaction. This preference is driven by the respondent's ability to diversify business risk and manages the high funding cost in the increasing interest rate environment. Also, the size and type of companies are observed to have an influence the companies' motivation to participate in such financing transaction. Small companies such as partnership and sole proprietorship place greater emphasis on the fair access to financing. While, large companies such as GLC appreciate the opportunity to share expertise and experience with the financial provider. In conclusion, this paper intends to facilitate the Islamic banking institutions in structuring *musyarakah* and *mudarabah* financing products by identifying the demographic profile of potential corporate customers and understanding their financial requirements and preferences.

Keywords: Risk sharing, Partnerships, Customer's behaviour, Islamic finance.

1. INTRODUCTION

Establishment of the first Islamic banking institution (IBI) in Malaysia, Bank Islam Malaysia Berhad (BIMB) in 1983 has enabled financial consumers to obtain financial products and services that comply with Shariah. In order to ensure comprehensive range of Shariah-compliant financial products and services are offered in the banking system, the Malaysian government and the Central Bank of Malaysia has developed relevant infrastructures such as Islamic interbank money market and Islamic capital market to facilitate efficient operation of the IBI. Several high-impact initiatives were introduced with an objective to increase the number and the diversity of industry players.

Interest-free Banking Scheme that allowed conventional bank to offer Shariah compliant banking products and services under the Islamic window concept. This scheme was introduced in 1993 and resulted in a huge success since financial consumers across the country can immediately have access to Shariah compliant banking products and services with a minimal establishment cost; and

Establishment of locally incorporated foreign Islamic banks such as Kuwait Finance House (KFH) and Al Rajhi Bank intended to provide broader options of Islamic banking products and services.

Islamic banking industry in Malaysia has progressed significantly over the past thirty years with total assets grew to RM581.6 billion[1] as at the December 2016. Financing is the largest component in the industry's total assets and continues to grow to RM428.1 billion, which represents 28.6%[1] of total loans in the banking system. Islamic banking financing products are structured based on various Shariah contracts such as sale-based contracts (*murabahah, bai bithaman ajil* and *istisna'*), lease-based contracts (*ijarah* and *ijarah thumma al-bai*) and equity-based contracts (*musyarakah* and *mudarabah*).

2. PROBLEM STATEMENT

Unlike conventional banks, Islamic banks have flexibility in structuring their financing product based on the different Shariah contracts, which would result in the different type of relationship between an Islamic bank and a customer as well as varying product features and risk profile. Nonetheless, as at December 2016, financing products that are mainly structured based on the sale-based contracts have accounted for 53.1% of the industry's total outstanding amount. From Islamic banks' point of view (supply-side), such concentration is due to the following reasons:

Familiar feature and risk profile of the financing products that are structured based on the sale-based contracts, which are akin to conventional financing products. Therefore, offering of the sale-based contract financing can easily leverage on the existing system, documentation and operational procedures; and

Low and insignificant demand for the different Shariah contracts, especially the equity-based contracts potentially due to the risk-sharing feature.

As a result, business entities and corporate bodies are currently offered with a limited choice of financing products by the Islamic banks. Potential benefits of the risk-sharing financing products are not fully explored by both sides: the Islamic banks and the business entities.

3. RESEARCH OBJECTIVES

Objectives of this study are as the following:

To assess the level of interest among the business entities and the corporate bodies towards the *musyarakah* and *mudarabah* financing products;

To understand the key factors that influence the business entities and the corporate bodies in deciding suitable financing products; and

To verify the potential reasons of the business entities and the corporate bodies to shy away from the *musyarakah* and *mudarabah* financing products.

4. DEFINITION & APPLICATION OF *MUSYARAKAH* AND *MUDARABAH* CONTRACTS

Differences between *musyarakah* and *mudarabah*

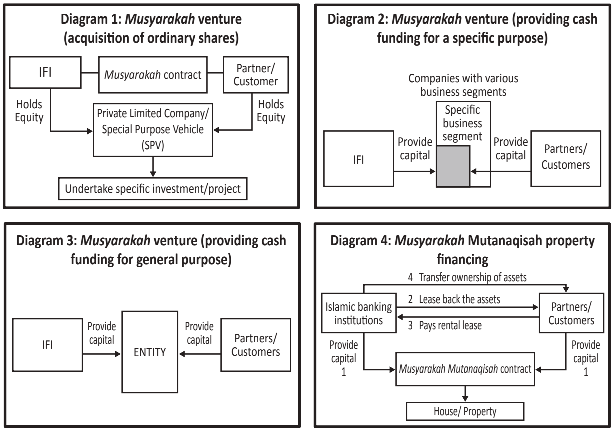
According to a policy document on *Musyarakah* and *Mudarabah* that are issued by the Central Bank of Malaysia in April 2015, *musyarakah* is defined as a partnership contract between two or more parties, where all parties contribute capital and agree to share the profit as well as bear the loss arising from

1 This data excludes Islamic banking assets of the development financial institutions (DFIs).

the partnership. While, *mudarabah* is defined as a contract between a capital provider (*rabbul mal*) and an entrepreneur (*mudarib*), where only the capital provider provides capital for the entrepreneur to manage and any profit arising from the partnership will be shared between both parties based on the pre-agreed profit sharing ratio (PSR). Financial losses are fully borne by the capital provider, unless such losses are due to the entrepreneur's misconduct (*ta'addi*), negligence (*taqsir*) or breach of specified terms (*mukhalafah al-shurut*).

Application of *musyarakah* and *mudarabah* contracts in Malaysia and other countries

In Malaysia, *musyarakah* contract is used to structure property financing (specifically for property under construction) as well as in the Islamic banks' investment activities (*musyarakah* venture). Under the *musyarakah* venture, Islamic banks may provide capital, either in the form of acquiring ordinary shares of the partnership or providing cash funding. Several Islamic banks use *musyarakah mutanaqisah* as the underlying Shariah contract to finance the acquisition of property under construction. Customer and Islamic bank jointly acquire and own the property. After the construction is completed, the customer will rent the property from the Islamic banks and gradually acquire ownership portion of the Islamic bank. The partnership ends once the property ownership has been fully redeemed by the customer. Application of *mudarabah* contract in Malaysia is insignificant and very limited to project financing, specifically for projects that secure direct funding or guarantee from the federal government. Such scenario could potentially due to the higher risk exposure assumed by the Islamic bank as a capital provider.



Successful application of *musyarakah* financing by Sudanese Islamic Bank (SIB) focuses on trade, industrial and agriculture sector. SIB provides capital to eligible farmer to acquire fixed assets such as tractors, water pumps, seeds and fertilizer, while the farmer contribute his land, labour and expertise. An evaluation survey of such financing scheme concluded that annual income of the farmers has increased, which potentially due to inherent motivation to enhance earnings that are now directly linked to performance of underlying economic activities. Based on such premise, some argue that risk-sharing financing product can be a powerful tool to spur growth and development of riskier sectors, smaller enterprises and hence the economy as a whole. A study by Sadr and Iqbal (2002) observed that investment in supervising and monitoring of *musyarakah* venture by Agricultural Bank of Iran (ABI) has improved its recovery rate (loan repayment). Although supervision and monitoring incur additional costs to ABI, but such costs were offset by the following benefits:

Effectively minimises moral hazards and adverse selection problems;

Obtains greater knowledge on market condition and environment; and

Ability to ensure optimal allocation of ABI's resources.

5. RESEARCH METHODOLOGY

A quantitative research approach was chosen to fulfil the objectives of the study. A questionnaire was designed to ensure its ability to elicit accurate data and information. The questionnaire is drafted in English and *Bahasa Malaysia* and has the following three sections:

Section 1 assesses respondents' acceptance of financing products that are offered by Islamic banks. This section consists of statements that can be answered by selecting more than one response as well as the "yes" or "no" questions;

Section 2 investigates the key characteristics and factors that influence respondents' acceptance of Shariah-compliant products. This section consists of statements, where respondents indicate their degree of importance (using a 4-point Likert scale ranging from 1 (least important) to 4 (very important); and

Section 3 analyses behaviour of respondents towards *musyarakah* and *mudarabah* financing. This section consists of the "yes" or "no" questions as well as statements where respondents indicate their degree of importance (using a 4-point Likert scale ranging from 1 (least important) to 4 (very important).

This study managed to obtain 231 responses; nonetheless only 178 respondents answered all the questions and are therefore usable for the study.

6. DATA ANALYSIS AND FINDINGS

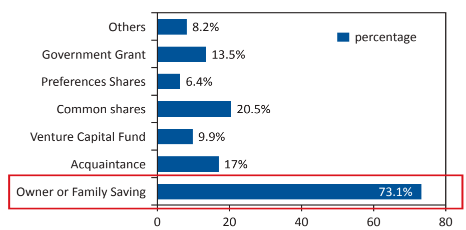
Current acceptance level of Shariah-compliant financing product

Source of capital for business entities during the initial years of business operation mainly comes from owner or family's savings (73.1%), shareholders (20.5%) and acquaintance or business colleague (17%). In terms of borrowing from banking institution, it is observed that majority rely on revolving credit facilities (44.9%), term financing (42.2%) and secured financing (34.7%) as a source of working capital.

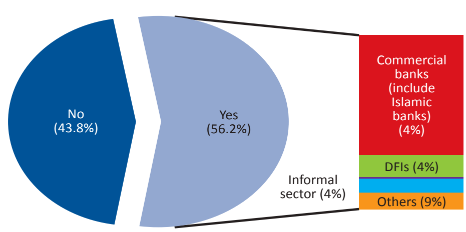
Respondents have fair access to business loan and financing products offered by banking institutions. Amount of financing obtained was fairly small, where 35% of total respondents received less than RM500,000 and only 16% obtained financing more than RM100 millions. This could be explained by majority of respondents have small business and fairly young in the industry, which limit their capacity to borrow from banking institution.

Consistent with the market share of total Shariah compliant financing, 25% of total respondents obtained financing from Islamic banks. It is observed that Islamic banks offered small number of large financing (more than RM100 millions). 55.6% of the listed companies borrow more than RM100 million, but only 7.7% were structured based on Shariah principles.

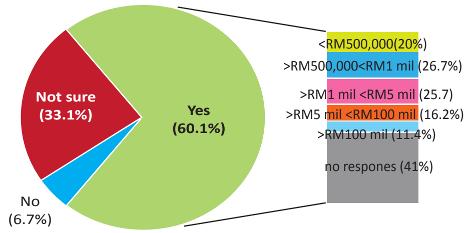
60% of total respondents indicate interest to apply for Shariah-compliant financing in the near future. Significant portion of this potential demand (46.7%) aims for financing of less than RM1 million. The potential demand mainly



Graph 1. What is Your Institution's Source of Capital During Start-up?



Graph 2. Has Your Institution Obtained any Business Loan/ Financing within the Last 5 Years? From which institution is the Loan/Financing Obtained?

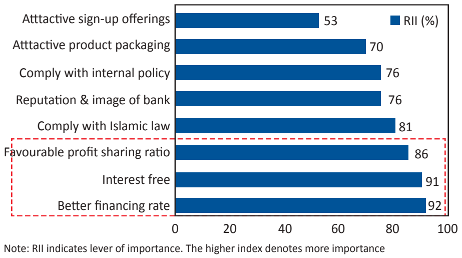


Graph 3. Does Your Institution Have Any Plan to Get Financing from IsIamic Bank in the Near Future? If yes, Please Select the Required Amount?

comes from Malaysia's key economic sectors such as manufacturing (13.3%), wholesale (10.5%) and communication (10.5%).

Financing cost remains as the key factor that influence financial decision making of the business entities and corporate bodies in Malaysia.

Compliance to Shariah requirements is secondary while attractive product packaging or sign-up offerings scored the lowest relative important index (RII)



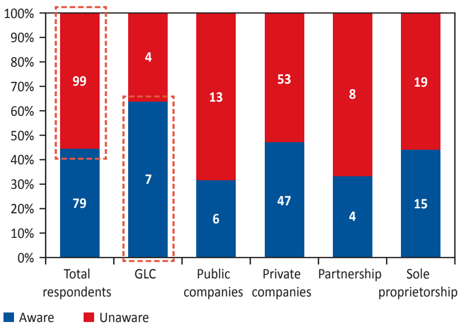
Graph 4. Please Indicate the Degree of Importance of the Following Factors in Influencing Your Institution to Choose Financing Products from an Islamic Bank?

Awareness on *musyarakah* and *mudarabah* financing

Respondents are fairly not aware of *musyarakah* and *mudarabah* financing. Only 44.4% of total respondent are aware of such Islamic banking offering. Government-linked companies (GLC) recorded the highest level of awareness. Following is the demographic details of respondents who are unaware of *musyarakah* and *mudarabah* financing:

Companies with large total assets (more than RM100 millions)

Companies that are listed in Bursa



Graph 5. Are You Aware of the Concept of Partnership Arrangement in Islamic Banking?

Nonetheless, based on the given description of *musyarakah* and *mudarabah* financing, majority indicated preference on the *musyarakah* and *mudarabah* financing (87.1%) over the conventional loan. Following is the demographic details of respondents who indicated preference on the *musyarakah* and *mudarabah* financing:

Partnership, private companies and sole proprietorship

Small size companies (total assets less than RM2.5 millions

Manufacturing, wholesale and communication sectors

Key attractive features of *musyarakah* and *mudarabah* financing

In general, respondents highlight that potential attraction of *musyarakah* and *mudarabah* financing are as follows:

The risk-sharing feature that allow respondents to diversify business risk (especially during the unfavourable economic environment) and mitigate high funding cost in the event where interest rate increases.

Stronger commitment and motivation from both parties (capital provider and entrepreneur) to ensure business success.

Table 1. Key Attractive Features of Musyarakah and Mudarabah Financing

Rank	Key Feature	RII (%)
1	Motivasi business expansion	88.2
2	Provides business support to management at times of economic depression	87.9
3	Both capital provider and entrepreneur would be committed to business success	87.8
4	Cost of funding is not exposed to movement of interest rate	87.5
5	Improves business efficiency	87.4

Respondents do not perceive *musyarakah* and *mudarabah* financing mostly suit for start-up business entities with small capital and risky business

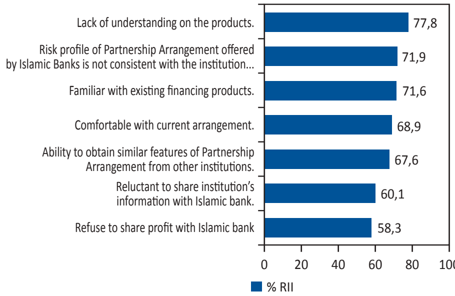
It is also observed that different type of business entities identify different key attractive features. Small companies such as private companies, partnership and sole proprietorship emphasise greater importance on the fair access to financing. While large companies such as GLC appreciate the opportunity to share expertise and experience with Islamic banks.

Table 2. The Comparison Result

Rank	Companies with small total assets (<RM 2.5 millions)	Companies with large total assets (>RM 100 millions)
1	Allows small business to obtain funds on a fair basis	Both capital provider and entrepreneur would be committed to business success
2	Moivates business expansion	Cost of funding is not exposed to movement of interest rate
3	Does not put heavy burden on entrepreneur at times of business slow down	Motivates business expansion
4	Provides business support to management at times of economic depression	Allows sharing of expertise and experience between capital provider and business owner
5	Both capital provider and entrepreneur would be committed to business success	Higher profit return as compared to conventional loan

Main factors that discourage participation in *musyarakah* and *mudarabah* financing

Lack of understanding on the *musyarakah* and *mudarabah* financing is the main reason for respondents to shy away from such products. In addition to the lack of understanding, the listed companies highlight that alternative financial product with similar features of *musyarakah* and *mudarabah* financing is available and currently offered by other institutions such as venture capitalist. "Reluctant to share companies' information with Islamic bank" and "Refuse to share profit with Islamic bank" scored the lowest RII, hence are not the key reason of refusal.



Graph 6. Main Factors That Discourage Participation in Musyarakah and Mudarabah Financing

7. CONCLUSIONS

In order to optimally benefit from the potential demand by business entities and corporates for Islamic financing products, Islamic banks must not only enhance customers' awareness on *musharakah* and *mudarabah* financing but also to ensure the product structuring is driven by customers' specific needs

and requirement. For instance, credit assessment of Islamic banks on financing application from private companies, partnership and sole proprietorship must be able to consider the potential of business activities in assessing the application's viability (in addition to creditworthiness of the applicant). For GLC, Islamic banks are expected to offer value-adding services such as provide business and financial advices (cash management). In addition, pricing of *musharakah* and *mudarabah* financing must be competitive and able to reflect performance of underlying business activities. This is a critical success factor since majority respondents consider financing cost as the main factor of consideration in choosing financing product.

From regulator's point of view, capability of Islamic banks to effectively supervise and monitor the partnership must be enhanced so that value propositions of the *musharakah* and *mudarabah* financing can be explored, without exposing the banking system to instability.

REFERENCES

Al-Harran, Saad A. S. (1993). Islamic finance: Partnership financing. Pelanduk Publications (M), Petaling Jaya

Sadr, Kazem & Iqbal, Zamer. (2002). Choice between debt and equity contracts and asymmetrical information: Some empirical evidence. In Munawwar Iqbal and David T. Llewellyn (eds.). Islamic banking and finance: New perspectives on profit-sharing and risk. UK: Edward Elgar.

Table 1.7 and 1.7.1 of the Monthly Statistical Bulletin December 2016, Central Bank of Malaysia.

Table 1.18.2 of the Monthly Statistical Bulletin December 2016, Central Bank of Malaysia.

Comparing the Technical Efficiency of Leading Baitul Maal wat Tamwil and Conventional Cooperatives in Indonesia[1]•

Ascarya
Islamic Economic and Finance Department, Bank Indonesia - ascarya@bi.go.id

Abstract

Microfinance institutions, either conventional cooperatives (CMFI) or Islamic cooperatives or BMT (IMFI) play important roles in empowering and developing the poor and micro enterprise in Indonesia by providing various micro financial services for their customer and member. Therefore, they must be sustainable, effective, efficient, affordable, and resulted in welfare impact to alleviate poverty and improve wellbeing. This study is aimed to compare and evaluate the technical efficiency - TE (including pure technical efficiency – PTE and scale efficiency - SE) of some leading conventional cooperatives (CMFI) and Islamic cooperatives or Baitul Maal wat Tamwil – BMT (IMFI) using Data Envelopment Analysis (DEA).

The results suggest that the efficiency of CMFI has been very impressive, especially in 2010 and 2011, but it has decreased in 2012 mainly due to the decline in PTE, while the efficiency of IMFI has been lower than CMFI in early years, but consistent increase in PTE and has made IMFI reached higher efficiency than CMFI in 2012. SE of CMFI has been stable at high level, while SE of IMFI has also been stable at high level and slightly increase in 2012. CMFI has started to decline, while IMFI has reached its optimum efficiency in 2012. The common problems of CMFI and IMFI are mostly in funding and equity as their main source of fund to extend micro-financing, including low in deposit taking, too much dependent on bank financing and low in member' equity (principal deposits and compulsory deposits). To resolve the funding problem, there should be national campaign "Let's go to Cooperatives" or "Let's go to BMT" and to make them community micro-bank. BMT could become cash-waqf manager to improve its equity and get rid of bank financing, while cash-waqf could also decrease mismatch and improve financial soundness of BMT.

Keywords: Microfinance, Islamic Microfinance, Cooperative, Baitul Maal wat Tamwil, Efficiency, Technical Efficiency

JEL Classifications: G210, G280, D610

1. INTRODUCTION

Micro enterprise (ME) in Indonesia, which summed to 55.9 million or 98.8% of all business entities, have absorbed 90.1% of employment in 2012 (Ministry of Cooperatives and SME data). The need of financing for micro enterprise has been served mostly by microfinance institution (MFI). Conventional microfinance institution (CMFI) is legally registered as conventional cooperatives, while Islamic microfinance institution (IMFI), which is known as Baitul Maal wa Tamwil (BMT), is registered as Islamic cooperatives (Koperasi Simpan Pinjam Pembiayaan Syariah or KSPPS).

BMT is IMFI combining microfinance and Islamic finance, therefore it has characteristic and basic model which is like mainstream MFI or cooperatives, but having mechanism, product, and services which comply to Islamic values, such as prioritizing justice and equity. Another distinct characteristic embedded to BMT including empowerment of customer (bottom of the triangle or the poorest of the poor) by providing support to run new ME in order to improve their living standard and prosperity. BMT is one type of approach which aims to alleviate poverty and to narrow the gap between the rich and the poor.

Considering the large portion of micro enterprise and high poverty rate in Indonesia, which approximately 40% of all people remain clustered around the national poverty line set at Rp330,776 or $22.6 per person per month in 2015 (World Bank data), the role of CMFI or conventional cooperatives

1 • The authors would like to thank Adinda R. N. Illiyyin as research assistant.

and IMFI or BMT as Islamic cooperatives are very important to provide variety of financial services in a much smaller scale. Cooperatives as MFI (conventional as well as Islamic) has to provide product and service which are in line with the need of the poor and ME member. To deliver excellent service while empowering the poor, cooperatives have to perform in effective and efficient way, as well as sustainable with the appropriate business model.

This study aims to determine and analyze the technical efficiency of major conventional cooperatives and Islamic cooperatives or BMTs in Indonesia by using Data Envelopment Analysis (DEA) to understand their existing conditions, their weaknesses, so that it can be used to improve their efficiencies in the future.

2. LITERATURE REVIEW

2.1. Microfinance

Microfinance can be defined as provision of financial products and/or services (such as, micro-credit, micro-savings, micro-equity, micro-transfers and micro-insurance) in sustained manner to the poor, marginalized people and/or low-income people whose low economic standing excludes them from formal financial systems. The main difference between microfinance and mainstream finance systems is its alternative approach to collateral that comes from the concept of Group Lending under Joint Responsibility (GLJR), since the poor do not have necessary asset for collateral, as well as personal approach with Pick-up Services or *banker meets customer*.

Microfinance products and services are usually provided by microfinance institution (MFI), conventional (CMFI) or Islamic (IMFI), which includes a wide range of providers that vary in their legal structure, mission, and methodology. However, all share the common characteristic of providing financial services to clients who are poorer and more vulnerable than traditional bank clients. An MFI can operate in the form of a commercial bank, rural bank, credit union, cooperatives, other non-bank financial institution (NBFI), development organization, or non-government organization (NGO).

Obaidullah (2008) confirmed that good microfinance programs were characterized by small, usually short-term loans; streamlined, simplified borrower and investment appraisal; quick disbursement of repeat loans after timely repayment; and convenient location and timing of services. Meanwhile, Islamic microfinance is one of Islamic approaches to alleviate and eradicate poverty, so that the main target is not only the poor, but more importantly is the poorest of the poor, which has always been left out by mainstream microfinance.

In Islam, poverty is in conflict with one of the primary objectives (maqasid) of Shariah, namely, "enrichment of self (nafs)". Moreover, Islamic jurists have unanimously held the view that it is the collective obligation (fard kifayah) of a Muslim society to take care of the basic needs of the poor. According to Obaidullah (2008), principles of Islamic approach to poverty alleviation include: a) charity; b) economic empowerment; c) debt avoidance; d) cooperation and solidarity; e) family cohesiveness; f) Shari'ah compliance of contracts; and g) Islamic norms and microfinance best practices.

2.2. Microfinance Models

Common models of microfinance in general can be distinguished into several types, namely: 1) Association model; 2) Bank Guarantee model; 3) Community Banking model; 4) Cooperative model; 5) Credit Union model; 6) Grameen Bank model; 7) Group model; 8) Individual model; 9) Intermediary model; 10) Non-Government Organization (NGO) model; 11) Peer Pressure model; 12) Rotating Saving and Credit Association (ROSCA) model; 13) Small Business model; 14) Village Banking model; 15) Self Help Group model; 16) Graduation model; and 17) Micro-banking Unit model. However, four microfinance models are stand out, namely Grameen Bank model, Village Bank model, Cooperative or Credit Union model, and Self-Help Group model (Obaidullah, 2008). Other important microfinance models are Graduation model introduced by Bangladesh Rural Advancement Committee (BRAC model), Grameen La Riba model, and Social Islami Bank Limited (SIBL) model introduced by MA. Mannan (2007). Among various microfinance models, some of them have been successfully implemented in Indonesia, namely: 1) Grameen bank model, represented by various non-government organizations or NGOs, cooperatives, Baitul Maal wa Tamwils or BMTs, or venture capitals; 2) Cooperative model, represented by cooperatives and BMTs; and 3) Community banking model, represented by conventional and Islamic rural banks; and 4) Micro-banking Unit, represented by conventional and Islamic commercial banks which establish micro-banking unit. The short description of each model is as follows.

Cooperatives is micro financial institution based on the concept of mutuality. According to Obaidullah (2008), Cooperatives model is in the nature of non-profit financial cooperative owned and controlled by its members. Cooperatives mobilize savings, provide loans for productive and provident purposes and have memberships which are generally based on some common bond. Cooperatives generally relate to an apex body that promotes primary credit unions and provides training while monitoring their financial performance. Credit Unions are quite popular in Asia, notably in Sri Lanka. A variant of Cooperatives is the member-based Islamic financial cooperatives in Indonesia called Baitul Maal wat Tamwil (BMT) which provide charity-based not-for-profit social services and margin-free loan or Qardh and/or Qardh Hasan (Bait ul-Maal function), as well as for-profit financing (Bait ut-Tamwil function).

2.3. Previous Studies

There are not so many studies measuring the technical efficiency of Islamic cooperatives or BMT or Islamic microfinance institution (IMFI), but there are a few studies comparing the efficiency of CMFI and IMFI. Ali and Ascarya (2010) measured and compare technical efficiency of two BMTs with many branches in East Java Indonesia using two-stage DEA and production approach. The results show that both BMT have not reached optimal efficiency, due to financing (for first BMT) and bank financing (for second BMT). Nasution (2014) measured technical efficiency of 14 BMTs in Java Indonesia using DEA and intermediation approach. The results show that BMT with highest efficiency could be due to the ability of BMT to educate its member to manage their wealth (especially their financing).

Abdelkader and Salem (2013) measured and compared technical efficiency of 14 IMFI and 51 CMFI in MENA (Middle East and North Africa) region using DEA and production approach, combining double bottom-line with social goal (outreach) and commercial goal (financial performance) as outputs produced by three inputs (total assets, number of employees and operating expenses). The results show that IMFI scored higher average efficiency in the period of 2005-2007, but lower in 2007-2009 and equal in 2010. However, there is no significant difference between the efficiency of IMFI and CMFI. Islamic approach or the application of Sharia compliant products do not affect the efficiency of MFI.

Mahmood, et al. (2014), measured and compare technical efficiency of 9 CMFI and 3 IMFI in Pakistan using DEA and production approach, with three inputs (number of employee, assets and cost per borrower) and two outputs (gross loan portfolio and number of active borrowers), so that overall technical efficiency of IMFI was higher than that of CMFI. The results show that 2 out of 3 IMFI and only 2 out of 9 CMFI were found to be in the efficient frontier. Moreover, total factor of production (TFP) of IMFI was found higher than that of CMFI. They conclude that IMFIs have been more efficient as compared to CMFIs in Pakistan.

Widiarto and Emrouznejad (2015) measured and compare social and financial efficiency of 232 MFIs (27 IMFIs and 205 CMFIs) from MENA 63, EAP (East Asia and Pacific) 29 and SA (South Asia) 113 using two-stage DEA and production approach, and then tested them to determine factors that contribute to efficiency. The results show that IMFI scored lower financial and social efficiency than those of CMFI. There exist MFIs that managed to relatively pursue outreach and sustainability objectives simultaneously. Thus, instead of focusing on trade-off between these objectives, the focus of MFIs should be on pursuing them simultaneously by emulating MFIs with similar characteristics that have managed to do so.

3. METHODOLOGY

3.1. Data

In line with DEA approach, this study uses annual data from balance sheet, income statement and number of labor in 2007-2012. Input variables for intermediation approach are: 1) Fixed Asset; 2)

Number of labor; 3) Deposits; 4) Bank Financing; and 5) Member's Equity. Output variable are: 1) Financing; 2) Other income. Sample are five major cooperatives and five major BMTs in Indonesia (see table 1).

Table 1: Sample of Cooperatives and BMTs

MODEL	Cooperatives	BMT
Grameen	KOMIDA	BMT ITQAN
	MBK VENTURA	
	SETIA BHAKTI WANITA	
Individual	KODANUA	BMT MARDLOTILLAH
		BMT IBADURRAHMAN
Micro-banking	KOSPIN JASA	KOSPIN JASA SYARIAH
		BMT UGT SIDOGIRI

In the Grameen model, MFI extends financing to member in group called group lending join responsibility (GLJR), therefore loan/financing is guaranteed by the group whose most member are women. In this model, financing activity aims to empower the poor. In individual model, MFI extends financing to individual member, while in micro-banking model, MFI extend financing to individual member and non-member (usually called as candidate of member).

3.2. Operational Definition of Variables

This study applies intermediation approach in calculating the technical efficiency of Islamic and conventional cooperatives, since this is the most appropriate approach to compare efficiency of different institutions. Input variables include deposits, number of labor, fixed asset, bank financing and member's equity, while output variables include financing extended and other income. The definition of each variable can be seen in table 2, while the complete input and output data can be seen in the Appendix 1.

Table 2: Input and Output Variables

INPUT	
X1: Deposits	Total deposits collected by MFI, recorded in the liabilities side of the balance sheet.
X2: Labor	Total number of labor employed by MFI.
X3: Fixed Asset	Total fixed assets of the MFI, recorded in the assets side of the balance sheet.
X4: Bank Financing	Total fund borrowed by MFI, which is financed by Islamic bank.
X5: Member's Equity	Total principal and compulsory deposits of MFI member, recorded as equity.
OUTPUT	
Y1: Financing	Total financing/loan extended by MFI to its members.
Y2: Other Income	Total other income from other micro finance services provided by MFI.

3.3. Methodology

The concept of efficiency rooted from the microeconomic concept, namely, consumer theory and producer theory. Consumer theory tries to maximize utility or satisfaction from individual point of views, while producer theory tries to maximize profit or minimize costs from producer point of views. In the producer theory, there is the 'S' curved production frontier line that describes the relationship between inputs and outputs of production process. This production frontier line represents the maximum output from the use of each input. It also represents the technology used by a business unit or industry. A business unit that operates on the production frontiers is technically efficient.

Performance measurement of financial institution has increasingly focused on frontier efficiency or X-efficiency (rather than scale efficiency), which measures deviation in performance of a financial institution from the best practices or costs-efficient frontier. Frontier efficiency stems from technical efficiency, which gauges the degree of friction and waste in the production processes, and allocative efficiency, which measures the levels of various inputs. X-efficiency is superior for most regulatory and other purposes to the standard financial ratios from accounting statements, such as ROA or cost/revenue ratio. X-efficiency is an objectively determined quantitative measure that removes the effects of market prices and other exogenous factors that influence observed performance.

Efficiency from production theory comprises of two components (Farell, 1957: 1) Technical efficiency (TE) describes the ability of a business unit to maximize output given certain amount of input (efficiency in terms of quantity); and 2) Allocative efficiency (AE) describes the ability of a business unit to utilize inputs in optimal proportion based on their price (efficiency in terms of price). When the two types of efficiency combined, it will produce economic efficiency or cost efficiency or Overall Efficiency (OE). A company is considered to be economically efficient if it can minimize the production costs to produce certain output within common technology level and market price level.

Overall Efficiency (OE) = Allocative Efficiency (AE) x Technical Efficiency (TE). Meanwhile, Technical Efficiency can be broken down into Pure Technical Efficiency (PTE) and scale efficiency (SE), so that Technical Efficiency (TE) = Pure Technical Efficiency (PTE) x Scale Efficiency (SE). Therefore, OE = AE x PTE x SE.

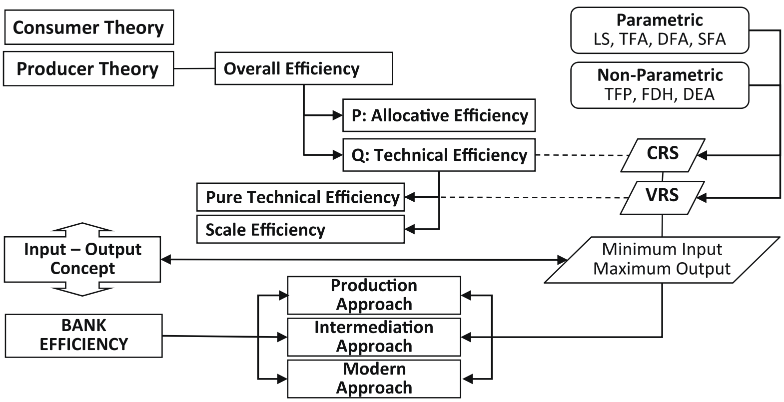


Figure 1. Theory of Efficiency

Efficiency can be measured using parametric approach or non-parametric approach. Parametric approach, such as stochastic frontier approach (SFA) and distribution free approach (DFA), measure efficiency using stochastic econometric and trying to eliminate the impact of disturbance to inefficiency. The advantage is that it allows for random error, so these methods are less likely to misidentify measurement error, transitory differences in cost, or specification error for inefficiency, while the disadvantage is that it has to impose more structure on the shape of the frontier by specifying a functional form for it.

Measure efficiency using non-stochastic approach, such as data envelopment analysis (DEA) tends to combine disturbance into inefficiency. DEA measures the efficiency of a decision making unit (DMU) relative to other similar DMUs with the simple restrictions that all DMUs lie on or below the efficiency frontier. DEA can also determine how a DMU can improve its performance to become efficient. Non-stochastic approach assumes that random errors do not exist and that all deviations from the frontier indicate inefficiency. The advantage is that it does not require an a priori assumption about the analytical form of the production function so imposes very little structure on the shape of the efficient frontier, so that there is no misspecification. The disadvantage is that it is sensitive to extreme observations and measurement error (the basic assumption is that random errors do not exist and that all deviations from the frontier indicate inefficiency), so that there exists a potential problem of "self-identifier" and "near-self-identifier".

DEA is a method for analyzing the relative efficiency and managerial performance of productive or decision making units (DMUs), having the same multiple inputs and multiple outputs. DEA allows us to compare the relative efficiency of banks by determining the efficient banks as benchmark and by measuring the inefficiencies in input combinations (slack variables) in other banks relative to the benchmark. DEA is a non-parametric, deterministic methodology for determining the relative efficient production frontier.

The efficiency measurement, parametric or non-parametric, of financial institution like banks can be approached from their activities. There are three main approaches to explain the relationship between input and output of banks. Two approaches, namely, production (or operational) approach and intermediation approach, apply the classical microeconomic theory of the firm. One approach, namely modern (or assets) approach applies modified classical theory of the firm by incorporating some specificities of banks' activities, namely risk management and information processing, as well as some form of agency problems, which are crucial for explaining the role of financial intermediaries (Freixas and Rochet, 1998).

Production approach describes banking activities as the production of services to depositors and borrowers using all available factors of production, such as labor and physical capital. This approach views financial institutions as a producer of a savings account and credit loans. This approach, initiated by Benston (1965) and Bell and Murphy (1968), considers banks as producer of deposit accounts to depositors and loans to borrowers. Therefore, this approach defines input as number of workforce, capital expenses on fixed assets and other materials, and defines output as the sum of all deposit accounts or other related transactions. This approach suits well for local branch level banks (Freixas and Rochet, 1998).

Intermediation approach describes banking activities as intermediary in charge of transforming the money borrowed from depositors (surplus spending units) into the money lent to borrowers (deficit spending units). In other words, deposits that are typically divisible, liquid, short-term, and risk less are transformed into loans that are typically indivisible, illiquid, long-term, and risky. This approach views financial institutions as intermediaries. These financial institutions transform and transferring financial assets from units with excess funds to units with lack of funds. Therefore, this approach defines input as financial capital (the deposits collected and the funds borrowed), and defines output as the volume of loans and investment outstanding. This approach suits well for main branch level banks (Freixas and Rochet, 1998).

4. RESULTS AND ANALYSIS

4.1. Profiles of BMT

There are five BMTs in the sample, including KOSPIN Jasa Syariah (KOSJs), BMT Ibaadurrahman (IBAAD), BMT Mardlotillah (MARD), BMT UGT (UGT) dan BMT Itqan (ITQAN), with time-series data of 2007-2012 period. Some important indicators of microfinance institution (MFI) include total assets, asset/branch, asset/staff, member's equity, number of staff, number of staff/branch, number of branches, total deposits, total financing, and some other important indicators. Assets and growth of assets of five BMTs represent the size and growth of the BMT. BMT UGT was the largest BMT with total assets Rp662.8 billion in 2012, with impressive average annual growth of 67% in the past five years, given it was just established in year 2000. The rapid growth of BMT UGT was supported by its network of Pesantren Sidogiri alumni who have been spread in many parts of Indonesia. By September of 2016, total assets of BMT UGT has reached Rp2.2 trillion with 277 branches and 762 thousands members.

KOSPIN Jasa Syariah was established in 2004 as a subsidiary of conventional KOSPIN Jasa, where its assets has reached Rp365.0 billion in 2012, with even more impressive average annual growth of 71% in the past five years. The rapid growth of KOSJS has been supported by the network of its parent conventional KOSPIN Jasa. Moreover, BMT Mardlotillah has been started since 1997 from scratch in a remote rural area of Sumedang, West Java, located inside the traditional market. Without any supporting network or social capital, its assets has quadrupled from Rp3.0 billion in 2007 to reach Rp12.5 billion in 2012, with average annual growth of 33% in the past five years.

The oldest BMT in the sample is BMT Ibadurrahman, which has been established since 1995 in suburb area of Ciawi, West Java, located close to traditional market. As one of the pioneers of BMT, IBAD was started from a small group of people conducting '*arisan*' (rotating savings) and financing activity among members (using *qordhul hasan* contract). Its assets has doubled from Rp3.0 billion in 2007 to Rp6.5 billion in 2012, with average annual growth of only 17% in the past five years. The growth of IBAD was lower than its peers, since the expansion strategy of IBAD is to spin-out branch to become independent BMT within the group. The youngest BMT in the sample is BMT Itqan, which was established in 2007, and it applied Islamic Grameen model since 2010. Its assets has jumped from Rp31 million in 2008 to Rp1.6 billion in 2012, especially after it applied Islamic Grameen model, with average annual growth of 123% in the past four years. By September of 2016, the total assets of BMT Itqan has reached Rp35 billion with 8 branches and 12 thousands members.

4.2. Results

DEA method delivers three efficiency measurements: 1) Technical Efficiency-TE; 2) Pure Technical Efficiency-PTE; and 3) Scale Efficiency-SE, where TE is a multiplication of PTE and SE. TE of BMTs in 2007-2009 has been slightly increased to 87.5%, but has been significantly increased in 2010-2012 to reach 98.1% in 2012 (see figure 2, left), which was better than cooperatives. The increase of TE mostly was supported by the increase in PTE. Meanwhile, TE of cooperatives has been increased from 87.1% in 2007 to 99.8% in 2010, and then it has been decreased to reach 95.5% in 2012 (see figure 2, right). Fluctuation of TE in cooperatives has been contributed mostly by PTE fluctuation. The TE level shows that cooperatives have been more efficient than BMT in early years, but by the end of 2012, BMTs have become more efficient than cooperatives.

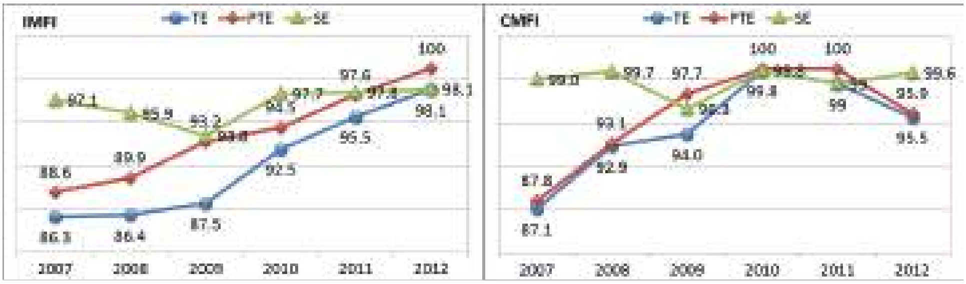


Figure 2. Efficiency of BMT (left) and Conventinal Cooperatives (right)

Technical efficiency of five BMTs on average has been increasing from 96.3% in 2007 to 98.1% in 2012. The best TE performer was KOSJs (KOSPIN Jasa Syariah) with only in 2009 its TE fell to less than 100%, followed by ITQN (BMT ItQan) with also only in 2009 its TE fell less than 100%. The lowest TE of BMT was UGT, which never reached 100% in any year. However, BMT UGT was the largest BMT with assets Rp662.8 billion and 192 branches in 2012, and it has still been expanding impressively (see table 3, left).

Technical efficiency of five Cooperatives on average has been increasing from 87.1% in 2007 to 99.8% in 2010, and then it has been decreasing to reach 95.5% in 2012. The best performer was KSBW (Koperasi Setia Bhakti Wanita) with TE of 100% in all years, followed by KODA (Kodanua) with TE less than 100% in 2010 and 2012. However, Kodanua and KOMD (Koperasi Mitra Dhuafa) should be alerted, since their technical efficiencies decreased to less than 100% in 2012, while other cooperatives reached 100% in 2012 (see table 3, right).

This page intentionally left blank

Friday, 24 March 2017

IPS27: Advances in Statistical Process Control Scheme For Risk-Free Monitoring

Control Charts for Attribute Control Based on Life Distributions with Applications

Amílcar Oliveira*
Universidade Aberta and CEAUL, Lisbon, Portugal – amilcar.oliveira@uab.pt
Teresa A. Oliveira
Universidade Aberta, Lisbon, Portugal – Teresa.oliveira@uab.pt

Abstract

Charts for attribute control will be explored, considering (i) the number of defective items, named np-charts; (ii) means of the proportion (p) of defective items, named p-charts; (iii) u-charts; (iv) c-charts. An update and some recent ideas on the subject based on lifetime distributions will be presented, as well as the advantages and disadvantages of each kind of these charts. Implementations in software R will be discussed through examples in different contexts.

Keywords: Life distributions, Control charts, P-charts, Np-charts, U-charts, C-charts, R software.

Control Charts for Simultaneous Monitoring of Unknown Parameters of a Shifted Exponential Distribution

Zhi Lin Chong*
Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia – chongzl@utar.edu.my
Amitava Mukherjee
XLRI – Xavier School of Management,
Jamshedpur-831001, Jharkhand, India – amitmukh2@yahoo.co.in

Abstract

Shifted exponential distribution is an important model for lifetime data, survival analysis and risk management. In the recent time, Mukherjee et al. (2015) proposed some interesting monitoring schemes for monitoring both the parameters of the shifted exponential distributions. Nevertheless, they assumed that the standards are known a-priori, which is not the case in practice. In the present context we assume that the standards are unknown and need to be estimated from the Phase-I samples. We develop some competitive control schemes to reduce risk involved with estimation bias from Phase-I data. Problems and challenges are outlined along with some recommendations for the practitioners.

Keywords: Control charts; Distance; Max; MLE; Phase-I; Phase-II; Shifted Exponential.

This page intentionally left blank

IPS28:
Financial Modelling and Analytics

Prediction of Personal Bankruptcy Using Data Mining Techniques

Yap Bee Wah*
Advanced Analytics Engineering Centre.
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia beewah@tmsk.uitm.edu.my
Sharifah Heryati Syed Nor
Faculty of Business Management, Universiti Teknologi MARA, Malaysia
heryati879@salam.uitm.edu.my
Shafinar Ismail
Faculty of Business Management, Universiti Teknologi MARA, Malaysia
shafinar@bdrmelaka.uitm.edu.my

Abstract

Personal bankruptcy is on the rise in Malaysia. The Insolvency Department of Malaysia reported that personal bankruptcy has increased since 2007 and the total accumulated personal bankruptcy cases stood at 131,282 in 2014. This issue is very alarming because the increase in the number of personal bankruptcy cases will have a negative impact on the Malaysia economy and the society. This paper evaluates three data mining methods which are logistic regression, decision tree and personal bankruptcy score card in classification of personal bankruptcy. The data was provided by the Debt Management Agency. In this study, personal bankruptcy is defined as distressed borrowers who failed to settle their loans. The sample comprises 24566 cases with 17% non-distressed and 83% distressed cases. The data includes a dependent variable, personal bankruptcy (Y=1(distressed), Y=0(non-distressed)) and 12 predictors. The sample data was partitioned into 70% for training and 30% for validation of the model. Model comparison results show that the validation classification accuracy for logistic regression (83.2%), decision tree (83.3%) and scorecard (82.2%) were quite similar. However, the precision was low and sensitivity was high for all three models. The sensitivity for scorecard model was (100%) with specificity 0% indicating that the credit scorecard model was affected by the imbalanced data. We then performed undersampling by selecting 4174 cases randomly from the 20372 terminated cases and evaluated the models again using the balanced sample of 8348 cases. The classification accuracy (LR=70.7%; DT=70.9%; SC=69.8%). sensitivity (LR=64.7%; DT=67.0%; SC=60.7%) and specificity ((LR=76.7%; DT=74.8%; SC=78.9%) indicate that decision tree is the best predictive model. The significant predictors based on the logistic regression model were age, race, gender, marital status, number of children, employment status, monthly income, location of residence, loan outstanding and number of loans. The models were developed using SAS Enterprise Miner 7.1. Predictive models can be deployed by relevant bodies to predict the likelihood of an individual who will become bankrupt due to financial difficulties.

Keywords: Personal bankruptcy; Data mining; Logistic regression; Decision tree, Credit scoring.

IPS29: Government Finance Statistics

Comparative Study of Government Finance Statistics Compilation and Utilization in Indonesia, Malaysia and Australia

Dr. Mei Ling *
Deputy Director of Statistics and Analysis of Financial Reports
Directorate General of Treasury, Ministry of Finance Republic of Indonesia
elinetj@gmail.com

Abstract

The fast pace of globalization has stimulated the need for a common language in form of international standards for public sector accounting and statistics, which will be crucial for stakeholder's decision making and country comparison. Similar to International Public Sector Accounting Standards (IPSAS) that are designed for use in preparing general-purpose financial statements, Government Finance Statistics (GFS) set uniform guidelines as an international standard for public sector statistics that has been compiled by countries worldwide. Many countries have already implemented GFS, and more have plans to implement GFS in the future. Both IPSAS and GFS enable comparison of data across countries and improve transparency. As the move to international standards continues, countries are starting to come to terms with the challenges involved. Australia and ASEAN countries such as Indonesia and Malaysia are three of the countries that compile and utilize GFS. Comparison study of those countries experience in the compilation and utilization of GFS, particularly in identifying challenges and finding solution to overcome them can become lessons learned for other countries in implementing GFS.

Keywords : Government finance statistics; Public sector accounting; Utilization.

CPS29:
Demography & Social Welfare Statistics (5)

The Multilevel Logistic Regression Analysis on Status of Migration of Population in Jawa Barat Province 2015

Budyanra*
Institute of Statistics, Jakarta, Indonesia
budy@stis.ac.id
Athiyyah Nafsu'l Muthmainnah
BPS-Statistics of Indonesia
12.7054@stis.ac.id

Abstract

Research in the social field usually involves a problem that connects the individual with the environment or group. The concept generally individuals interact with the surrounding social groups, so that individuals are influenced by social group, and the group is also affected by individuals who are in the group. Hierarchies data arise because individuals who nested within social groups respectively. For hierarchies data, the analysis of multilevel is very suitable to be applied and will certainly generate an appropriate model in seeing relationships between variables. Multilevel logistic regression can describe the relationship and the tendency of the influence of the explanatory variables in the form of individual variables and contextual variables affect the response variables in the form of categorical variables. The aims of this paper to implement the use of multilevel logistic regression with random intercept model in explaining the factors that affect the status of migration in Jawa Barat province in 2015. From the random intercept model at the multilevel logistic regression obtained that the individual variables, namely: marital status, highest level of education, access to information, and the contextual variables, namely: the level of unemployment, population density and the minimum wage districts / cities significantly affect status of migration in Jawa Barat province. In a subsequent study is expected to examine the hierarchy data with more levels of analysis

Keywords: Hierarchy data, Multilevel logistic regression, Status of migration, Population

1. INTRODUCTION

Hosmer and Lemeshow (2000) states that the logistic regression method is an integral component of the analysis of data that describing the relationship between the response variable is categorical scale with one or more explanatory variables that can be categorical or continuous. Over the past ten years, the logistic regression model became standard analytical methods that are commonly used in various fields. If the response variable generated in the form of two categories, then the type of logistic regression analysis was used is a binary logistic regression analysis. Categorical scale data generated from an identical experiment that produces two types of responses, the success or failure where the odds of success in each trial is the same and the response generated between values do not affect each other (independent). The problem is if there is a hierarchy of structured data, which according to Goldstein (1995) hierarchical structured data has emerged for individuals who collected / nested within their groups respectively (Goldstein, 1995). Most of the research in the social field usually involves a problem that connects the individual with the environment or group. The concept generally, individuals interact with the surrounding social groups, so that individuals are influenced by social group, and the group is also affected by individuals who are in the group. Condition like this is called hierarchical data conditions. That is, the units on the hierarchical structured data that observed in the same group generally have similar characteristics compared with other units that observed in different groups. So that the correlation between units in the same group will be stronger, or in other words it can be said that the data is structured hierarchies are not completely independent.

In the state data analyzed turned out to have a form of hierarchy, then binary logistic regression method is not appropriate for use (Hox, 2010). So this paper will aim to apply the methods that can accommodate logistic regression analysis on the data with a hierarchical structure by using multilevel binary logistic regression analysis. In this paper the application of multilevel logistic regression method using the data migration status of the population in Jawa Barat province in 2015

2. THE MULTILEVEL BINARY LOGISTIC REGRESSION

Goldstein (1995) introduced a multilevel regression model that aims to solve the problems of hierarchy structured data. Data hierarchies arise because individuals who collected / nested within social groups respectively. Multilevel analysis is very suitable to be applied to the data or data group with the cluster as well as in the case of data obtained from a multistage sampling (Goldstein, 1995). A survey conducted by sampling gradually in fact have the effect of a group that oversees the elements examined or the effects of cluster called the design effect and it was indicated by the value of interclass correlation. This analysis observe, measure, and analize the variables at multiple levels simultaneously so the research that considers the relationship between individual characteristics and the characteristics of the group is a type of research that can be analyzed using multilevel analysis. Hox (2010) gives an example of data structured hierarchies used in research in the field of education where the population consists of schools and students in each school, design sampling in this study consisted of two stages: first, a couple of samples drawn from schools, and then samples drawn students from the students that exist in the schools selected as a sample in this example, students nested in schools.

In this paper the multilevel binary logistic regression analysis is used as the response variable was divided into two categories: migrated and does not migrated to the explanatory variables structured hierarchy. Interpretation of parameters in multilevel binary logistic regression analysis did not differ by binary logistic regression analysis of the level. However, in a multilevel binary logistic regression analysis to estimate variance components. Varian intergroup shows the effect of units on level 2^{nd} to level unit 1^{st} . This effect is referred to as random effects (random effect).

Multilevel research will divide the model into two: null models and conditional models. Null model is a condition in which the explanatory variables have not been incorporated into the model, both explanatory variables at level 1^{st} or level 2^{nd}. Meanwhile, conditional models is a condition in which the explanatory variables, both explanatory variables at level 1^{st} or level 2nd, has been incorporated into the model. So that in these conditions the model can also be called a multilevel binary logistic regression model. Hox (2010) mentions that the multilevel regression model can be classified into two basic forms, namely:

1. Model with random intercept multilevel

This model is a model where the intercept is modeled as a random effect of variables at level 2nd on the assumption that each group has intercept different (not fixed as ordinary regression), but has the inclination or slope the same so that the effect of each explanatory variable to variable same response for each group.

2. The multilevel model with random slope

This model is a model where the coefficients of explanatory variables at a lower level is modeled as a random effect of variables at level 2nd on the assumption that each group has a slope or a slope that is different (not fixed as ordinary regression) to enable the effect of explanatory variables on the response variable is different for each group. In this study, the model used is the multilevel model with random intercept because it assumes the influence of each explanatory variable on the response variable is the same for each group.

In general, according to the Hox (2010) binary logistic multilevel models with random intercept model at level 1^{st} can be written as follow

$$\ln\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right]=\beta_{0j}+\sum_{p=1}^{p}\beta_{pj}X_{pij}+\varepsilon_{ij} \tag{1}$$

Intercept in a random intercept models assumed to vary between groups. This variation is modeled by Z_j variables at level 2^{nd} with the following equation:

$$\beta_{0j}=\beta_0+\sum_{q=1}^{Q}\gamma_{0q}Z_{qj}+u_{0j} \tag{2}$$

Where:

π_{ij} = probability of event "success" for-i^{th} at level 2^{nd} in group of j^{th}

β_{0j} = value of random intercept

β_{pj} = fixed effects for all p-explanatory variables at level 2nd for group jth
X_{pij} = the explanatory variables with subscript p is for the p^{th} explanatory variables at level 1st (p = 1 . . . P), the subscript i is for all individuals at level 1st in the group jth at level 2nd (i = 1 . . . nj)
Z_{qj} = the explanatory variables with subscript q is for the q^{th} explanatory variables at level 2nd (q = 1 . . . Q) in the group jth at level 2nd
u_{0j} = the residual errors for the group jth at level 2nd (assumed in normal distribution with N~0,6$^2_{u0}$)
ε_{ij} = the residual errors for the individual ith at level 1st in the group jth at level 2nd
Based on the general equation above, the multilevel binary logistic regression model is as follows:
Level 1st (Individuals)

$$ln\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right]=\beta_{0j}+\beta_{1j}\ X_{1ij}+\beta_{2j}\ X_{2ij}+\cdots+\beta_{pj}\ X_{pij}+\varepsilon_{ij} \tag{3}$$

Level 2nd (Groups)

$$\beta_{0j}=\gamma_{00}+\gamma_{01}\ Z_{1j}+\cdots+\gamma_{0q}\ Z_{qj}+u_{0j} \tag{4}$$

where:

$$\beta_{1j}=\gamma_{10}\ .....\beta_{pj}=\gamma_{p0}$$

By substituted two models above, the final model is as follows:

$$ln\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right]=\gamma_{00}+\gamma_{10}\ X_{1ij}+\cdots+\gamma_{p0}\ X_{pij}+\gamma_{01}\ Z_{1j}+\cdots+\gamma_{0q}\ Z_{qj}+u_{0j}+\varepsilon_{ij} \tag{5}$$

Where :
γ_{00} = *intercept* in overall average
γ_{p0} = fixed effect/fixed slope for the explanatory variable p^{th} at level 1st with p=1,2..P
γ_{0q} = ixed effect/fixed slope for the explanatory variable q^{th} at level 2nd with q=1,2..Q

3. THE STAGES IN MULTILEVEL BINARY LOGISTIC REGRESSION ANALYSIS

In conducting multilevel binary logistic regression analysis needs to be done several stages as follows:
Estimation of parameter model
Hox (2010) explains that the estimator commonly used in multilevel regression analysis was estimator maximum likelihood estimation (MLEs). One form models using maximum likelihood estimation method is a form of Generalized Linear Models (GLM). GLM is a modern approach that is used when a problem occurs in the form of a variable that is not normal distributions. For instance, a commonly used generalized linear model for dichotomous data is the logistic regression model specified by (Hox, 2010):

1. the probability distribution is binomial (μ) with mean μ,
2. the linear predictor is the multiple regression equation for η, e.g., $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
3. the link function is the logit function given by η = logit(μ).

Note that this specification does not include a term for the variance of the error distribution. In the binomial distribution, the variance is a function of the mean, and it cannot be estimated separately. The estimation method in generalized linear models is a maximum likelihood procedure that uses the inverse of the link function to predict the response variable. The inverse function for the logit used above for binomial data is the logistic transformation given by $g(x) = e^x/(1 + e^x)$. The corresponding regression model is usually written as:

$$y=\frac{e^{(\beta 0\ +\ \beta 1X1\ +\ \beta 2X2)}}{1+e^{(\beta 0\ +\ \beta 1X1\ +\ \beta 2X2)}}$$

This regression equation is sometimes written as:
y = logistic ($\beta_0 + \beta_1 X_1 + \beta_2 X_2$).
In the process, MLEs had the principle will always maximize the likelihood function by using the inverse of the link fuction to predict the response variable. MLEs in multilevel modeling procedures resulting from the iterative process that begins with the parameter value estimates will increase in each successive iteration so that the parameter value will change during the iteration process.

b.Testing significance of random effect (likelihood ratio test)

Hox (2010) mentions that in the hierarchy of structured data contained random residual effect that is on both the level 1st and level 2nd . Testing significance of random effect is done by using the difference in deviance, namely likelihood ratio (LR) as -2 ln likelihood of a model. This test is used to determine whether there are differences between units in level 2nd using hypotheses:
H_0:σ^2_{u0} = 0 (random effect are significant)
H_1:σ^2_{u0} > 0 (random effect are not significant)
by using the test statistic Likelihood Ratio Test the form as follows:

$$LR = -2 \ln \left[\frac{\textit{likelihood logistic model without random effects}}{\textit{likelihood logistic model with random effects}} \right] \sim \chi^2 \quad (6)$$

If the value of LR > $\chi^2_{\alpha;1}$ or a p-value < α, then it will be rejected H_0
If the value of LR < $\chi^2_{\alpha;1}$ or a p-value ≥ α, then it will be accepted H_0
When LR > $\chi^2_{\alpha;1}$ then H_0 rejected, and will give the conclusion that there is a difference between groups at level 2nd so that an analysis using multilevel logistic model of better use than the logistic regression model one level.

c. Define value of Interclass Correlation Coefficient (ICC)

Multilevel logistic regression analysis has a calculated value as the intraclass correlation coefficient (ICC), where the value of ICC is a value that can be interpreted as an alleged relationship between the two individuals in the same group (Hox, 2010). ICC equation stated as follows:

$$\text{ICC} = \frac{\sigma^2_{u0}}{\sigma_{u0}{}^2 + \sigma_{e0}{}^2} \quad (7)$$

where σ^2_{u0} is the variance of error in level 2nd and σ^2_{e0} is the variance of error at an individual level. ICC value ranging from 0 to 1. When the value of the ICC is getting bigger or getting close to 1, it can be said that the diversity of the response variable due to the very large individual variation between groups and the effects of grouping data clearly visible so more suitable multilevel analysis applied to the data.

d. Parameters simultaneous test (G-Test)

Testing parameters simultaneously performed to test the suitability of the model simultaneously by using the test statistic G (Hosmer and Lemeshow, 2000). It is also used to test the overall parameters or determine the explanatory variables in the model simultaneously (Hox, 2010). This test can be done using the hypotheses:

$$H_0 : \gamma_{10} = \cdots = \gamma_{i0} = \cdots = \gamma_{01} = \cdots = \gamma_{0j} = 0$$

(No influence of explanatory variables on the response variable)
H_1:at least there is one γ_{ij}≠0; where i = 1, ..., p and j = 1, ..., q
(At least there is one explanatory variables that influence the response variable)
by using the test statistic G (Hosmer & Lemeshow, 2000):

$$G = -2 \ln \left[\frac{L_0}{L_p} \right] \sim \chi^2 \quad (8)$$

Information:
L_0 = the value of likelihood without the explanatory variables (null model)
L_p = value of likelihood with p explanatory variables (conditional model)
Statistics G follows the Chi-square distribution with degrees of freedom p, where p is the number of explanatory variables that are at level 1st and level 2nd .
If the value of G > $\chi^2_{\alpha;1}$ or p-value <α, it H_0 will be rejected
If the value of G < $\chi^2_{\alpha;1}$ or p-value ≥ α, then it H_0 will be accepted
When H_0 was rejected, it will give the conclusion that at the level of confidence (1-α) percent contained at least one explanatory variable in models which influence the response variable.

e. Parameters partial test (Wald-Test)

Hox (2010) states that the partial test aims to determine the significance of each parameter contained in the model using the hypotheses:

$H_0:\gamma_{pq}=0$ (No influence of explanatory variables to-p at level 1st or explanatory variables of q at level 2nd on the response variable)

$H_1:\gamma_{pq}\neq 0$ (There are significant influence of explanatory variables to-p at level 1st or explanatory variables of q at level 2nd on the response variable)

by using the test statistic of Wald (W):

$$W=\left[\frac{\hat{\gamma}_{pq}}{Se(\gamma_{pq})}\right]^2 \sim Z_{\alpha/2} \tag{9}$$

Where: $\hat{\gamma}_{pq}$ = estimator for γ_{pq} ; Se($\hat{\gamma}_{pq}$) = estimator for standard error of γ_{pq}.

If values of W> $Z_{\alpha/2}$ or p-value <α, it H_0 will be rejected, If the value < $Z_{\alpha/2}$ or p-value ≥ α, then it H_0 will be accepted. When H_0 was rejected, it will give the conclusion that the explanatory variables to-p at level 1st or explanatory variables to- q at level 2nd partially affect the response variable.

f. Odds Ratio

The odds ratio is widely used as a measure of association as it approximates how much more likely or unlikely (in terms of odds) it is for the outcome to be present among those subjects with x = 1 as compared to those subjects with x = 0 (Hosmer & Lemeshow, 2000)

Odds ratios are used to determine how likely one category than the other categories of explanatory variables in the variable specific response to the response variable for comparison. odds ratio estimate how successful the likelihood of an incident between observation x = 1 compared with observation x = 0 (Hosmer & Lemeshow, 2000).

The possible values of the logistic probabilities from a model containing a single dichotomous covariate coded 0 and 1. The *odds* of the outcome being present among individuals with x = 1 is $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with x = 0 is $\pi(0)/[1 - \pi(0)]$. The *odds ratio*, denoted OR, is the ratio of the odds for x = 1 to the odds for x = 0, and is given by the equation:

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \frac{\frac{\exp(\beta_0+\beta_j)}{1+\exp(\beta_0+\beta_j)}\frac{1}{1+\exp(\beta_0)}}{\frac{1}{1+\exp(\beta_0+\beta_j)}\frac{\exp(\beta_0)}{1+\exp(\beta_0)}} = \frac{\exp(\beta_0+\beta_j)}{\exp(\beta_0)} = \exp(\beta_j) \tag{10}$$

That is, the likelihood of risk of events in the category y = 1 at the category of x = 1 is equal to exp (β_j) times the risk / likelihood of events y = 1 in the category x = 0 (Nachrowi and Usman, 2006)

4. DATA AND METODOLOGY

The data used in this study is data collected from secondary data in the form of raw data and also the data from publication. The raw data used came from Population Survey Intercensal (SUPAS) 2015 conducted by the Statistics of Indonesia (BPS-Indonesia). Meanwhile, the publication of data used came from Survey National of Labor (Sakernas) the second half of 2015 and Jawa Barat in Figures 2016 issued by BPS and data on minimum wage labor on regency / city in Jawa Barat in 2015. Intercensal Population Survey (SUPAS) is done for the purpose of estimate the population and demographic indicators between the two time population census. A survey conducted in 2015. The binary logistic regression analysis in this study used multilevel to clarify the effect of explanatory variables in the form of individual variables and contextual variables on the response variable in the form of population migration status aged 15 years and over. The explanatory variables used in this research is divided into individual variables and contextual variables. Individual variables include respondents information regarding gender, marital status, the highest level of educational, and access to information. While contextual variables include information region (district) in the form of level of unemployment (TPT), population density, and the minimum wage regency / city (UMK). This study assumes the influence of each explanatory variable to variable responses is the same for each group, the model used is the model with random intercept multilevel. The statistical program STATA/MP version 13th were used for all of the inferential analysis in this study.

5. CONCLUSIONS

The stages from application of the multilevel logistic regression analysis on status of migration of population in Jawa Barat province 2015 with data of SUPAS 2015 was as follows:

a. Estimation of parameter model

Based on the results of data processing SUPAS 2015 by using statistical software Stata / MP version 13, the results are as follows:

1. The LR test vs. Logistic Regression: chibar2 (01) = 28916.36> chibar2 = 0.0000
2. Log Likelihood (null model) = -79,289.731
3. Log Likelihood (conditional model) = -72,470.398
4. ICC = 0.22362

Table 1. The Result of Processing Data

Variables	Coefficient (γ)	*Standard Error*	Wald (Z)	*p-value*	*Odds Ratio*
-1	-2	-3	-4	-5	-6
Gender	-0,0165	0.0135	-1.23	0.22	0.9836
Marital Status	-1.1936	0.0172	-69.59	0.000*	0.3031
Highest Level of Educational					
Middle school/lower	0.5283	0.02	26.37	0.000*	1.6959
High school/lower	1.1005	0.0187	58.85	0.000*	3.0056
High School/higher	1.6118	0.0244	66.18	0.000*	5.0117
Acces to Information	0.5759	0.021	27.39	0.000*	1.7788
Level of Unemployment	-0.1066	0.0535	-1.99	0.046*	0.8988
Density of Population	0.1161	0.0237	4.9	0.000*	1.1231
Minimum Wage	0.6268	0.201	3.12	0.002*	1.8715
_cons	-3.0232	0.4309	-7.02	0.000*	0.0486

Source : STATA output

So the multilevel binary logistic regression model for the status of migration of population in Jawa Barat province was as follows :

$$\ln\left[\frac{\widehat{\pi_{ij}}}{1-\widehat{\pi_{ij}}}\right] = -3.0232 - 0.0165\,Gender_{ij} - 1.1936 Marital\ Status^{*}_{ij} + 0.5283 Highest\ Level\ of\ Educational1^{*}_{ij}$$
$$+ 1.1005 Highest\ Level\ of\ Educational2^{*}_{ij} + 1.6118 Highest\ Level\ of\ Educational3^{*}_{ij}$$
$$+ 0.5759 Access\ to\ Informations^{*}_{ij} - 0.1066 Level\ of\ Unemployment^{*}_{ij}$$
$$+ 0.1161 Density\ of\ Population^{*}_{ij} + 0.6268 Minimum\ Wage^{*}_{ij}$$

*) significant at the 95% confidence level

b. Testing significance of random effect (likelihood ratio test)

From the processing data obtained LR value that follows the Chi-square distributions amounting to 28916.36, and the p-value of 0.0000. Based on the equation (6), the value $\chi^2_{(0.05,1)}$ was 3.84, it's meaning LR is greater than $\chi^2_{(0.05,1)}$. So that the decision can be taken that H_o is rejected, because there is a random effect significant migration status. This gives the conclucion that the use of multilevel logistic regression model (logistic models with random effects) is better than ordinary logistic regression model to analyze the migration status of the population aged 15 years and above.

c. Define value of Interclass Correlation Coefficient (ICC)

The results of data processing SUPAS 2015 using equation (7) obtained in the form of null models ICC value was 0.22362. This figure shows that 22.362 percent of the variation migration of the population aged 15 and above in Jawa Barat province was influenced by differences in the characteristics of individuals in each regency / city.

d. Parameters simultaneous test (G-Test)

Results of processing data SUPAS 2015 indicates output as follows:

L_0=-79 289,731; L_p= -72 470,398; $\chi^2_{(0,05;7)}$ =14,07;

$$G = -2Ln\left[\frac{L_0}{L_p}\right] = -2Ln\left[\frac{-79.289{,}731}{-72.470{,}398}\right] = (158.579{,}462 - 144.940{,}796) = 13.638{,}666$$

The counting process that has been done by using equation (8) obtained the G-Test value was amounted 13638.666, where the value is much greater than $\chi^2_{(0{,}05;7)}$ which is only at 14.07. In order to take decisions that H_o is rejected, or in other words there is minimum one of explanatory variables that influence population migration status variables age 15 years old and above in Jawa Barat Province.

e. Parameters partial test (Wald-Test)

From the data processing SUPAS 2015 in tables 1 note that the explanatory variables such as marital status, highest level of education, access to information, level of unemployment, density of population, and the minimum wages used in the model has the absolute value of the Wald test (| Z |) is greater than the value $Z_{\alpha/2}$ ($Z_{(0{,}025)}$ = 1.96). In order to take decisions that Ho is rejected, or in other words the explanatory variables are variables affecting migration status of the population aged 15 years and above partially.

f. Odds Ratio

The odds ratio for marital status was 0.3031, which means that people who had married was 0.3031 times more likely to migrated than people who still single. People with level of education middle school/ lower was 1.6959 times more likely to migrated than people with elementary school/lower, people with level of education high school/lower was 3.0056 time more likely to migrated than people with elementary school/lower, people with level of education high school/higher was 5.0117 times more likely to migrated than people with elementary school/lower. People with has access to informations was 1.7788 times more likely to migrated than people with has not access to informations. The odds ratio of level of unemployment was 0.8988 which means that the odds for people who had migrated was decreased 0.8988 times by one percent increased of the level of unemployment. The odds ratio of density of population was 1.1231 which means that the odds for people who had migrated was increased 1.1231 times by one thousand persons/km^2 increased of the population density. The odds ratio of minimum wage 1.8715 which means that the odds for people who had migrated was increased 1.8715 times by one million rupiah increased of the minimum wage.

REFERENCES

BPS. (2015). *Jawa Barat dalam Angka* (Jawa Barat in Figures) *2015*. Jakarta: BPS.

BPS. (2016). *Penduduk Indonesia Hasil Survei Penduduk Antar Sensus 2015*.Jakarta: BPS.

BPS. (2016). *Jawa Barat dalam Angka (*Jawa Barat in Figures) *2016*. Jakarta:BPS

Goldstein, Harvey. (1995). *Multilevel Statistical Models.* London: Arnold London.

Hosmer, David W. & Lemeshow, Stanley. (2000). *Applied Logistic Regression (2nd ed)*. New York: John Wiley & Sons, Inc.

Hox, Joop J. (2010). *Multilevel Analysis (Techniques and Applications): Second Edition*. New York: Routledge.

Nachrowi, Djalal & Usman, Hardius (2006) *Pendekatan Populer* Dan *Praktis Ekonometrika Untuk Analisis Ekonomi* Dan *Keuangan*. Jakarta: Lembaga Penerbit Fakultas Ekonomi Universitas Indonesia

Surat Keputusan Gubernur Jawa Barat Nomor: 561/ Kep. 1581-Bangsos/ 2014 Tentang Upah Minimum Kabupaten/Kota di Jawa Barat Tahun 2015.

Internet Access and Digital Divide in Indonesia

Eni Lestariningsih*
Directorate of Finance, IT, and Tourism Statistics, BPS-Statistics Indonesia,
Jakarta, Indonesia – elen@bps.go.id
Atika Nashirah Hasyyati
Directorate of Finance, IT, and Tourism Statistics, BPS-Statistics Indonesia,
Jakarta, Indonesia – atika.hasyyati@bps.go.id
Karmila Maharani
Directorate of Finance, IT, and Tourism Statistics, BPS-Statistics Indonesia,
Jakarta, Indonesia – karmila@bps.go.id

Abstract

Information society era can be seen through the rapid growth of Information and Communication Technology (ICT), but there are still gaps among countries and within countries which is known as digital divide. Digital divide is also as a gender problem. Digital divide as a problem of women including access to the internet and others lack of access to ICTs, and also digital divide in terms of advanced ICT skills. The analysis includes all provinces in Indonesia in 2015. Internet penetration data was calculated from on National Socio Economic Survey (Susenas) data. Meanwhile, ICT Development Index and its sub-index, percentage of poor people, Human Development Index (HDI), and Gross Regional Domestic Product (GRDP) per capita were taken from Statistics Indonesia website. Plot does not show a high relationship between Gini ratio and internet access (in women and men cases), while plot shows a positive relationship between GRDP per capita and internet access (in urban and rural cases). This study used regression models to investigate the differences between men and women internet access on GRDP per capita. The results show that internet access by women influence GRDP per capita significantly, if internet access by women increase 10%, GRDP per capita will increase by 0.51%. Internet access by men also influence GRDP per capita significantly, if internet access by men increase 10%, GRDP per capita will increase by 0.43%. In addition, this study also used canonical correlation analysis, the canonical correlation shows that the relationship between socioeconomic variables and digital divide was statistically significant.

Keywords: Digital divide; Internet access; GRDP per capita, ICT Development Index.

JEL Classification: C21, C39, O11, 047

1. INTRODUCTION

Information society era can be seen through the rapid growth of Information and Communication Technology (ICT), but there are still gaps among countries and within countries which is known as digital divide. The Internet Society and TRPC (2015) mention that the digital divide within nations is between the 'haves' and the 'have-nots' in terms of internet connectivity and affordable usage, while the digital divide between nations is about bandwidth (capacity) and transmission speeds (usage). Digital divide has been a problem of the world today especially among ASEAN countries. Across ASEAN countries internet penetration varies from as much as 73% of the population in Singapore to little more than 1% in Myanmar (The Internet Society and TRPC, 2015). Based on World Bank Indicators 2014, Indonesia categorized in the third cluster as low internet access. It was different compare to Singapore, Malaysia, and Brunei Darussalam which were in the first cluster as majority access, while Thailand, Philippines, and Vietnam in the second cluster as partial access.

Digital divide is also as a gender problem. Digital divide as a problem of women including access to the internet and others lack of access to ICTs, and also digital divide in terms of advanced ICT skills. ITU (2013) estimated that some 200 million fewer women are online, compared with men. In 2003, Commission on the Status of Women had directly focused on the issue of ICT and the empowerment of women. The commission adopted agreed conclusions which addressed women's equal access to ICT-based economic activities and employment (United Nations, 2005).

Economic development can be achieved by enhancing some social economic indicators such as human development and ICT. ICT can accelerate the development of economics which in turn impact to social economic life. ICT Development Index as a measure of the development of Information and Communication Technology (ICT) so that it needs more understanding and concern. ICT Development Index is very sensitive index. It is because it depends on data source, definition and method in collecting data, ideal value, and area, national or sub-national level. ICT Development Index is a composite index so that the figures depend on each component index and the weight of each component. Sub-directorate of Communication and Information Technology Statistics, Statistics Indonesia, have been trying to develop ICT Development Index in sub-national level in Indonesia.

ICT development index as a composite index which combine eleven indicators into one benchmark measure, consists of three sub-index. The three sub-index of ICT development index are ICT access and infrastructure, ICT use, and ICT skill. Based on the International Telecommunication Union (ITU) data, sub-index use has the smallest contribution to ICT development index in term of average value in 2015. By improving internet access, digital divide can be reduced.

Factors affecting economic development can be classified into two, they are economic factors and non-economic factors. Investment in human capital could have a more permanent impact on the growth process if high skills and training go hand-in-hand with more intensive research and development and a faster rate of technological progress, or if the adoption of new technologies is facilitated by a highly skilled workforce (Bassanini and Scarpetta, 2001). Robert J. Barro (1996) found that the growth rate is enhanced by higher initial schooling and life expectancy, lower fertility, lower government consumption, better maintenance of the rule of law, lower inflation, and improvements in the terms of trade.

The Solow growth model is designed to show how growth in the capital stock, growth in the labor force, and advances in technology interact in an economy as well as how they affect a nation's total output of goods and services (Mankiw, 2009). According to the Solow model, technological progress can lead to sustained growth in output per worker where it is persistently rising living standards. Neo-classical growth theory, based on the expanded Solow model, suggests that long run growth is dependent upon labor force growth and technological progress (Grossman and Helpman, 1994). This study enhance the literature of Solow growth model and Neo-classical growth theory by using internet penetration as technological progress.

By providing access to information, connecting people to business everywhere, and opening up new markets, the internet can transform the very nature of an economy and support economic development (Deloitte, 2014). Internet enhanced speed and quality of information flows result in reduced transaction costs, internet enhanced access to financial capital with services such as mobile banking, internet access enhances the productivity of labour and capital, internet improved human resource qualification and specialization, internet can be facilitation of entrepreneurship and business expansion so it opens the access to new markets, and there will be greater innovation and adoption of new organizational models and business processes (Deloitte, 2014).

There are a number of factors that make it difficult for people to obtain access to the internet such as poverty; high device, data, and telecommunications charges; infrastructure barriers; digital literacy challenges; and policy and operational barriers (West, 2015). In addition, West (2015) mentioned that lacking disposable financial resources makes it difficult to purchase devices or gain access to digital services.

According to Ernst & Young, over the next decade, the impact of women on the global economy – as producers, entrepreneurs, employees and consumers – will be at least as significant as that of China's or India's one –billion-plus populations, if not greater (UNESCO, UN Women, ITU, and Microsoft, 2013). Ann Mei Chang in ITU and UNESCO (2013) stated that access for women is often correlated with the development of a country, implying that the gender gap will fall away as an economy matures. Furthermore, World Bank (ITU and UNESCO, 2013) has estimated that a 10% increase in broadband adoption will result in a 1.38% increase in economic growth which is also intuitively obvious, as access to the Internet can enable women to increase their productivity, access new markets, improve their education, find better jobs, and contribute to the innovation economy. In sum, the main question of this study are whether women and men internet access can influence economic development or not in Indonesia, and this study also try to investigate the relationship between socioeconomic variables and digital divide in Indonesia in 2015. The first part of this paper about the difference in internet access of women and men that indicate the digital divide in Indonesia. The second part of this paper analyse the digital divide in Indonesia in 2015 based on the relationship between digital divide and socioeconomic variables. This study use percentage of poor people, Human Development Index (HDI), skill sub-index, unemployment rate, and Gross Regional Domestic Product (GRDP) per capita as proxy of other social indicators.

2. DATA AND METHODOLOGY

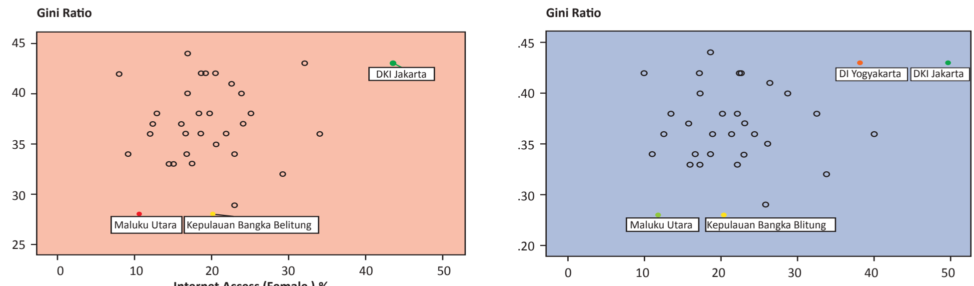
The analysis include all provinces in Indonesia in 2015. Internet penetration data was calculated from National Socio Economic Survey (Susenas). Meanwhile, ICT Development Index and its sub-index, percentage of poor people, Human Development Index (HDI), and Gross Regional Domestic Product (GRDP) per capita were taken from Statistics Indonesia website. In order to investigate the influence of internet penetration on economic development, we obtained linear regression model.

This study also uses the canonical correlation to analyse the relationship between socioeconomic variables and digital divide. Canonical correlation is a method that enables the assessment of the relationship between two sets of multiple variables (Hair, et.al., 2010). In applying canonical analysis, it is helpful to think of one set of variables as independent and the other set as dependent, however, it does not imply that they share a causal relationship (Hair, et.al., 2010).

3. INTERNET ACCESS IN INDONESIA

In 2015, the percentage of Internet users in Indonesia amounted to 21.98 pesen. Among 34 provinces, DKI Jakarta as the capital city of Indonesia has the highest percentage of internet user, 46.63 %. The infrastructure and facilities of the internet are well provided in this city. While the lowest percentage of internet users was in Papua, 8.98%. Papua is located in the eastern part of Indonesia. The development of the infrastructure, the skill of the people, and other socioeconomic aspects in Papua are left from the other provinces.

The internet access by gender is believed as one of the factor that influence the socioeconomic condition in a province. Gini ratio describes the income inequality in a province. The plot between internet access by gender and gini ratio of 34 provinces is shown in the plot. The plot does not show a high relationship between gini ratio and internet access by female. DKI Jakarta has the highest internet access by female and also quite high gini ratio. It means that DKI Jakarta still has high income inequality although many of the female in this city accessing the internet. While Maluku Utara has the lowest internet access by female and the lowest gini ratio. It means that the income in Maluku Utara is relatively equal when there are less female accessing the internet. Kep. Bangka Belitung has a low gini ratio and slowest speed in the increasing of internet access percentage.



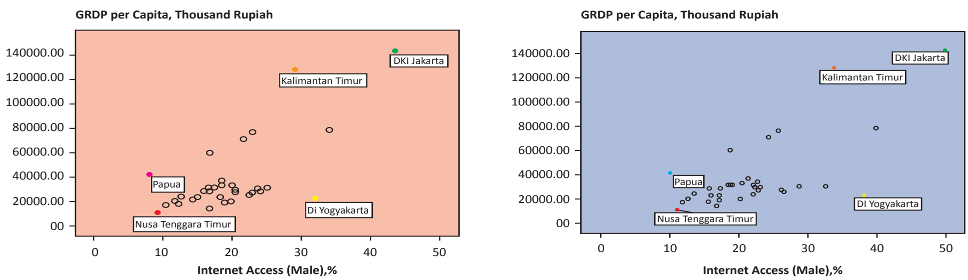
Source: Author's Calculation, BPS-Statistics Indonesia

Figure 1. Plot of Gini Ratio and Internet Access (Female and Male), 2015

The same pattern is shown in the plot between internet access by male and gini ratio. The plot does not show a high relationship between gini ratio and internet access (male). DKI Jakarta has the highest internet access (male) and gini ratio, and the lowest both in internet acccess (male) and gini ratio was Maluku Utara. Kep. Bangka Belitung has a low gini ratio and slowest speed in the increasing of internet access percentage.

Besides gini ratio, GRDP percapita can also show the socio economic condition in a province. The plot shows a positive relationship between GRDP per capita and internet access (female). DKI Jakarta has the highest internet access (female) and GRDP per capita. Kalimantan Timur is the second province with high GRDP and the internet access by female also quite high. Nusa Tenggara Timur has the lowest internet access (female) and GRDP per capita. While Papua has a quite high GRDP per capita, but very low in internet access (female).

The plot between GRDP per capita and internet access (male) also shows a positive relationship. The pattern is similar with the plot of female, that is DKI Jakarta has the highest internet access (male) and GRDP per capita. Nusa Tenggara Timur has the lowest internet access (male) and GRDP per capita. As for Papua has a quite high GRDP per capita, but very low in internet access (male).



Source: Author's Calculation, BPS-Statistics Indonesia

Figure 2. Plot of GRDP per Capita and Internet Access (Female and Male), 2015

Based on regression analysis, in 95% confidence level, 42.5% variation of GRDP per capita can be explained by internet access (female). Internet access by female influence GRDP per capita significantly, if internet access by female increase 10%, GRDP per capita will increase by 0.51%. the result is a bit different with male that in 95% confidence level, 42.8% variation of GRDP per capita can be explained by internet access (male). Internet access by male influence GRDP per capita significantly, if internet access by male increase 10%, GRDP per capita will increase by 0.43%. Ann Mei Chang in ITU and UNESCO (2013) stated that access for women is often correlated with the development of a country, implying that the gender gap will fall away as an economy matures. Furthermore, World Bank (ITU and UNESCO, 2013) has estimated that a 10% increase in broadband adoption will result in a 1.38% increase in economic growth which is also intuitively obvious, as access to the Internet can enable women to increase their productivity, access new markets, improve their education, find better jobs, and contribute to the innovation economy.

$$Ln(\widehat{GRDPpC}_i) = 9.366+0.051^{*}\ Internet\ access\ (female)_i$$
$$R^2 = 0.425$$

$$Ln(\widehat{GRDPpC}_i) = 9.419+ 0.043^{*}\ Internet\ access\ (male)_i$$
$$R^2 = 0.428$$

4. THE DEVELOPMENT OF ICT IN INDONESIA

The development of ICT in Indonesia can be figured by ICT Development Index as an composite index which combine three sub-index. Three sub-index of ICT Development Index are access and infrastructure, use, and skill. ICT Development Index of Indonesia were increase year by year during 2012 to 2015, in 2012 ICT Development Index of Indonesia was 4.24, increase to 4.5 in 2013, 4.59

in 2014, and 4.83 in 2015. In line with the index, its sub-index were also increase year by year during 2012-2015.

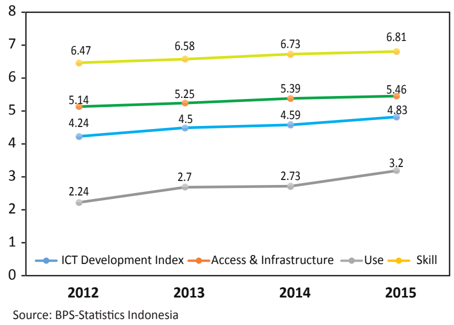


Figure 3. ICT Development Index of Indonesia and Its Sub-Index, 2012-2015

In order to investigate the relationsip between group of variables X (socioeconomic variables) and group of variables Y (ICT development) in 2015, the canonical correlation analysis was obtained. In this case, group of variables X (socioeconomic variables) consist of skill sub-index of ICT Development Index, health (percentage of population having health complaint during the last month), unemployment rate, percentage of poor people, and Gross Regional Domestic Products (GRDP) per Capita. In overall, Omnibus test shows that all variables used were statistically significant. Meanwhile, group of variable Y (ICT development) consist of access and infrastructure sub-index and use sub-index of ICT Development Index.

According to the canonical correlation analysis, the canonical correlations for the first function was statistically significant. This first function has result with a canonical correlation of 0.89474 (Wilks lambda = 0.16625 ; F = 7.84363; df = 10 ; p = 0.000). The canonical correlations for the second function was not statistically significant. The second canonical correlation function produce a canonical correlation of 0.40788 (Wilks lambda = 0.83363 ; F = 1.39698; df = 4 ; p = 0.261). The highest correlation was between digital divide and GRDP per capita (0.826).

Variance explained by canonical variables of dependent variables (socioeconomic variables) tell us 25.58 percent of variance between dependent variables accounted for by the dependent canonical variate, and a 20.48 percent of variance between the dependent variables is accounted for by the covariate canonical variate. In addition, variance explained by the canonical variables of the covariates (ICT development) tell us 77.09 percent of variance among the covariates accounted for by the dependent canonical variate, and a 96.296 percent variance among the covariates is accounted for by the covariate canonical variate.

5. CONCLUSIONS

Internet access by female influence GRDP per capita significantly. As well as the model of female internet access, internet access by male also influence GRDP per capita significantly. The regression model of female shows that if female internet access increase it will lead to the increasing of economic development, this result was better than the model of male. Moreover, the result of the canonical correlation indicate a statistically significant relationship between socioeconomic variables and digital divide. The effect of GRDP per capita was more than the effects of other socioeconomic variables on digital divide. This study has indicated that socioeconomic variables are correlated with the reduction of digital divide in Indonesia.

REFERENCES

Barro, Robert J. (1996). NBER Working Paper Series, Determinants of Economic Growth: A Cross-Country Empirical Study, NBER Working Paper 5698. Cambridge: National Bureau of Economic Research.

Bassanini, Andrea and Stefano Scarpetta. (2001). The Driving Forces of Economic Growth: Panel Data Evidence for the OECD Countries, OECD Economic Studies, No. 33, Vol. 2.

Deloitte. (2014). Value of Connectivity: Economic and Social Benefits of Expanding Internet Access. London: Deloitte LLP.

Grossman, Gene M. and Elhanan Helpman. (1994). Endogenous Innovation in the Theory of Growth, The Journal of Economic Perspectives, vol. 8, no. 1, pp. 23–44.

Hair, Joseph F. et.al. (2010). Canonical Correlation: A Supplement to Multivariate Data Analysis Multivariate Data. Pearson Prentice Hall Publishing.

ITU and UNESCO. (2013). Doubling Digital Opportunities: Enhancing the Inclusion of Women & Girls in The Information Society. Geneva: Broadband Commissions.

Mankiw, N. Gregory. (2009). Macroeconomics [Seventh Edition]. Cambridge: Wroth Publishers.

Statistics Indonesia. (2015). Welfare Statistics 2014. Jakarta: BPS, Statistics Indonesia.

The Internet Society and TRPC. (2015). Unleashing the Potential of the Internet for ASEAN Economies. The Internet Society (ISOC) and TRPC Pte Ltd.

UNESCO, UN Women, ITU, and Microsoft. (2013). Girls in STEM and ICT Careers: The Path toward Gender Equality. Accessed in www2.tku.edu.tw/~tfstnet/upload/file/20130704153742d38d9.pdf, February, 26th 2016.

United Nations. (2005). Women 2000 and Beyond: Gender Equality and Empowerment of Women through ICT. New York: United Nations.

West, Darrel M. (2015). Digital Divide: Improving Internet Access in the Developing World through Affordable Services and Diverse Content. Washington: The Brookings Institution.

A Panel Data Analysis of The Role of Human Development Index in Poverty Reduction in Papua 2010 – 2015

Faisal Arief*
Statistics of Ternate City, Ternate, Indonesia – faisal.arief@bps.go.id
Erli Wijayanti Prastiwi
Statistics of West Sumba Regency, West Sumba, Indonesia – erliwijayanti@bps.go.id

Abstract

Papua is a province in Eastern Indonesia having the highest percentage of poor people in the country for years. Although decentralization had been embarked for 16 years, Papua remains mired in poverty and marginalized. The large size of the island, the topography, and the distance with other region give Papua major challenges to reduce poverty not only in economic and physical development but also in human development. In economic and physical development, government has done several things. Infrastructure has been built, some cities are being more modern and great amount of aid also has been given. In human development matters however, Papua has the lowest Human Development Index (HDI) among other 34 provinces in Indonesia for years. Besides, it is critical to make people in Papua fight poverty and ready for the economic and physical development. This paper has two major aims. First, to analyse the role of HDI on poverty reduction in Papua by using Panel Data Analysis. Second, to examine which components and indicators of HDI have significant role on poverty reduction in Papua by using the same methodology. The data set used in this paper are HDI and percentage of poor people in 26 city or regencies in Papua province 2010 – 2015. The best model chosen in the statistical test is Fix Effect Model (FEM). The results show that HDI is significantly affected the poverty reduction in Papua. The components of HDI which significantly affected poverty reduction in Papua are component of education (knowledge) and decent standard of living which are mean years of schooling, expected years of schooling, and purchasing power parity. The principal contribution of this paper are to give new perspective for the government and for researchers that the development in Papua is not only about the economic and physical infrastructure but also about the human. All stakeholders have to ensure that the people in Papua are ready for all the development. It is essential to increase Human Development Index in Papua especially the component of education and decent standard of living. For the government, this paper provides a new reference about which indicators in component of education have the biggest contribution on poverty reduction, so that need to be place in the first priority. In other words, the facilities at schools, the number of teachers, the capability of the teachers, as well as capacity building for the educators is exceptionally required. Besides that, the government also should be focus on efforts to increase the purchasing power of the poor through cash transfer and labour-intensive program.

Keywords: Papua, Panel, Econometrics, Education

1. INTRODUCTION

In 2001, the Indonesian government issued Law No.21 of 2001 on Special Autonomy for Papua Province. The goals of the special autonomy is to give greater authority to the province and the people of Papua to govern and regulate usage of Papua province's natural resources for the prosperity of the people. On the other hand, the special autonomy for Papua Province is made to reduce the gap between the Papua Province and other provinces and improve the human life in the province. Since established as a special autonomous region, Papua has some specificly and significantly different rights than other provinces. One of these special rights are on the side of the financial budget. First, the percentage of equalization funds from the mining of petroleum and natural gas by 70 percent over the year-to-1 up to the 25th year and to 50 percent for the 26th year onwards. Second, a special reception in the implementation of special autonomy equivalent to 2 percent of the ceiling of the General Allocation Fund for education funding (30 percent) and health (15 percent). Third, additional funds for infrastructure development.

Based on these descriptions, it is certain that the central government budget that gave to Papua Province is the biggest compared to other provinces. However, until now, Papua has the lowest of welfare compared with other provinces. The quality of human resources, education level and health are still very low. Thus, this condition can lead to poverty. And it is proven that Papua still had the greatest percentage of poor people in all over Indonesia.

Additionally, Papua is a province wide and topography of mountains and forests, the distance between districts far enough, modes of transportation are limited and costly, difficult access, as well as the high price of basic necessary is also an obstacle for local governments and the people to achieve prosperity. This is a major challenge for the government of Papua to how to overcome these constraints in order to achieve wellfare for all in Papua.

With the enactment of legislation of special autonomy for Papua, this shows the seriousness of the government to increase the welfare of the people of Papua. Infrastructure development has been built,, several cities in Papua has become the center of the economy, a wide range aid also been given. Seeing this reality, the development in Papua is still limited to the physical and economic development eventhough it is is still not affect the quality of human resources. So this also sparked poverty in Papua. Based on BPS data, HDI Papua Province during the past five years to occupy the lowest rank. In addition, the province holds the highest percentage of poor people in Indonesia.

2. LITERATURE REVIEW

Poverty

According to the 2000 World Development Bank (WDR), "poverty is pronounced deprivation in wellbeing". In other hand, poverty is multifaceted, manifested by conditions that include malnutrition, inadequate shelter, unsanitary living conditions, unsatisfactory and insufficient supplies of clean water, poor solid waste disposal, low educational achievement and the absence of quality schooling, chronic ill health, and widespread common crime (UNSD, 2005).

In a broad sense, poverty means the inability of a person to fullfil their needs in accordance with the relative perception of himself. To measure the poverty, Statistics Indonesia (BPS) uses the concept of ability to fullfil the basic needs (basic needs approach). With this approach, poverty is seen as an economic inability to fullfil the basic needs of food and non-food which is measured from the expenditure side. So the Poor is the population had an average monthly per capita expenditure below the poverty line.

Conceptually, poverty can be distinguished according to the assessment standards relative poverty and absolute poverty. Relative poverty is a poor conditions specified and determined subjectively by the local community and locally. Meanwhile, absolute poverty is defined as assessed by the poor condition of the minimum living standard of a society that is realized in the poverty line.

The poverty line is the minimum living standards to fullfil the basic needs. BPS calculate poverty lines using two components, namely food and non-food components. Of the two components yields a food poverty line and non-food poverty line. Thus, the poverty line can be calculated by summing food poverty line and the non-food poverty line.

Human Development

Human development is a process of enlarging people's choices (HDR, 1990). The most critical of these wide-ranging choices are to live a long and healty life, to be educated and to have access to resources needed for a decent standard of living. Additional choices include political freedom, guaranteed human rights, and personal self-respect.

In the system of measurement and monitoring of human development, in 1990 UNDP introduced Human Development Index (HDI) as a measurement of human development achievements. This index uses a three-dimensional approach are a long and healthy life, knowledge, and a decent standard of living. Of these three dimensions, there are four indicators used in the calculation. In the new method of calculating the HDI, an indicator used the life expectancy rate (LER), mean years of schooling (MYS), expected years of schooling (EYS), purchasing power parity (PPP) (BPS, 2015)

The Linkage of Poverty and Human Development

Improved human development will lead to prosperity and people's lives for the better (Hidayat, N. K., 2008). In other words, when the quality of the human being be better, the human choices will be more extensive, so the human has a greater chance to be success. In simple terms, when health, education, and financial condition of a person in good condition, then the person has a lot of opportunities, one of them is obtaining a better job. By obtaining a better job, then welfare will be assured, and be spared from the poverty problem. This statement is also supported by the results of research Singh (2012) that the human development index and per capita income have influence on the reduction of poverty.

3. METODOLOGY

This paper uses panel data which is a combination of cross-section data and time series data. Cross-section data consist of 26 districts/cities in Papua province, while the time series data as much as 6 series, ie from 2010 to 2015. The variables used in this paper are the percentage of poor people (poverty), HDI, LER, MYS, AYS, and PPP. The data was obtained from the Statistics Indonesia publications are Human Development Report and Data and Information Poverty by Regency/City in Indonesia. This paper uses the panel data regression analysis which aims to determine the influence of the human development on poverty in Papua. The dependent variable is poverty, and the independent variables are HDI, LER, MYS, AYS, and PPP.

Panel Data Analysis

Panel data, also known as longitudinal data, have both time series and cross-sectional dimensions (Brooks, 2008). The panel data analysis can be defined as a technique which uses cross data of the time dimension to predict the economical relations (Greene, 2003). The panel data analysis uses the affects of the time as much as it uses the affects of the cross sections (Wooldridge, 2002). They arise when we measure the same collection of people or objects over a period of time. Econometrically, the setup is :

$$y_{it} = \alpha + \beta x_{it} + u_{it} \tag{1}$$

where y_{it} is the dependent variable, α is the intercept term, β is a $k \times 1$ vector of parameters to be estimated on the explanatory variables, x_{it}; $t = 1, \ldots, T$; $i = 1, \ldots, N$. The advantage using the panel data regression is accommodate individual heterogeneity among others, the data is more informative and varied, accommodate dynamic changes, increase the degrees of freedom, increase the precision, and the estimation becomes more efficient (Brooks, 2008; Baltagi, 2005).

In the panel data regression, there are several steps that must be done, ie select the estimation model, determine the estimation method, assumptions testing and goodness of fit test. On the panel data regression analysis, the estimation model generally used there are three types, as follows:

Pooled Model

Pooled Model is the estimation model with the same intercept and same slope for every individual. There isn't inter-individual heterogeneity (α equal). The model uses the same assumptions as in cross-section data regression. Ordinary Least Squares (OLS) produce most efficient and consistent estimators (Park, 2011).

Fixed Effect Model

A fixed group effect model examines individual differences in intercepts, assuming the same slopes and constant variance across individual (group). Since an individual specific effect is time invariant and considered a part of intercept, individual effect is allowed to be correlated with other regressor. This fixed effect model is estimated by least squares dummy variable (LSDV) regression (OLS with a set of dummies) and within effect estimation methods.

Random Effect Model

A random effect model assumes that individual effect is not correlated with any regressor and the estimates error variance specific to groups (or time). Hence, individual effect is an individual specific random heterogeneity or a component of the composite error term. This is why a random effect model is also called an error component model. The intercept and slopes of regressors are the same across individual. The difference among individuals (or time periods) lies in their individual specific errors, not in their intercepts.

A random effect model is estimated by generalized least squares (GLS) when a covariance structure of an individual is known. The feasible generalized least squares (FGLS) or estimated generelized least squares (EGLS) method is used to estimate when a covariance structure is not known. There are various estimation methods for FGLS including the maximum likelihood method and simulation (Park, 2011).

Next, conduct appropiate formal test to examine individual group and/or time effects. If the null hypothesis of the LM test is rejected, a random effect model is better than the pooled model. If the null hypothesis of the F-test is rejected, a fixed effect model is favored over pooled model. If both hypothesis are not rejected, fit the pooled model. Conduct the Hausman test when both hypothesis of the F-test and LM test are all rejected. If the null hypothesis of uncorrelation between an individual effect and regressors is rejected, go for the robust fixed effect model, otherwise, stick to the efficient random effect model (Park, 2011).

After the estimation model is selected, then select appropriate estimation method according to the variance-covariance structure residuals. This step is necessary because using panel data regression causing variance-covariance residuals will be complex than using simple regression with cross-section or time series data. In addition, the selection of estimation methods can also be used to accommodate the classic assumption violation on panel data regression. Greene (2003) and Gujarati (2004) divides the structure of the variance-covariance residuals according to condition 1) Homoscedastic without correlation between the units, 2) heteroscedastic without correlation between the units, 3) heteroscedastic with correlation between the units and 4) heteroscedastic with autocorrelation. Some of these conditions causing estimation methods become different. Ekananda (2005) says that the estimation method using OLS (condition 1), weight least square (WLS) cross-sectional weight or GLS (condition 2), Seemingly uncorrelated Regression (SUR) weight or FGLS (condition 3), and FGLS with the residual autoregressive (condition 4).

The next stage is test the assumptions and goodness of fit models. The classic assumption test is done to resolve of the violation of basic assumptions. In general, the classical assumption test performed when the method of estimation using OLS. The Classical assumption tests include, normality test, homoscedasticity test, non-autocorrelation test, and non-multcolinearity test. And the final stage is testing the significance of regression models with look at the the best criteria. the criteria include coefficient of determination (*adjusted R-squared*), simultaneous test (*F-Statistics*), and partial test (*t-Statistic*).

4. RESULT

After going through the stages of model selection, examination of the structure of the variance-covarians residual, classic assumption test and testing the feasibility of the model, and the best model to analyze the influence of the human development index on poverty is a fixed effect model with cross-section weight and cross-section SUR (PCSE) for model I and model II. Summary of panel data regression output listed in Table 1.

Table 1. Summary of Output Panel Data Regression

<table>
<tr><th colspan="7">Model I
(independent variable is HDI, while dependent variable is poverty)</th></tr>
<tr><th>Variable</th><th>Coefficient</th><th>Std. Error</th><th>Prob t-Statistic</th><th>Prob F-Statistic</th><th>R-squared</th><th>Adjusted R-squared</th></tr>
<tr><td>C</td><td>115.3087</td><td>12.6668</td><td>0</td><td rowspan="2">0</td><td rowspan="2">0.9864</td><td rowspan="2">0.9837</td></tr>
<tr><td>HDI?</td><td>-1.545</td><td>0.2323</td><td>0</td></tr>
<tr><th colspan="7">Model II
(independent variables are LER, MYS, EYS, PPP, while dependent variable is poverty)</th></tr>
<tr><th>Variable</th><th>Coefficient</th><th>Std. Error</th><th>Prob t-Statistic</th><th>Prob F-Statistic</th><th>R-squared</th><th>Adjusted R-squared</th></tr>
<tr><td>C</td><td>161.8074</td><td>32.4655</td><td>0</td><td rowspan="5">0</td><td rowspan="5">0.9881</td><td rowspan="5">0.9854</td></tr>
<tr><td>LER?</td><td>-0.7118</td><td>0.4548</td><td>0.12</td></tr>
<tr><td>MYS?</td><td>-1.7131</td><td>0.3848</td><td>0</td></tr>
<tr><td>EYS?</td><td>-1.5283</td><td>0.3043</td><td>0</td></tr>
<tr><td>PPP?</td><td>-0.0092</td><td>0.0023</td><td>0</td></tr>
</table>

Note : all statistic test using significance level of 5 percent

Based on Table 1, we get the information that the human development has influence on poverty reduction in Papua Province. This information is reflected on both models. **Model I** used to analyze the influence of HDI on poverty. Table shows that adjusted R-squared is 0,9837. This result implies that on the average about 98,37 percent of variance in poverty is explained by changes in HDI. This model also have probability value of less than 5 percent simultanously (F-statistic = 0.000). This output indicates that HDI have influence on poverty statistically. The regression result also shows that the impact of HDI on poverty is negative. That result showed by coefficient of HDI is -1.54. The result given in Model I implies that, if HDI increases is 1 percent, poverty decreases approximately 1,54 percent.

Model II used to analyze the effect of HDI's indicators (LER, MYS, EYS, PPP) on poverty. Table shows that adjusted R-squared is 0,9854. This result implies that on the average about 98,54 percent of variance in poverty is explained by changes in HDI's indicators. This model also have probability value of less than 5 percent simultanously (F-statistic = 0.000). This output indicates that HDI's indicators have influence on poverty statistically.

If seen partially, all variables have influence on poverty at 95 percent confidence level, except LER. The regression result also shows that the impact all of HDI's indicators on poverty are negative. That table show that coefficient of MYS, EYS, and PPP respectively is -1,71; -1,52; -0,009. The result given in Model II implies that, if MYS increases is 1 percent, poverty decreases approximately 1,71 percent; if EYS increases is 1 percent, poverty decreases about 1,52 percent; and if PPP increases is 1 percent, poverty decreses about 0,009 percent.

Based on the result of both of models, that human development can be reduce poverty in Papua. These results are conform with the study Hidayat (2008) which reveal that increasing the human development can be decreasing the poverty. Increasing the accessibility to meet the quality of education and increase the purchasing power parity can reduce poverty. The good education will be ensure a person to have the ability of individuals that are reliable so that it has a lot of opportunities and a high power bargaining for search for jobs, the good purchasing power can be increase the ability of individuals in fullfil their needs. And the finally, these conditions will be create wealth for the peoples, and the number of poor people in Papua will decline.

5. CONCLUSIONS

The above empirical analysis clearly revealed that human development index and HDI's indicators have influence on the reduction of poverty. The effect of human development index and HDI's indicators on poverty reduction is found significant in this paper, except life expectancy rate (LER). This study also provides the information that the biggest contribution on poverty reduction in Papua is mean years of schooling. And the smallest contribution is purchasing power parity.

The principal contribution of this paper are to give new perspective for the government and for researchers that the development in Papua is not only about the economic and physical infrastructure but also about the human. All stakeholders have to ensure that the people in Papua are ready for all the development. It is essential to increase Human Development Index in Papua especially the component of education and decent standard of living. For the government, this paper provides a new reference about which indicators in component of education have the biggest contribution on poverty reduction, so that need to be place in the first priority. In other words, the facilities at schools, the number of teachers, the capability of the teachers, as well as capacity building for the educators is exceptionally required. Besides that, the government also should be focus on efforts to increase the purchasing power of the poor through cash transfer and labour-intensive programme. With proper cash transfer and selective, it will assist the poor peoples to have much money, so the ability to buy goods and services will be increase. While, the labor-intensive program will be maintain the job availability for the poor people. Not only that, the government can also provide skills training or informal education, so they are will have the creativity that can be used to increase their income.

REFERENCES

Badan Pusat Statistik. (2015). Indeks Pembangunan Manusia 2014 Metode Baru. Badan Pusat Statistik, Jakarta

Baltagi, Badi H. (2005). Econometric Analysis of Panel Data, New York, John Wiley & Sons.

Brooks, C. (2008). Introductory Econometrics for Finance : Second Edition, New York, USA, Cambridge University Press.

Ekananda, M. (2005). Analisis Data Panel : Estimasi dengan Struktur Varians-Covarians Residual, Jakarta, Program Pasca Sarjana Universitas Indonesia.

Greene, William. (2003). Econometrics Analysis : Fifth Edition, USA, Prentice Hall.

Gujarati, Damodar. (2004). Basic Econometrics : Fourth Edition, Mc. Graw Hill Companies.

Hidayat, N. K. (2008). Analisis Hubungan Komponen Indeks Pembangunan Manusia dengan Kemiskinan di Provinsi Jawa Barat, Bogor, Program Studi Ekonomi Pertanian dan Sumberdaya

Park, H. M. (2011). Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata. International University of Japan, Japan

Singh, R. (2012). Human Development Index and Poverty Linkage, International Journal of Marketing and Technology, Vol. 2, No. 5, (2012), page 219-230.

United Nations Development Programme. (1990). Human Development Report 1990, Oxford University Press, New York, USA

United Nations Development Programme. (2014). Human Development Report 2014, Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience, Communications Development Incorporated, Washington DC, USA

United Nations Statistics Division. (2005). Handbook on Poverty Statistics : Concepts, Methods and Policy Use. United Nations, New York

Wooldridge, Jeffrey M. (2002). Econometrics Analysis of Cross Section and Panel Data, London, MIT Press Cambridge, Massachussets.

World Bank. (2001). World Development Report 2000/2001: Attacking Poverty. Oxford University Press, New York

CPS30:
Macrofinancial Statistics (5)

Implementation of Cox Proportional Hazard in Discontinuities Payment at Risk Management of Insurance Premiums

Jazi Munjazi *
Universitas Gadjah Mada, Yogyakarta, Indonesia – munjazzy@gmail.com
Farhan Trunna Mahadika
Universitas Gadjah Mada, Yogyakarta, Indonesia – farhantrunnamahadika@gmail.com
Danardono
Universitas Gadjah Mada, Yogyakarta, Indonesia – danardono@ugm.ac.id
Danang Teguh Qoyyimi
Universitas Gadjah Mada, Yogyakarta, Indonesia – qoyyimi@ugm.ac.id

Abstract

The risk is the possibility that something unpleasant or unwelcome will happen. Necessary companies that are able to reduce the risk, for example insurance companies, both individual or corporate business customers. This is due to an insurance company attempt to account for the risks that would be faced by customers. In reducing a person's risk, the insurance company also has the risk that the customers do not continue to pay the insurance premiums. Therefore, the insurance company needs to know the behavior of their costumers to determine the value of risks. Survival analysis using cox proportional hazard model is one of the methods to analyze risk of discontinuities payment of insurance premiums. In this case the research data is being analyzed from insurance products of PT Asuransi Jasa Indonesia (Jasindo) Yogyakarta branch.
It was previously assumed that customer's category, the price of insurance, vehicle's brand, and vehicle's model year are variables factors that affect the continuity payment of insurance premiums. Based on Cox proportional hazard analysis, only customer's category which affecting the continuity payment of insurance premiums. Leasing customers have greater discontinuities payment risk of insurance premiums 2.79 times compared with direct customers, banking customers have greater discontinuities payment risk of insurance premiums 1.95 times compared with direct customers, and leasing customers have greater discontinuities payment risk of insurance premiums 1.42 times compared with banking customers. Thus, leasing customer's category are the most customers at discontinuities payment risk of insurance premiums, followed by banking customer' category and then direct customer's category.

Keywords: Cox proportional hazard, Discontinuities payment risk, Insurance premiums.

1. INTRODUCTION

The risk is the possibility that something unpleasant or unwelcome will happen. Risk management is an important part to reduce or eliminate the risk of certain kinds of events happening. Risk management is a process for identifying, assessing, and prioritizing risks of different kinds. Insurance is a risk treatment option which involves risk sharing. It can be considered as a form of contingent capital and is akin to purchasing an option in which the buyer pays a small premium to be protected from a potential large loss.

Insurance risk is often taken by insurance companies, both individual or corporate business customers. Who then bear a pool of risks including market risk, credit risk, operational risk, interest rate risk, mortality risk, longevity risk, etc. Its has the risk that the customers do not continue to pay the insurance premiums. Therefore, the insurance company needs to know the behavior of their costumers to determine the value of risks. Based on this, the insurance companies could make certain policies so can continue to survive the competition between insurance companies.

There are various type of insurance such as life insurance, general insurance, and reinsurance. In this case the research data is being analyzed from general insurance, especially vehicle insurance at PT Asuransi Jasa Indonesia (Jasindo) Yogyakarta branch. Survival analysis using cox proportional hazard model is one of the methods to analyzed risk of discontinuities payment of insurance premiums in this case.

2. RISK MANAGEMENT AND INSURANCE

Risk is a word that usually has a negative connotation that something unpleasant or unwelcome will happen. The definition of risk itself is different, such as degree of uncertainty occurence thing or deviation from an expectation or harmful events at a certain period. The type of risk is classified into some dimension, such as pure risk versus speculative risk, subjective risk versus objective risk, and static risk versus dinamic risk. Risk management is the act or practice of dealing with risk. It include planning for risk, assessing (identifying and analyzing) risk areas, developing risk-handling option, monitoring risks, to determine how risks have changed, and documenting the overall risk management program.

Insurance is a risk treatment option which involves risk sharing. Insurance business divided into two broad categories, which are life insurance and property and casualty insurance or general insurance. Primary insurance and reinsurance are another way to classify insurance business. Life insurance covers term life, whole life, endowment, universal life, etc. General insurance covers property(fire, flood, earthquake, windstorm), casualty(general liability, third party damage, employers liability). Therefore reinsurance covers treaty and facultative.

Insurance can be considered as a form of contingent capital and it is similiar to purchasing an option in which the buyer pays a small premium to be protected from a potential large loss. An insurance premium is the amount of money that an individual or business must pay for an insurance policy. The insurance premium is considered income by the insurance company once it is earned, and also represents a liability in that the insurer must provide coverage for claims being made against the policy.

3. SURVIVAL ANALYSIS

Survival analysis and duration models originate in biostatistics, in which the survival time is the time until death or until the relapse of an illness. In recent years, these techniques have also gained popularity in social sciences to model the length of unemployment spells and strike duration. A comprehensive overview of the methods and models used in survival analysis is given by Therneau and Grambsch (2000) and by Klein and Moeschberger (2005).

Let the random variable T denotes the *survival time*. The distribution function of T is defined by the equation

$$F(t)=P(T<t) \tag{1}$$

and measures the probability of survival up to time t. Because T is a continous random variable, its density function $f(t)$ can be computed as the first derivative of the distribution function.

The *survival function* $S(t)$ denotes the probability of surviving until t or longer and is given by

$$S(t)=P(T\geq t)=1\text{-}F(t) \tag{2}$$

The limit

$$h(t)=\lim_{\delta\to 0}\frac{P(t\leq T<t+\delta|T\geq t)}{\delta} \tag{3}$$

represents the risk or proneness to death at time t. The function $h(t)$ is usually called the (*instantaneous*) *hazard function* or *the failure rate* and measures the instantaneous death rate given survival until time t. Greater values of the hazard function can also be interpreted as higher potential for the event to occur.

Because any of the functions $F(t)$, $S(t)$, $h(t)$ may be expressed with the help of any of the remaining three functions, one may decide to model any one of them and estimate the others from the derived equations.

4. COX PROPORTIONAL HAZARD MODEL

The Cox proportional hazards model is a semi-parametric method of analyzing the effects of different covariates on the hazard function. A detailed discussion of the Cox models can be found in Kleinbaum (2005) and in Hosmer and Lemeshow (2003). Assuming and individuals under observation, the Cox proportional hazards model is of the form

$$h_i(t|x) = e^{x_i'\beta}.\,h_0(t) = c_i.\,h_0(t), i = 1,2, \dots , n, \tag{4}$$

In which $x_i=(x_{i1},x_{i2},\ldots,x_{ik})'$ is the vector of k covariate values for individual i, $\beta=(\beta_1,\beta_2,\ldots,\beta_k)^\wedge{}'$ is the vector of regression coefficients, $h_i\,(t|x)$ is the hazard function of individual i, and $h_0\,(t)$ is the *baseline hazard*. Thus, the baseline hazard corresponds to an observation with $x_i=0$. The effect of the covariates on the hazard function in the Cox proportional hazards model does not depend on time because the ratio $\frac{h_i\,(t)}{h_0\,(t)}$ is equal to the constans c_i. Consequently, the baseline hazard determines the shape of the hazard function. The ratio of the hazard functions of individuals i and j, namely $HR=\frac{h_i\,(t)}{h_j\,(t)}$, is called the hazard ratio. This quotient is equal to

$$HR=\frac{h_i\,(t)}{h_j\,(t)}=\frac{e^{x_i'\beta}.h_0\,(t)}{e^{x_j'\beta}.h_0\,(t)}=e^{(x_i-x_j)'\beta} \tag{5}$$

The hazard ratio is the ratio of covariate effects for both individuals and is thus independent of time. This is called the proportional hazards assumption. The interpretation of the hazard ratio is similar to the odds ratio interpretation for logistic regression. A hazard ratio lower than 1 indicates decreased risk, whereas a ratio higher than 1 signals increased risk. Suppose that the vectors of covariates x_i and x_j differ only in the value of the p-th covariate and only for one unit. In this case, the hazard ratio

$$HR=\frac{h_i(t)}{h_j(t)}=e^{\beta_p} \tag{6}$$

measures the change of the hazard function for a unit change in the p-th covariate (if the covariate is a numerical variable). The hazard ratio is said to be statistically significant at the given level, when its confidence interval excludes 1. In this case, the null hypothesis that the variable is not related to survival can be rejected. This is the basis for the interpretation of the Cox regression results. By using Cox's partial likelihood estimator, it is possible to estimate the parameter vector without specifying and estimating the baseline hazard.

5. DISCUSSIONS

The data is being analyzed in this research from vehicle insurance of PT Asuransi Jasa Indonesia (Jasindo) Yogyakarta branch in 2013-2015. Honda is vehicle's brand which most dominant insured with the percentage of 75.1%. Banking customers and leasing customers respectively 16.3% and 4.2% insurance customers, the rest of customers are direct customers. Throughout 2012 to 2015, only 32.9 % vehicle is insured in the same year when it is manufactured. It was previously assumed that customer's category , the price of insurance, vehicle's brand, and vehicle's model year are variables factors that affect the continuity payment of insurance premiums. Consumer category consists of three categories, which are direct customer (reference category), leasing customer and banking customer. The price of insurance is vehicle price value. Vehicle's brand consist of Honda (reference category), Kawasaki, Suzuki, Yamaha, and Others. Vehicle's model year is year when vehicle was manufactured. With that assume, these survival analysis using Cox proportional hazard model with software R,

Table 1. Wald Chi-Square Test Summaries

Model	DF	Wald Chi-Square	Critical Value ($\chi^2_{\alpha:n}$)	P-value	Conclution
1	8	25.35	15.5073	0.001356	Model significant
2	7	25.33	14.0671	0.0006636	Model significant
3	6	25.23	12.5916	0.0003101	Model significant
4	2	22.94	5.99146	1.05E-05	Model significant

From Table 1, the result all Wald chi-square test are greater than critical value or p-value < α (0.05), it means all model are significant. Furthermore, each parameter model can be analyzed by Wald z partial test.

Cox proportional hazard model 1 is obtained from variable customer's category , the price of insurance, vehicle's brand, and vehicle's model year,

Table 2. Partial Test Cox PH Model 1

Variable	P-value (> |z|)	Conclution
Customer's category : leasing customer	0.005895	Parameter significant
Customer's category : banking customer	0.000268	Parameter significant
Vehicle's brand: Kawasaki	0.799752	Parameter insignificant
Vehicle's brand: Suzuki	0.565806	Parameter insignificant
Vehicle's brand: Yamaha	0.340226	Parameter insignificant
Vehicle's brand: Others	0.627092	Parameter insignificant
Vehicle's model year	0.740029	Parameter insignificant
The price of insurance	0.945825	Parameter insignificant

Table 2 shows that parameter of customer's category is both leasing customers or banking customers are significant. But the other parameters are insignificant. Because the price of insurance has the greatest p-value, so it is excluded from model.

Cox proportional hazard model 2 is obtained from variable customer's category , vehicle's brand, and vehicle's model year,

Table 3. Partial Test Cox PH Model 2

Variable	P-value (> |z|)	Conclution
Customer's category : leasing customer	0.005884	Parameter significant
Customer's category : banking customer	0.000239	Parameter significant
Vehicle's brand: Kawasaki	0.787882	Parameter insignificant
Vehicle's brand: Suzuki	0.340820	Parameter insignificant
Vehicle's brand: Yamaha	0.627039	Parameter insignificant
Vehicle's brand: Others	0. 337711	Parameter insignificant
Vehicle's model year	0.745180	Parameter insignificant

Table 3 shows that parameter of customer's category both leasing customers or banking customers are significant. But the other parameters are insignificant. Because vehicle's model year has the greatest p-value, so it is excluded from model.

Cox proportional hazard model 3 is obtained from variable customer's category and vehicle's brand,

Table 4. Partial Test Cox PH Model 3

Variable	P-value (> |z|)	Conclution
Customer's category: leasing customer	0.006046	Parameter significant
Customer's category: banking customer	0.000129	Parameter significant
Vehicle's brand: Kawasaki	0.800075	Parameter insignificant
Vehicle's brand: Suzuki	0.359497	Parameter insignificant
Vehicle's brand: Yamaha	0.625707	Parameter insignificant
Vehicle's brand: Others	0.350894	Parameter insignificant

Table 4 shows that parameter of customer's category is both leasing customers or banking customers are significant. But the other parameters are insignificant. Because vehicle's brand has the greatest p-value, so it is excluded from model.

Cox proportional hazard model 4 is obtained from variable customer's category only,

Table 5. Partial Test Cox PH Model 4

Variable	P-value (> |z|)	Conclution
Customer's category: leasing customer	0.000725	Parameter significant
Customer's category : banking customer	0.000128	Parameter significant

Table 5 shows that parameter customer's category is both leasing customers or banking customers have p-value < α (0.05), it means all parameter significant including to the model.

$$h(t|x) = h_0(t)\exp\,(1.0243Leasing + 0.0.6702Banking) \tag{7}$$

From the model is obtained risk factor of customer's category or discontinuities payment risk of insurance premiums based on customer's category .

$$\begin{aligned} HR_1 &= \frac{h(t|Leasing\ customer)}{h(t|Direct\ customer)} \\ &= \frac{h_0(t)\exp(1.0243(1) + 0.0.6702(0))}{h_0(t)\exp(1.0243(0) + 0.0.6702(0))} \\ &= 2.79 \end{aligned} \tag{8}$$

Leasing customers have greater discontinuities payment risk of insurance premiums 2.79 times compared with direct customers.

$$\begin{aligned} HR_2 &= \frac{h(t|Banking\ customer)}{h(t|Direct\ customer)} \\ &= \frac{h_0(t)\exp(1.0243(0) + 0.0.6702(1))}{h_0(t)\exp(1.0243(0) + 0.0.6702(0))} \\ &= 1.95 \end{aligned} \tag{9}$$

Banking customers have greater discontinuities payment risk of insurance premiums 1.95 times compared with direct customers.

$$\begin{aligned} HR_3 &= \frac{h(t|Leasing\ customer)}{h(t|Banking\ customer)} \\ &= \frac{h_0(t)\exp(1.0243(1) + 0.0.6702(0))}{h_0(t)\exp(1.0243(0) + 0.0.6702(1))} \\ &= 1.42 \end{aligned} \tag{10}$$

Leasing customers have greater discontinuities payment risk of insurance premiums 1.42 times compared with banking customers. Thus, leasing customer's category is the most customers at discontinuities payment risk of insurance premiums, followed by banking customer's category and then direct customer's category.

6. CONCLUSIONS

From All customers of Jasindo Yogyakarta in 2012 until 2015, Honda is vehicle's brand the most dominant insured with the percentage of 75.1%. Banking customers and leasing customers respectively 16.3% and 4.2% insurance customers, the rest of customer's category is direct customers. Throughout 2012 to 2015, only 32.9 % vehicle insured in the same year when it is manufactured. It was previously assumed that customer's category , the price of insurance, vehicle's brand, and vehicle's model year are variables factors that affect the continuity payment of insurance premiums. Leasing customers have greater discontinuities payment risk of insurance premiums 2.79 times compared with direct customers, banking customers have greater discontinuities payment risk of insurance premiums 1.95 times compared with direct customers, and leasing customers have greater discontinuities payment risk of insurance premiums 1.42 times compared with banking customers. Thus, leasing customers category are the most customers at discontinuities payment risk of insurance premiums, followed by customer's category and then banking customers direct category. Because many of customers insured vehicles with vehicle's brand Honda, Jasindo could cooperate with these vehicle companies.

REFERENCES

Cox, DR. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*. B, Vol. 34, No. 2. Pp. 187–220

George E. Rejda. 1995. *Principle of Risk Management and Insurance*. HarperCollins College Publishers

Hosmer, D.W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis*. Wiley & Sons, Inc.

Klienbum, D.G. dan Klien, M. (2005). *Survival Analysis, A Self-Learning Text*. Springer Science Bussiness Media, Inc.

Therneau, Terry M., Grambsch, Patricia M. (2000). *Modelling Survival Data:Extending the Cox Model*. Springer Science Bussiness Media, Inc., USA.

Trieschman, Gustavson, Hoyt. 2001. *Risk Management and Insurance. S*outh-Western College Publishing

Economic Policy Uncertainty and Financial Market Volatility: Evidence from Japan

Takayuki Morimoto*
Kwansei Gakuin University, Sanda, Japan – morimot@kwansei.ac.jp

Abstract

In this study, we show a relationship between economic policy uncertainty and financial market volatility in Japanese financial market. Uncertainty is measured by the index of economic policy uncertainty (EPU) based on newspaper coverage, frequency newly developed by Baker et al. Volatility is calculated as a sum of squared intraday returns, which is known as the realized volatility (RV). The EPU and RV are combined with the mixed data sampling (MIDAS) approach in order to investigate how economic policy uncertainty shocks are associated with the Japanese financial market volatility. The result will contribute to financial market research and economic policy studies.

Keywords: Economic policy uncertainty index; Realized volatility; GARCH-MIDAS model; DCC-MIDAS model; Japanese financial market.

1. INTRODUCTION

Asgharian et al. (2016) investigate US and UK stock market movements using the economic policy uncertainty indices of Baker et al. (2016) in combination with the mixed data sampling (MIDAS) approach. They find that The long-run US-UK stock market correlation depends positively on US economic policy uncertainty shocks while the US long-run stock market volatility depends significantly on the US economic policy uncertainty shocks but not on UK shocks while the UK depends significantly on both.

In this research, we follow Asgharian et al. (2016) and apply their method to Japanese stock market. Specifically, we investigate the relation between Nikkei225 which is the stock index for the Tokyo Stock Exchange (TSE) and individual stocks comprised in TOPIX100 which is composed of Top 100 stocks traded on TSE in light of economic policy uncertainty and stock market volatility. Uncertainty is measured by the index of economic policy uncertainty (EPU) based on newspaper coverage, frequency newly developed by Baker et al. (2016) Volatility is calculated as a sum of squared intraday returns, which is known as the realized volatility (RV). The EPU and RV are combined with the mixed data sampling (MIDAS) approach proposed by Ghysels et al. (2004, 2006) in order to investigate how economic policy uncertainty shocks are associated with the Japanese financial market volatility.

Meanwhile, Engle et al. (2013) use the MIDAS approach to link macroeconomic variables to the long-term component of volatility. They incorporate a mean reverting unit daily heteroscedastic volatility process with a MIDAS polynomial that applies to long-term macroeconomic variables, which is called the generalized autoregressive conditional heteroscedasticity model with MIDAS (GARCH-MIDAS) approach. Now we replace a macroeconomic variable with a monthly EPU in GARCH-MIDAS model following Asgharian et al. (2016). Furthermore, Colacito et al. (2011) introduce a novel component model for dynamic correlations which is called the dynamic conditional correlation model with MIDAS (DCC-MIDAS) approach. DCC-MIDAS model is a natural extension of GARCH-MIDAS model to DCC model advocated by Engle (2002). We also use DCC-MIDAS model to capture the dynamic correlation of volatilities between the market index and individual stocks in TSE.

The EPU index of Japan which can be downloaded on the web site: www.policyuncertainty.com is based on frequency counts of articles in Japan's newspapers, Asahi and Yomiuri. It counts the number of news articles containing the terms uncertain or uncertainty, and one or more policy terms. Policy terms are the Japanese equivalents of `tax', `policy', `spending', `regulation', etc. To capture `spending' by the government, they use a set of four terms: `saishutsu', `kokyo jigyohi', `kokyo toushi', and `kokuhi', see the web site for more details.

Our specification employs monthly EPU index of Japan as an explanatory variable in the variance equation of a unit daily GARCH-MIDAS model, which we refer to the model as GARCH-MIDAS-EPU. In our empirical analysis, we first estimate the parameters the GARCH-MIDAS-EPU model pair of two stock returns. After that, we obtain the estimated DCC-MIDAS parameters with the standardized residuals from the GARCH-MIDAS-EPU model using the quasi-likelihood method.

2. MODELS

In this section, we briefly introduce GARCH-MIDAS-EPU and DCC-MIDAS models which are mentioned above, following Colacito et al. (2011), Asgharian et al. (2016) and Conrad et al. (2014).

Let us assume that the vector of returns $\mathbf{r}_t = [r_{1,t}, \ldots, r_{n,t}]'$ follows the process:

$$\mathbf{r}_t \sim N(\mu, H_t)$$
$$H_t = D_t R_t D_t$$

where μ is the vector of unconditional means, H_t is the conditional covariance matrix and D_t is a diagonal matrix with standard deviations on the diagonal. Furthermore, we also assume that:

$$R_t = E_{t-1}[\xi_t \xi_t']$$
$$\xi_t = D_t^{-1}(\mathbf{r}_t - \mu)$$

where $E_{t-1}[\cdot]$ is the expectation at time t - 1 given the observations until time t - 1. Then we have $\mathbf{r}_t = \mu + H_t^{\frac{1}{2}} \xi_t$ with $\xi_t \sim N(0, I_n)$. We refer to g_i and m_i as the short and long run variance components respectively for asset i and denote by N_v^i the number of days that m_i is held fixed. The superscript i indicates that this may be asset-specific and the subscript v differentiates it from a similar scheme that will be introduced later for correlations. In particular, while $g_{i,t}$ moves daily, $m_{i,\tau}$ changes only once every N_v^i days. We assume that for each asset i = 1 . . . . , n, univariate returns follow the GARCH-MIDAS process:

$$r_{i,t} = \mu_i + \sqrt{m_{i,\tau} \cdot g_{i,t}} \xi_{i,t}, \quad \forall t = \tau N_v^i, \ldots, (\tau+1) N_v^i$$

where $g_{i,t}$ follows a GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha_i - \beta_i) + \alpha_i \frac{(r_{i,t-1} - \mu_i)^2}{m_{i,\tau_t}} + \beta_i g_{i,t-1}$$

while the MIDAS component $m_{i,\tau}$ is a weighted sum of K_v^i lags of realized variances (RV) over a long horizon:

$$m_{i,\tau} = \bar{m}_i + \theta_i \sum_{l=1}^{K_v^i} \varphi_l(\omega_v^i) RV_{i,\tau-l}$$

where the RV involve N_v^i daily squared returns. Namely:

$$RV_{i,\tau} = \sum_{j=(\tau-1)N_v^i+1}^{\tau N_v^i} (r_{i,j})^2$$

where N_v^i could for example be a quarter or a month. The above specification corresponds to the block sampling scheme as defined in Engle et al. (2013), involving so called Beta weights defined as:

$$\varphi_l(\omega_v^i) = \frac{(1 - l/K_v^i)^{\omega_v^i - 1}}{\sum_{j=1}^{K_v^i} (1 - j/K_v^i)^{\omega_v^i - 1}}$$

where the parameters N_v^i and K_v^i are independent of i, i.e. the same across all series.

We use the two-step DCC-MIDAS model of Colacito et al. (2011) extended to allow for exogeneous variables influencing the long-run volatility and correlation as in Asgharian et al. (2016). The first step consists of estimating separate GARCH-MIDAS models for the stock returns for day $i = 1, \ldots, N_t$ in month t as:

$$r_{i,t-1} = \mu + \sqrt{\tau_t g_{i,t} \varepsilon_{i,t}} \quad \varepsilon_{i,t} \sim (0,1)$$

where the total stock variance $\sigma_{i,t}^2$ is separated into a short-run component $g_{i,t}$ and a long-run component τ_t such that $\sigma_{i,t}^2 = \tau_t g_{i,t}$. A GARCH (1,1) process describes the short-run component:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha \frac{(r_{i,t-1} - \mu)^2}{\tau_t} + \beta g_{i-1,t}$$

where $\alpha > 0$ and $\beta \geq 0$, $\alpha + \beta < 1$. The long-run component is described by a MIDAS regression where the lagged EPU shocks of the EPU_{t-k} are included over $k = 1, \ldots, 24$:

$$\tau_t = \theta_0 + \theta_1 \sum_{k=1}^{K} \varphi_k EPU_{t-k}$$

where the weighting scheme is described by a beta lag polynomial:

$$\varphi_l(\omega_v^i) = \frac{\left(1 - l/K_v^i\right)^{\omega_v^i - 1}}{\sum_{j=1}^{K_v^i} \left(1 - j/K_v^i\right)^{\omega_v^i - 1}}$$

where the parameter θ_1 measures the effects of the economic policy uncertanity shocks on the long-run volatility. We fix $w_1 = 1$ to ensure higher weights to the most recent observations as with Asgharian et al. (2016).

Colacito et al. (2011) propose the DCC-MIDAS model which is a natural extension of the GARCH-MIDAS model to the Engle (2002) DCC model. Using the standardized residuals $\xi_{i,t}$, it is possible to obtain a matrix Q_t whose elements are:

$$q_{i,j,t} = \bar{\rho}_{i,j,t}(1 - a - b) + a\xi_{i,t-1} + bq_{i,j,t-1}$$

$$\bar{\rho}_{i,j,t} = \sum_{l=1}^{K_c^{ij}} \varphi_l \left(\omega_r^{ij}\right) c_{i,j,t-l}$$

$$c_{i,j,t} = \frac{\sum_{k=t-N_c^{ij}}^{t} \xi_{i,k}\xi_{j,k}}{\sqrt{\sum_{k=t-N_c^{ij}}^{t} \xi_{i,k}^2}\sqrt{\sum_{k=t-N_c^{ij}}^{t} \xi_{j,k}^2}}$$

where we could have used simple cross-products of $\xi_{i,t}$ in the above formulation of $c_{i,j,t}$. The normalization allow us to discuss regularity conditions in terms of correlation matrices. Correlations can then be computed as:

$$\rho_{i,j,t} = \frac{q_{i,j,t}}{\sqrt{q_{i,i,t}}\sqrt{q_{j,j,t}}}$$

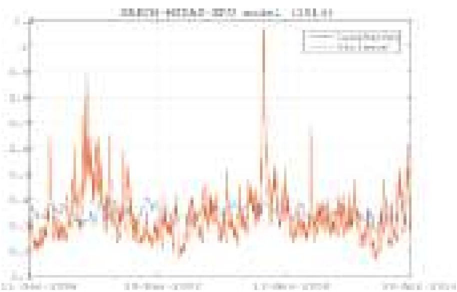
where we regard $q_{i,j,t}$ as the short run correlation between assets i and j, whereas $\bar{\rho}_{i,j,t}$ is a slowly moving long run correlation. Rewriting the first equation of system as

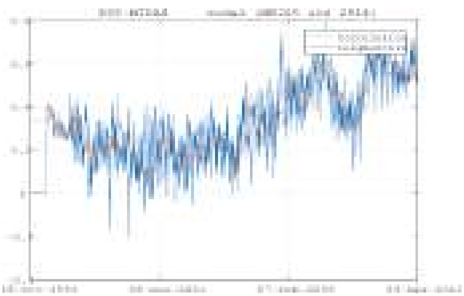
$$q_{i,j,t} - \bar{\rho}_{i,j,t} = a\left(\xi_{i,t-1}\xi_{j,t-1} - \bar{\rho}_{i,j,t}\right) + b\left(q_{i,j,t-1} - \bar{\rho}_{i,j,t}\right)$$

conveys the idea of short run fluctuations around a time-varying long run relationship. The idea captured by the DCC-MIDAS model is similar to that underlying GARCH-MIDAS. In the GARCH-MIDAS the short run component is a GARCH component, based on daily returns, that moves around a long-run component driven by realized volatilities computed over a monthly basis, see Colacito et al. (2011).

3. EMPIRICAL ANALYSIS

We apply the DCC-MIDAS with GARCH-MIDAS-EPU model to Nikkei225 and TOPIX100 data listed on TSE from June 1988 to April 2016 in order to investigate the relation between economic policy uncertainty and financial market volatility in Japanese financial market. Here is an example of the results of our empirical analysis. Figures below show the plots of estimated short- and long-run variances and correlations for Japan Tobacco Inc. (2914) and Nikkei225.





Graph 1. The Plots of Estimated Short- and Long-run Variances and Correlations for Japan Tobacco Inc. (2914) and Nikkei225.

4. CONCLUSIONS

The result our empirical analysis will contribute to financial market research and economic policy studies.

REFERENCES

Asgharian, H., Christiansen, C., Gupta, R., and Hou, A. J. (2016). Effects of Economic Policy Uncertainty Shocks on the Long-Run US-UK Stock Market Correlation (October 3, 2016). Available at SSRN: https://ssrn.com/abstract=2846925 or http://dx.doi.org/10.2139/ssrn.2846925.

Baker, S.R., Bloom, N., and Davis, S.J. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131: 1593-1636.

Colacito, R., Engle, R.F., and Ghysels, E. (2011). A Component Model for Dynamic Correlations. Journal of Econometrics 164: 45-59.

Conrad, C., Loch, K., and Rittler, D. (2014). On the macroeconomic determinants of long-term volatilities and correlations in US stock and crude oil markets. Journal of Empirical Finance 29: 26-40.

Engle, R. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. Journal of Business and Economic Statistics 20: 339-350.

Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. The Review of Economics and Statistics 95: 776-797.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models, CIRANO Working Paper 2004s-20.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. Journal of Econometrics 131: 59-95.

This page intentionally left blank

CPS31:
Macroeconomic Statistics (6)

Nowcasting Household Consumption and Investment

Tarsidin, Idham*, Robbi Nur Rakhman*
Bank Indonesia, Jakarta, Indonesia – idham_a@bi.go.id or robbi_nurrakhman@bi.go.id

Abstract

It is imperative for the Central Bank to know the current state of the economy as the basis underlying projections of future economic conditions. To that end, current economic conditions, in this case household consumption and investment, could be predicted using nowcasting. In this research, a nowcasting model was developed for the two aforementioned macroeconomic variables using a Dynamic Factor Model (DFM). The indicators used when nowcasting household consumption included: motor vehicle sales, total deposits, the lending rate on consumer loans, M1 and the rupiah exchange rate (NEER), while the indicators used for nowcasting investment included: cement sales, motor vehicle production, electric energy consumption, outstanding loans and M1. Accuracy testing showed that the nowcasting model for household consumption using DFM was sound, while the forecast error for nowcasting investment was significant but remained below the benchmark.

Keywords: Nowcasting, Mixed Frequency Regression, Dynamic Factor Model
JEL Classification: C38, C53

1. INTRODUCTION

According to the forecasting and policy analysis (FPAS) framework, Bank Indonesia utilises the ARIMBI core model and several satellite models, such as SOFIE, MODBI, ISMA AND BIMA. The five models are fundamentally macroeconomic models used for short-term (up to two years) and medium-term (2-5 years) projections. In addition, Bank Indonesia also has several indicator models, including GDP, inflation and exchange rate models, which are used to produce near-term forecasts (for the current quarter and subsequent period) as well as nowcasting (current quarter).

Indicator models are crucial considering the importance of the central bank knowing the current state of the economy as the basis underlying projections of future economic conditions in line with international best practices at other central banks, including the Riksbanken and Bank of England.[1] Furthermore, projections based on the indicator models are used as inputs for the macroeconomic models. Therefore, projections for the current quarter and subsequent period are fundamentally generated by the indicator models.

Considering the importance of indicator models, further development was required. Currently, Bank Indonesia only has indicator models for GDP, inflation and the exchange rate but is lacking models for GDP components, such as consumption, investment and exports/imports. This research developed an indicator model for nowcasting household consumption and investment. The model was expected to improve the accuracy of Bank Indonesia's economic projections and, therefore, support more precise monetary policymaking.

Nowcasting has developed rapidly over the past decade. The most commonly used methods include: Bridge Equation, MIxed DAta Sampling (MIDAS) developed by Ghysels, et al. (2004)), Mixed Frequency VAR (MF-VAR developed by Mariano and Murasawa (2010) as well as Schorfheide and Song (2013)) and the Dynamic Factor Model (DFM, otherwise known as the Mixed Frequency Factor Model (MF-FM), developed by Mariano and Murasawa (2003) and Giannone, et al. (2008). Nowcasting GDP in Indonesia has been performed by Kurniawan (2014) using the MIDAS approach along with the Mixed Frequency Factor Model, as well as by Luciani, et al. (2015) using the Dynamic Factor Model. Nowcasting has been proven to effectively bolster economic assessments in Indonesia.

This research aimed to develop indicator models with a nowcasting function for household consumption and investment, for which the data is released quarterly by BPS-Statistics Indonesia. The model was developed using indicators for which monthly data is available as well as higher frequency data. The model was developed using the Dynamic Factor Model Method and data for the period of 2003-2015. Pseudo out-of-sample testing was also performed using 2015 data. The DFM method was selected because it is currently the most popular nowcasting approach utilised due to the inherent advantages.

1 As cited by Angelini, et al. (2008), Andersson and Reijer (2015) as well as Bell, et al. (2014) amongst others.

2. THE MODEL AND DATA

2.1. The Dynamic Factor Model

Nowcasting using the Factor Model is commonly used by central banks to predict current quarter GDP, including the Riksbanken[2], Norges Bank, European Central Bank (ECB)[3], Bank Indonesia[4] and by international organisations including ADB[5]. The capacity to accommodate numerous indicators ensures the popularity of the Factor Model. Nowcasting household consumption and investment is not dissimilar to nowcasting GDP. Therefore, the nowcasting method typically used for GDP can also be applied to nowcasting other lower-frequency variables (including household consumption and investment), as suggested by Bańbura, et al (2011b).

The preferred method in this research is the Mixed Frequency Dynamic Factor Model (abbreviated to DFM). There are a number of significant variations to the estimates and projections using DFM. In this research, the authors were inclined to follow the approach of Giannone, et al. (2008), developed by Bańbura and Rünstler (2011a) and modified by Kurniawan (2014). The model specification can be expressed as follows:

$$z_t = \Lambda \beta_t \qquad\qquad u_t \sim N(0, \Sigma_u) \tag{3.1}$$

$$\beta_t = F\beta_{t\;1} + \varepsilon_t \qquad\qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{3.2}$$

where:

z_t : consists of reference series (y_t) and monthly indicators (x_t)
f_t : common factors
Λ : factor loadings
u_t : idiosyncratic component
β_t : consists of common factors (f_t) and idiosyncratic component (u_t)
F : autoregressive cofficients
ε_t : white noise error

Equations 3.1 and 3.2 were estimated in two stages, known as the two-step estimator. The first stage involved estimating the representative parameters of the state space using the principal components (from the balanced panel of monthly indicators[6]), thus estimating the magnitude of the common factors. In the second stage, the common factors were re-estimated by applying a Kalman smoother to the entire information set. The reference series (household consumption and investment) was subsequently projected by regressing the latent variables in the form of year-on-year household consumption growth and year-on-year investment growth against the common factors.

2.1. Data

The data used in this research was for the period from 2003-2015. The estimations were based on data from 2003-2015 and the pseudo out-of-sample testing used 2015 data. The reference series were household consumption and investment, while the component series were various representative indicators selected from a range of candidate indicators. The reference series had a quarterly frequency and the data is published by BPS-Statistics Indonesia. Meanwhile, the indicators in the component series were mixed frequency, with the majority in the form of monthly data. Not all data for the indicators was available from 2003-2015. Therefore, the data was ragged data with differing availability across the period.

Since 2015, BPS-Statistics Indonesia has released data using 2010 as the base year for 2010-2015 data. Previous BPS data used 2000 as the base year. In a departure from the previous data, household consumption data using 2010 as the base year does not contain the consumption of non-profit institutions. Data with a base year of 2010 was applied in this research, therefore the data did not

2 Refer to Andersson and Reijer (2015).
3 Refer to Angelini, et al. (2008).
4 Refer to Kurniawan (2014).
5 Refer to Luciani, et al. (2015).
6 As mentioned by Bańbura, et al. (2012), x_t only contains observable monthly indicators, thus omitting quarterly series of household consumption and investment, as well as monthly latent series.

contain the consumption of non-profit institutions. For earlier years (2003-2009), however, backcasting was applied to the growth in order to obtain household consumption without the consumption of non-profit institutions. Similarly, for investment the data used had a base year of 2010, so for previous years backcasting was again applied.

3. STAGES OF NOWCASTING

3.1. Indicator Selection

Although the modelling was not required to find causality between the component series and reference series, this does not imply that any indicator could be inputted to the model. The indicators should have a close relationship with their reference series, in this case household consumption and investment (PMTB). OECD (2012) found that their economic relevance must be considered when selecting the indicators (economically significant and a broad scope). During the initial stage, several candidate indicators were selected that meet the requirements in terms of economic relevance and practical considerations. The candidate indicators were subsequently whittled down using several criteria in order to observe the correlation with the reference series (based on the coefficient of correlation) and similarity of the common factors with the reference series (based on Principal Component Analysis). Several indicators were thus selected for use in the estimations.

3.2. Filtering and Transformation

Prior to use in the estimations, the data must be filtered using seasonal adjustments with X-12 or TRAMO/SEATS. More robust estimations are possible by omitting seasonal factors from the reference series and component series. Nonetheless, when the data was used for projections, the prediction was the actual value that contains seasonal factors. Therefore, the results must be multiplied or added to the seasonal factors for comparison with the realised data. Another common approach is to exclude the seasonal adjustment. Therefore, direct comparisons with the realised data are possible. This approach is appropriate if similar seasonal factors are found in the reference series and component series. The estimations may be compromised, however, if the seasonal factors are different in the reference series and component series, leading to less robust estimations.

Considering that the data on household consumption and investment released by BPS-Statistics Indonesia has not had the seasonal factors removed, seasonal adjustments were preferred in this research. Data may be presented in the form of level, percentage, index or unit. Level data was based on constant prices (real value). Data in level, index and unit were transformed into year-on-year growth. Percentage data, on the other hand, was transformed into the difference with the previous year. The reference series and indicators were then standardised for comparison.

3.3. Nowcasting Exercise

As mentioned previously, the estimations were made using the Dynamic Factor Model and data for the period from 2003-2015. In this case, a nowcasting exercise was conducted systematically using various combinations of component series, namely consisting of 4, 5 and 6 indicators from the range of candidate indicators selected. The best model was chosen based on pseudo out-of-sample testing of the data from 2015 to predict household consumption and investment in each quarter of the year. The model with the smallest RMSE was selected, demonstrating that the prediction model was more accurate.

A more rigorous test was conducted for the fourth quarter of 2015, namely weekly predictions for the first, second and third months in line with data availability (released data). To that end, the approach of Camacho and Perez-Quiros (2010) was used to evaluate the results of the nowcasting exercise by using real-time conditions and revised data. Consequently, a real-time dataset was built consisting of historical data available for each of the nowcasting data points.

To enhance the prediction capabilities, Giannone, et al. (2008) as well as Bańbura and Rünstler (2011) proposed uncertainty measures (that show the marginal gain on the prediction precision) to assess the role of latest data releases. Furthermore, Bańbura and Rünstler (2011) stressed the need for different weightings on each indicator in every prediction period. Therefore, when using the prediction model, different indicators could be used or even large datasets (consisting of timely indicators and/or indicators containing important information).

3.4. Model Performance Evaluation

The next stage involves evaluating the performance of the selected model. The evaluation takes into account a comparison with the benchmark models, including the Bridge Equation and ARIMA model. A comparison with the Bridge Equation was chosen due to the simplicity of the model. Meanwhile, the ARIMA model is known for its near-term forecasting accuracy compared to naïve methods and other simple models.

4. RESULTS AND ANALYSIS

4.1. Indicator Selection

As mentioned previously, there were two requirements of indicator selection, namely economic relevance and practical considerations. Based on those considerations, a number of candidate indicators were selected for household consumption and investment that met both requirements, at least to some extent. From there, a number of candidate indicators were subsequently selected in order to observe their coefficient of correlation and contribution to explain the reference series for potential use in the estimations.

4.1.1. Indicators of Household

Household consumption is the largest component of GDP. During the observation period from 2003-2015, household consumption accounted for 54-63% of GDP. A number of indicators were found to have a high coefficient of correlation to household consumption as follows:

Indicators representing the magnitude of household consumption, including motor vehicle sales and the retail sales index;

Indicators reflecting consumer opinion of economic conditions that underlie their consumption decisions, including the consumer confidence index (composite, current economic condition and expectation) as well as consumer tendency index;

Economic performance indicators, which ultimately impact the level of household income and consumption, including the industrial production index, export YTD, import YTD, loading/unloading of cargo (through domestic and international ports) and the stock price index at the Indonesia Stock Exchange; and

Indicators related to banking, economic liquidity and exchange rates, which influence consumers' consumption decisions, including total deposit, credit (consumer and total), 1, 3 and 6-month term deposit rates, the consumer lending rate, M1, M2 as well as the Rupiah exchange rate against the US dollar and the effective exchange rate, both nominal and real.

Of the various candidate indicators mentioned, several indicators with the highest coefficient of correlation were selected. Indicators with a correlation coefficient value of 0.70 or more and a year-on-year growth correlation coefficient value of 0.15 or more as well as the correct correlation sign were selected for the evaluation stage. Based on those criteria, the following 11 indicators were selected: motor vehicle sales, retail sales index, consumer confidence index - current economic condition, industrial production index, stock price index at the Indonesia Stock Exchange, total deposit, credit, consumer lending rate, M1, M2 and nominal effective exchange rate (NEER). To reduce the potential number of indicator combinations, a number of indicators were discarded due to their close resemblance to other indicators (with a higher correlation coefficient).

We also consider some practical aspects of data like frequency, revision, and publication lag. All indicators have a monthly frequency, no (significant) data revisions and a publication lag of 7 days to 3 months. Despite a publication lag of 2-3 months, the industrial production index was included due to the important information contained in the indicator. A lag of three months was rare, with two months the norm. The candidate indicators selected based on the correlation of coefficient were subsequently analysed using Principal Component Analysis (PCA) to identify similarities between the common factors and household consumption. No evidence was found of certain indicators having an extremely close correlation with household consumption. Therefore, all the 11 indicators selected by coefficient of correlation analysis were used in the DFM estimation.

4.1.2. Indicators of Investment

Investment, or gross fixed capital formation (GFCF), is also a significant component of GDP. Although the contribution to GDP is not massive, in the range of 25-35% during the observation period of

2003-2015, investment plays a salient role in an economy. Investment consists of construction and non-construction investment (including machinery and equipment, vehicles, other equipment, cultivated biological resources and intellectual property products). Several indicators correlate closely with investment, which can be categorised as follows:
Indicators representing the size of investment, including capital spending (state budget), cement sales and motor vehicle production;
Indicators reflecting corporate and household expectations of economic conditions, which underlie their investment decisions, including the consumer confidence index (composite, current economic condition and expectation), selling prices (realization and expectation) as well as usage of labor (realization and expectation);
Indicators of corporate conditions, which indicate corporate ability to invest, including corporate financial conditions and access to credit;
Economic performance indicators, which ultimately influence corporate and household ability to invest, including the industrial production index, production capacity utilization, electricity consumption, export YTD, import YTD, loading/unloading of cargo (through domestic and international ports), market capitalization and stock price index at the Indonesia Stock Exchange; and
Indicators related to banking, economic liquidity and exchange rates, which influence corporate and household investment decisions, including credit (working capital, investment and total), lending rates (working capital and investment), M1, M2 as well as the Rupiah exchange rate against the US dollar and the effective exchange rate, both nominal and real..
Several indicators with the highest coefficient of correlation were selected from the various candidates. Indicators with a correlation coefficient value of 0.70 or more and a year-on-year growth correlation coefficient value of 0.15 or more as well as the correct correlation sign were selected for the evaluation stage. Based on those criteria, the following 14 indicators were selected: cement sales, motor vehicle production, the consumer confidence index - state of the economy, industrial production index, electricity consumption, export YTD, import YTD, credit (working capital, investment and total), lending rates (working capital and investment), M1 and the nominal effective exchange rate (NEER). To reduce the potential number of indicator combinations, a number of indicators were discarded due to their close resemblance to other indicators.
Similar to the household consumption, we also consider some practical aspects of data like frequency, revision, and publication lag. All indicators are shown to have a monthly frequency, no (significant) data revisions and a publication lag of 7 days to 3 months. Despite a publication lag of 2-3 months, the industrial production index was included. The candidate indicators selected based on the correlation of coefficient were subsequently analysed using Principal Component Analysis (PCA) to identify similarities between the common factors and household consumption. No evidence was found of certain indicators having an extremely close correlation with investment. Therefore, all the 14 indicators selected by coefficient of correlation criteria were used in the DFM estimation.

4.2. Nowcasting Exercise
An exercise was performed using various different combinations of indicators to obtain the best nowcasting model. Based on the results of pseudo out-of-sample testing, the model with the smallest RMSE was selected.

4.2.1. Household Consumption
Through the exercise, several combinations of the 11 selected indicators were tested, including 330 combinations of four indicators, 462 combinations of five indicators and 462 combinations of six indicators. The exercise produced a component series of the best indicators:

- motor vehicle sales
- total deposit
- the consumer lending rate
- narrow money (M1) and
- the Rupiah exchange rate (NEER)

Based on the RMSE, in the first month, the performance of the best combination of indicators was no more robust than several other combinations of indicators, but in the second and third months that combination ranked in first place. The selection of the best combination showed that the five aforementioned indicators had the highest resemblance of common factors to household consumption. Nonetheless, the results of the exercise also demonstrate that the difference in performance amongst the 10 best indicator combinations was almost negligible, evidenced by the small difference in RMSE. Therefore, a number of other indicators were also shown to be sufficiently robust for nowcasting household consumption.

The results of nowcasting household consumption show the RMSE value of 0.02 and MAPE of 0.29% (based on the evaluation in the third month). Nowcasting was performed for each quarter of 2015 and the exercise showed that the nowcasting model was adequately robust from the first month, with no significant gains in accuracy found in the subsequent months. In fact, the results in the third month of the second quarter were no more accurate than the previous month.

4.2.1. Household Consumption

Similar to household consumption, several combinations of the 11 selected indicators were tested in the exercise. In total, 1,254 combinations were tested, consisting of 330 combinations of four indicators, 462 combinations of five indicators and 462 combinations of six indicators. The exercise produced a component series of the best indicators:

- cement sales
- motor vehicle production
- electric consumption
- outstanding loans and
- narrow money (M1)

Based on the RMSE, the performance of the best combination of indicators ranked second in the first month but ranked first in the second and third months, while performing best in the second month when the RSME value was smallest. Nonetheless, the results of the exercise also demonstrate that the difference in performance amongst the 10 best indicator combinations was almost negligible, evidenced by the small difference in RMSE. Departing from the findings of the nowcasting exercise for household consumption, however, the RMSE of the investment nowcasting model was comparatively high at 1. The authors expect that this was linked to fluctuating investment data (in year-on-year growth). When nowcasting investment, the indicators of cement sales, motor vehicle production and electric consumption were nearly always selected as components of the best combination

The results of nowcasting investment show the RMSE value of 1.03 and MAPE of 15% (based on the evaluation in the third month). Nowcasting was performed for each quarter of 2015 and the exercise showed that the nowcasting model was more robust in the second month. In the third month, however, the results of the nowcasting exercise improved in the first and second quarters but actually deteriorated in the third and fourth quarters. Furthermore, a tendency to overshoot was also found when nowcasting the first and second quarters, contrasting the propensity to undershoot in the third and fourth quarters. The model was, however, the best one that could be produced. The proclivity to over- and undershoot must be taken into consideration when using the model in order to minimise the error or deviation.

4.3. Model Performance Evaluation

As mentioned previously, model performance was evaluated by comparing the nowcasting results to the benchmark models, in this case the Bridge Equation and ARIMA model. As presented in Table 4.1, the comparison of model accuracy for nowcasting household consumption showed that the forecast error of the Dynamic Factor Model was smaller than the forecasting error of the Bridge Equation and ARIMA. Therefore, nowcasting household consumption using the DFM model was shown to be the most robust.

Table 4.1. Comparison with Other Models – Household Consumption

Nowcasting		Realization	Nowcasting Result [and Error/ Deviation]					
			Dynamic Factor Model		Bridge Equation		ARIMA	
2015	I	5.01	5.00	(0.01)	4.33	(0.68)	4.98	(0.03)
	II	4.97	4.99	0.03	4.64	(0.33)	4.90	(0.06)
	III	4.95	4.96	0.01	4.50	(0.45)	4.88	(0.07)
	IV	4.92	4.94	0.01	4.43	(0.50)	4.87	(0.06)
			RMSE	0.02	RMSE	0.50	RMSE	0.06

On the other hand, the comparison of model accuracy for nowcasting investment showed that the forecast error of the Dynamic Factor Model was smaller than the forecasting error of the Bridge Equation and ARIMA, as presented in Table 4.2. The forecast error of the DFM model was considered significant but smaller than that of the benchmark models.

Table 4.2. Comparison with Other Models – Investment

Nowcasting		Realization	Nowcasting Result [and Error/ Deviation]					
			Dynamic Factor Model		Bridge Equation		ARIMA	
2015	I	4.63	5.06	0.48	3.10	(1.53)	5.38	0.75
	II	3.88	4.73	0.75	3.97	0.09	5.43	1.55
	III	4.79	4.63	(0.19)	4.00	(0.79)	4.78	(0.01)
	IV	6.90	5.06	(1.84)	4.31	(2.59)	5.68	(1.22)
			RMSE	1.03	RMSE	1.55	RMSE	1.06

5. CONCLUSIONS

The most robust nowcasting models for household consumption and investment were built using the Dynamic Factor Model (DFM) method based on the various testing and exercise stages. The indicators used for nowcasting household consumption were motor vehicle sales, total deposit, consumer lending rate, M1 and the Rupiah exchange rate (NEER), while the indicators used for nowcasting investment included cement sales, motor vehicle production, electricity consumption, total credit and M1.

Accuracy testing revealed that the forecast error of the household consumption nowcasting model using the DFM method was small and, therefore, robust in terms of predicting the level of household consumption. Meanwhile, the forecast error of the investment nowcasting model using the DFM method was large enough but smaller than the benchmark models (Bridge Equation and ARIMA).

REFERENCES

Andersson, Michael K. and Reijer, J.J. den. "Nowcasting." *Sveriges Riksbank Economic Review*, No. 2015:1, 2015.

Angelini, Elena; Camba-Méndez, Gonzalo; Giannone, Domenico; Rünstler, Gerhard; and Reichlin, Lucrezia. "Short-Term Forecasts of Euro Area GDP Growth." *ECB Working Paper Series*, No. 949, October 2008.

Bańbura, Marta and Rünstler, Gerhard. "A Look into the Factor Model Black Box: Publication Lags and the Role of Hard and Soft Data in Forecasting GDP." *International Journal of Forecasting*, 27, 2011(a).

Bańbura, Marta; Giannone, Domenico; and Reichlin, Lucrezia. "Nowcasting." *The Oxford Handbook of Economic Forecasting*, July 2011(b).

Bell, Venetia; Co, Lai Wah; Stone, Sophie; and Wallis, Gavin. "Nowcasting UK GDP Growth." *Bank of England Quarterly Bulletin Q1-2014*, 2014.

Camacho, Maximo and Perez-Quiros, Gabriel. "Introducing the Euro-Sting: Short-Term Indicator of Euro Area Growth." *Journal of Applied Econometrics*, 25(4), 2010.

Ghysels, Eric; Santa-Clara, Pedro; and Valkanov, Rossen. "The MIDAS Touch: Mixed Data Sampling Regression Models." June 2004.

Giannone, Domenico; Reichlin, Lucrezia; and Small, David. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data." *Journal of Monetary Economics*, 55, 2008.

Kurniawan, Ferry. "Nowcasting Indonesian Economy". Bank Indonesia, October 2015.

Luciani, Matteo; Pundit, Madhavi; Ramayandi, Arief; and Veronese, Giovanni. "Nowcasting Indonesia." *FRB Finance and Economics Discussion Series*, No. 2015-100, September 2015.

Mariano, Roberto S. and Murasawa, Yasutomo. "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series." *Journal of Applied Econometrics*, 18(4), 2003.

Mariano, Roberto S. and Murasawa, Yasutomo. "A Coincident Index, Common Factors, and Monthly Real GDP." *Oxford Bulletin of Economics and Statistics*, 72(1), 2010.

OECD System of Composite Leading Indicators. OECD, April 2012.

Schorfheide, Frank and Song, Dongho. "Real-Time Forecasting with a Mixed-Frequency VAR." *NBER Working Paper Series*, No. 19712, December 2013.

Modelling Regional Economic Growth in East Java Province 2009-2014 Using Spatial Panel Regression Model

Akmad Thoifur and Erni Tri Astuti*)
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia
*) erni@stis.ac.id

1. INTRODUCTION

Economic growth is an increase in the capacity of an economy to produce goods and services, compared from one period of time to another at a certain country. Traditionally, aggregate economic growth is measured in terms of gross national product (GNP) or gross domestic product (GDP), although alternative metrics are sometimes used. Economic growth can be measured either at national level or regional level, especially after enactment of regional autonomy. Regional economic growth is very important indicator for evaluating the performance of local government. Government, both central and local will seek to encourage the acceleration of economic growth, so that public welfare is expected to increase.

East Java Province is one of the biggest provinces in Indonesia consist of 29 regencies and 9 cities. During 2009 to 2014, although it had achieved the 2nd biggest regional GDP in 2014, which contributed almost 14.40 percent of the national GDP, the average economic growth of East Java province is still 5.95 percent per year, below the average national economic growth 6.18 per year. Moreover, if we look up at regency level, there are unbalanced growth among regencies. For example, Kota Batu and Surabaya can achieved economic growth up to 7.18 percent and 6.90 percent per year, while others can only achieved 5.31 percent per year (Sampang Regency) or 4.98 percent per year (Kota Kediri). Due to various characteristics, regions will grow and develop variously, some are fast growing and some are depressed. In facts, region that have high economic growth tend to cluster, as well as regions with low economic growth. The tendency of regional clustering indicates spatial dependency of economic growth in East Java Province.

Perroux in Pasaribu (2015) said that the process of economic development in a region are related to the geographical position of the region. Tobler's law of Geography said that anything are interrelated, but something close has bigger relation than far ones. Directly or indirectly, economic development in a region will affect and be affected the surrounding area. It is proven by the regional input output tables, migration and urbanization flow, and cross-regional flow of commodities.

There are some concept about impact of growth center area to surrounding area as follows (Pasaribu 2015):

Spread effect (trickle-down-effect or positive spillover) is the impact of growth center that pull up surrounding area because it will expand the distribution of resources in the surrounding area

Backwash effect (negative spillover) is the impact of the growth center that harm the surrounding area because it will absorb resources from the surrounding area.

Net spillover effects, the impact of growth center initially absorb other local resources, but in the long run will pull up surrounding area.

Theoretically, there are many factors that can affect economic growth in certain region, such as natural resources, human resources, capital stocks, investments, technological developments, and labor efficiency (Todaro and Smith, 2012). From statistical modelling point of view, economic growth plays as response variable and others are explanatory variables. Regression model is a tool that can used to reveal the relationship between response variable with some explanatory variables. However classical regression modeling based on ordinary least square estimation (OLS) required some restricted assumptions. One of these assumptions is non autocorrelation, meaning that each observations are independent with others. However, this assumption is often violated, especially when the observational unit concerning area or spatial data. Campbell (2012) said that the main principle of spatial data are autocorrelation and spatial heterogeneity. Autocorrelation or spatial interactions will lead to spatial linkages. This will violating non autocorrelation assumption, and OLS will not efficient anymore. Addressing the conditions, it is necessary to apply regression model that can accommodate spatial autocorrelation, namely spatial regression model.

The purpose of this study is to determine factors that influenced economic growth in East Java Province and also incorporating the spatial aspects.

2. MATERIAL AND METHODS

This study cover East Java Province which consists of 29 regencies and 9 cities during 2009 and 2014 period of time. The data is in form of panel data. The term panel data refers to multi-dimensional data frequently involving measurements over time. Panel data contain observations of multiple phenomena obtained over multiple time periods for the same region. Economic growth as a response variable is measured by real regional gross domestic product (RGDP) taken from annual report published by BPS Statistics Indonesia. Because of the limitation of the data, we only used three explanatory variables suspected to affect economic growth, including human resources involved in the process of production, investment, and technological developments. Human resources variable is approached by number of labor (L), investment is approached by capital expenditure spending of local governments (K), while technological developments is approached by human development index (HDI). In spatial analysis, position of any region to others is very important because it will included in the model in terms of spatial weight. Spatial weight w_{ij} typically reflects the "spatial influence" of region j on region i. There are many criteria to determine the value of spatial weight w_{ij}, and in this study we measured the weight by queen contiguity criteria. These weights simply indicate whether spatial units share a boundary or not. Queen Contiguity defines a neighbor when at least one point on the boundary of one region is shared with at least one point of its neighbor (common border or corner).

Before doing the analysis, the existence of spatial autocorrelation was checked by calculating Moran's Index (*I*). Positive value of *I* means that region with high value of economics growth tend to be surrounding by others region that have high economic growth, vice versa. Zero I means that thera is no spatial autocorrelation (Anselin, 1988). Sometimes Moran Scatter plot can help to determine this spatial relationship. Moran scatterplot divide the observations or regions into 4 distinct quadrant. Quadrant 1 is called hotspot, consists of fast growing regions surrounded by fast growing regions or high-high clustering. Quadrant 2 consist of slow growing regions surrounded by fast growing regions. Quadrant 3 is called cold spot or low-low clustering, consists of slow growing regions surrounded by slow growing regions. Quadrant 4 consist of fast growing regions surrounded by slow growing regions. Quadrant 2 and 4 are called spatial outlier. Furthermore, a statistical test for the existences of spatial autocorrelation is conducted.

The next step is doing model selection in panel data analysis, between spatial error model (SEM) and spatial lag or autoregressive model (SAR). There is also a selection process fixed effect and random effect model. Fixed effects is used when we want to control for omitted variables that differ between regions but are constant over time. Meanwhile, random effects is the model to use when we want to control for omitted variables that change over time but are constant between regions. Hausman test by Mutl dan Pfaffermayr (2008) is applied to choose which effect is appropriate.

After selection process, in this study we used spatial auto regression model with fixed effect defined as

$$\ln \mathrm{RGDP}_{it} = \lambda \sum_{j=1}^{38} w_{ij} \ln \mathrm{RGDP}_{jt} + \beta_1 \ln K_{it} + \beta_2 \ln L_{it} + \beta_3 \ln \mathrm{HDI}_{it} + \mu_i + \rho \sum_{j=1}^{38} w_{ij} u_{jt} + \varepsilon_{it}$$

Where:

λ : autoregressive coefficient
β_k : slope of k^{th} variable (*k*=1,2,3)
ρ : Spatial error coefficient
w_{ij} : Spatial weight between i^{th} and j^{th} regions

u_{it} : autocorrelation error of i^{th} region t^{th} year (i=1,2,…,38), (t=2009,2010,…,2014)
μ_i : Spatial specific effect of i^{th} region
ε_{it} : error regions i year t

3. RESULT

Figure 1 reflect the economic growth of East Java Province during 2009 to 2014. After four years of outstanding increment and reached the highest value up to 6.95 percent in 2012, the economic growth become decline caused by global financial crisis. Region with the highest economic growth is Kota Batu which can reached 7.18 percent in average per yaar. In the other hands, Kota Kediri has a lowest scored by only 4.98 percent in average. There are 18 regions that have economic growth rate higher than the average growth of East Java Province, while 20 regions were below it.

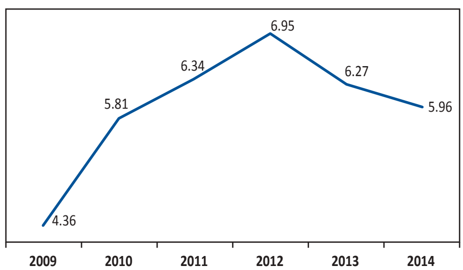


Figure 1. Economic growth of East Java 2009-2014

For examining the spatial relationship between these regions in terms of economic growth, we used Moran scatter plot as shown in Figure 2. There are 11 regions in quadrant I, 6 regions in quadrant II, 15 regions in quadrant III, and 6 regions in quadrant IV. Regarding to the location, regions in central and northern coast of East Java Province clustered in the hotspot areas (quadrant 1). In these regencies or cities, there are many economic activity such as center of industrial, trade and education activity. Not only strategic, but these regions also supported by good infrastructure. Meanwhile, almost all regions at the eastern coast and Madura island clustered in cold spot area (quadrant 3). They have low growth and be surrounded by low growth area as well. These condition may be caused less strategic region, lack of infrastructure, and economic activity that relies on primary sector. On the other hands, regions in south-west of East Java are more heterogenous. Most of them have low economic growth because of unfavorable geographic condition, lack of infrastructure, and economic activity that relies on primary sector.

Further investigation by Moran test, where the results is shown in Table 1, in total and in every year as well, Moran Index are significant at α = 5 percent. These conditions reflected the significance spatial autocorrelation among regencies in term of economic growth in East Java Province. Economic development in a region is influenced not only by internal factors in the regions concerned but also influenced by economic growth in surrounding areas. Positive Moran I means positive autocorrelation. Areas that achieved high economic growth tend to be surrounded by areas with high rates of economic growth as well, vice versa.

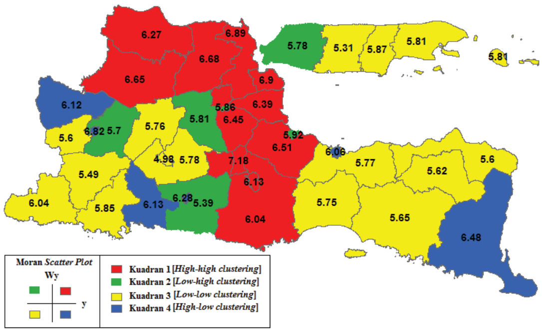


Figure 2. Distributions of Regions in East Java based Quadrat in Moran Scatterplot

Table 1. Result of Moran test

Year	I	E(I)	Var(I)	Z	*p-value*
-1	-2	-3	-4	-5	-6
2009	0.199923	-0.027027	0.016288	1.7783	0.0377
2010	0.200573	-0.027027	0.016288	1.7834	0.0373
2011	0.202782	-0.027027	0.016288	1.8007	0.0359
2012	0.204444	-0.027027	0.016288	1.8137	0.0349
2013	0.207709	-0.027027	0.016288	1.8393	0.0329
2014	0.209403	-0.027027	0.016288	1.8526	0.032
Total	0.213293	-0.004405	0.002954	4.0055	0

The result of Lagrange Multiplier test in Table 2 showed that spatial lag model (spatial autoregressive) is more suitable for modelling regional economic growth in East Java than spatial error model. Furthermore, from the Hausman fixed effect is more appropriate for the data rather than random effect with p-value less than 5 percent.

Table 2. LM Test Result

LM Test	Test Statistic	*df*	*p-value*
-1	-2	-3	-4
LM Error	3.324	1	0.06828
LM Lag	16.533	1	0.00004

Parameters estimates from spatial autoregressive regression model are then determined by applying maximum likelihhod estimation. The estimated parameter are shown in Table 3.

Table 3. Estimation of Parameter Model

Variables	Koeficient	*Std Error*	t-stat	*p-value*
-1	-2	-3	-4	-5
Intercep	-0.9261	0.15998	-5.7888	0.000**
Spatial Lag	0.92089	0.01418	64.9225	0.000**
Ln(K)	0.00377	0.00177	2.1308	0.016**
Ln(L)	0.01784	0.01358	1.3132	0.094*
Ln(HDI)	0.32828	0.06908	4.7523	0.000**

*) significant at α = 10 percent
**) significant at α = 5 percent

The final model can be written as

$$\ln \widehat{RGDP}_{it} = (-0{,}92610 + \mu_i) + 0{,}92089 \sum_{j=1}^{38} w_{ij} \ln RGDP_{jt}$$
$$+ 0{,}00377 \ln K_{it} + 0{,}01784 \ln L_{it} + 0{,}32828 \ln HDI_{it}$$

Spatial autoregressive coefficient indicate significant relationship between the regions in economic growth. The fast growing areas tend to be surrounded by fast growing areas as well, and vice versa. The positive influence of neighboring area to the region in accordance with the theory of trickle-down effect. Growth center-area will provide positive spillover (spread effect) for the neighboring area (buffer). In other words, growth center-area pull up buffer zone to grow.

Capital expenditure spending and HDI have positive and significant impact to economic growth (at α = 5%). While, number of labor has positive and significant impact to economic growth at α = 10%. These are suitable with the theory announced by Adam Smith, Todaro, Solow-Swan, and others.

In spatial panel regression model with fixed effect approach, there is spatial specific effect μ_i attached to each unit of spatial observation. It shows the characteristics of the region effect on economic growth yet not included in the model, such as natural resources, economic structure, geographic condition, private investment, infrastructure, etc. The spatial specific effect μ_i is displayed below.

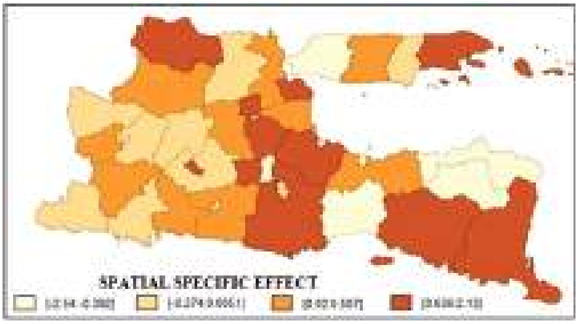


Figure 3. Spatial Specific Effect any Districts in East Java

4. CONCLUSION

During 2009 to 2014, there are significant evidence of spatial correlation between regencies or cities in East Java Province in term of economic growth. Positive autocorrelation prove that homogenous regions tend to clustered. The fast growing region tends to be surrounded by fast growing region as well. Conversely, areas of low economic growth tend to be surrounded by areas that have low economic growth anyway. Linkages among regions are getting stronger and stronger because of infrastructure and transportations progress.

Investment (capital expenditure spending), number of labor, and technological development (approached by HDI) have positive significant impact to economic growth. Fast growing regions give positive *spillover* to neighboring regions. In other words, center-of-growth area pull up buffer area to grow. Closer areas give greater influence than far ones, as said by Tobler's theory.

REFERENCES

Anselin, Luc (1988) *Spatial Econometrics: Methods and Models*. Springer: Science+Business Media Dordrecht

Campbell, Jonathan dan Shin, Michael. (2012) *Essentials to Geographic Information Systems.* New York: Flat World Knowledge, Inc.

Elhorst, J. Paul (2014) *Spatial Econometrics from Cross-Sectional Data to Spatial Panels.* Springer

Lloyd (2007) *Local Models for Spatial Analysis*. CRC Press

Mutl, Jan dan Pfaffermayr, Michael (2008) "The Spatial Random Effects and the Spatial Fixed Effects Model: The Hausman Test in a Cliff and Ord Panel Model" *Economics Series 229.* Vienna: Institute for Advanced Studies

Pasaribu, Ernawati (2015) *Dampak Spillover Dan Multipolaritas Pengembangan Wilayah Pusat-Pusat Pertumbuhan Di Kalimantan* [Desertation]. Bogor:IPB

Sukirno, Sadono. (2011) *Makroekonomi, Teori Pengantar Edisi Ketiga*. Jakarta : Rajawali Press

Estimation of Environmental Kuznets Curve for CO2 Emissions and Methane Emissions: Empirical Analysis for Indonesia

Debita Tejo Saputri[1] and Budiasih[2]
[1] Statistical College, Jakarta, Indonesia – 13.7552@stis.ac.id
[2] Statistical College, Jakarta, Indonesia – budiasih@stis.ac.id

Abstract
The Environment Kuznets Curve (EKC) is named after Kuznets (1955) who hypothesized a relationship between various indicators of environmental degradation and income per capita. This study focuses on the relationship between per capita income, per capita energy, coal, oil, and openness on environment degradation (per capita CO_2 emissions and methane emissiosn) in Indonesia by the period 1971 to 2015. Findings from Error Correction Model (ECM) method is as follows: theory of EKC for Indonesia showed statistically significant for per capita of CO_2 emissions while for methane emissions does not. The sign the output indicates the level of per capita CO_2 emissions and methane emissions initially goes up but eventually goes down with the rise in GDP per capita over time. The coefficient of error-correction terms of all models has a correct sign (negative) and is statistically significant at 5 percent, which indicates the adjustment takes place towards the long run equilibrium. It suggests the validity of long-run equilibrium relationship among the variables. Thus, it is recommended to look for ways to promote green growth in the country.

Keywords : CO_2 Emissions; Methane Emissions; EKC; ECM
JEL : C22, Q13, O56

1. INTRODUCTION

The concept of the SDGs was established at the UN Conference on Sustainable Development, Rio+20, Brazil in 2012. The objective was to produce a set of universally applicable goals that balances the three dimensions of sustainable development: environmental, social, and economic. Indonesia is a significant global player both in terms of economic performance and environmental emissions. Indonesia continues to be a rising power both in the Association of Southeast Asian Nations and the G20 with a Gross Domestic Product of $861,9 billion in 2015 (World Bank, 2015). However, the country's environment becomes worse under such continuous expansion. The combination of high population density and high levels of biodiversity, together with 17,500 islands, makes Indonesia one of the most vulnerable countries to the impacts of climate change (Yusuf AA and HA Francisco. 2009). Like other developing countries, environment emissions has become a major problem because this country was the sixth position in the world's top ten emitters with 2,160.64 $MtCO_2e$ total Greenhouse Gass (GHG) (World Resources Institute, 2013).

Emissions in Indonesia has gotten the world's attention when the Government of Indonesia hosted the UN Climate Change Conference in Bali 2007 resulting the Bali Road Map, a variety of decisions and programs that will reach toward a safer climate future. Indonesia has ambitious goals to reduce GHG emissions. On the September 2009, G-20 meeting in Pittsburgh, former President Susilo Bambang Yudhoyono laid out a vision where Indonesian government was crafting a policy that would cut emissions by 26 percent by 2020 from business as usual levels. Five years after that, in 2015, president "Jokowi" when delivering a statement on the UN Framework Convention on Climate Change, Paris, has revised the target become 29 percent reduction in greenhouse gas emissions by 2030.

A study from Lashof and Dillip (1990) and Climate Analysis Indicator Tool (2010) found that the main greenhouse gasses (GHGs) in the Earth's atmosphere contains carbon dioxide (70 percent), methane (20 percent), nitrous oxide (8 percent) and other (3 percent). This study exemines the relationship among GHGs emissions (per capita CO_2 emissions and methane emissions), per capita income, per capita energy, coal, oil, and openness based on the environmental Kuznets Curve (EKC) Hypothesis for Indonesia during period 1971-2015. The issue of environmental pollutants is a progressive trend in developing countries as they require more energy consumption for higher economic development. Consequently, they suffer from the environmental problems. Foreign trade affects the environmental quality through scale, composition and technical effect (Chen Honglei, et al., 2011). The presence of the EKC hypothesis will provide adequate information to policy makers on whether and in which intervention is required to reduce CO_2 emissions and methane emissions, while at the same time maintain economic growth. The remaining of the paper is organized as follows: Section 2 discusses the data and methodologies employed, while Section 3 presents the empirical result and analysis. Finally, Section 4 concludes, offering certain policy recommendations.

2. DATA AND METHODOLOGY

2.1. Data

The data employs an annual time series data for the period of 1971–2015 collected from World Development Indicator. According to the mechanism of the relationship among per capita CO_2 emissions and methane emissions, per capita income, per capita energy, coal, oil, and openness, the regresion models are as follows:

Dependent variable CO_2 emissions

$$Ln\_CO_{2t} = \beta_0 + \beta_1\, Ln\_GDPC_t + \beta_2\, Ln\_GDPC_t^2 + \beta_3\, Ln\_energy_t + \beta_4\, coal_t + \beta_5\, oil_t + \beta_6\, openness_t + \boldsymbol{\varepsilon_t} \tag{1}$$

Dependent variable methane emissions

$$Ln\_methane_t = \beta_0 + \beta_1\, Ln\_GDPC_t + \beta_2\, Ln\_GDPC_t^2 + \beta_3\, Ln\_energy_t + \beta_4\, coal_t + \beta_5\, oil_t + \beta_6\, Openness_t + \boldsymbol{\varepsilon_t} \tag{2}$$

Equation shows variables from the models : CO_2 (metric tons per capita) and methane (thousand metric tons of CO_2 equivalent) as environment ambient; GDPC as real GDP per capita; energy (kg of oil equivalent per capita); rent coal (% of GDP); rent oil (% of GDP), and openness is trade as a share of GDP, the sum exports and imports divided by the value of GDP in current U.S. dollars. For the empirical exercise are presented in the next section, all data are presented in a logarithmic form except coal, oil, and openness. For the validity, the EKC theory requires in the quadratic model to Ln_GDPCt be positive and significant while Ln_GDPCt^2 has to be negative and significant. The EKC hypothesis argues that the initial phase of economic growth has negative consequences to the environment, but the negative consequences diminished as the growth rate surpasses a certain point.

2.2. Methodology

This study employs an error correction model (ECM) because it goes beyond cointegration regression in revealing how short-run dynamics work in order to revert to the long-run relationship between the quality of the environment, income, coal, oil, energy, and openness. As in standard ECM estimation, the starting point is to establish the non-stationarity of the variables used in the study and then proceed with an estimation. Dickey-(DF) test is used to examine stationary data.

$$\mathrm{Y_t} = \alpha_0 + \alpha_1 \mathrm{Y_{t-1}} + \varepsilon_\mathrm{t} \tag{3}$$

Second, estimated long-run equilibrium relationship, by estimating the regression all of the cointegrated variables (Thomas, 1997). *Augmented Engle–Granger (AEG) Test* is used to find out a long run among the variables. Further, if there is an existence of the long-run relationship between then Error Correction test is applied to find out a short-run relationship because of the possibility of disequilibrium in the short-run. The cointegration technique pioneered by Granger and Engle (Gujarati, 2004) permits long run components of variables to conform long-run equilibrium relationships of the short-run components having a flexible dynamic specification. Cointegration for two (or more) time series variable indicates that there is a long-run relationship or equilibrium between these variables. After a time series relationship between long-run and mutually cointegrated, short-run disequilibrium in the relationship between variables can be estimated. Under the ECM specification, the error term to its long-run value and short-run equations are given by equations (4) and (6), respectively:

$$\mathrm{Y_t} = \beta_0 + \beta_1 X_\mathrm{t} + \mathrm{u_t} \tag{4}$$

$$u_t = \mathrm{Y_t} - \beta o - \beta_1 X_t \tag{5}$$

$$\Delta Y_t = \alpha_0 + \alpha_1 dX_\mathrm{t} + \alpha_2\, \mathrm{u_{t-1}} + \varepsilon_\mathrm{t} \tag{6}$$

$$\mathrm{u_{t-1}} = \mathrm{Y_{t-1}} - \beta o - \beta_1 X_{t-1} \tag{7}$$

Yt (the dependent variable) is an indicator of the quality of the environment, u_t is a white error term, while ECT or u_{t-1} is lagged value of the error term. The ECM will have a valid model specification if it has a negative and statistically significant Error Correction Term (ECT) coefficient. The probability value of ECT coefficient that is less than 5 percent significance level will indicate its significance. This study also conducts some diagnostic test for checking the reliability of the model such as the assumption of normality, homoskedasticity, non-autocorrelation, and non-multicollinearity.

3. EMPIRICAL RESULT AND ANALYSIS

CHG Emissions In Indonesia

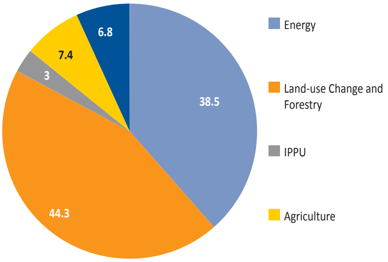


Figure 1. GHG Emissions by Sector Indonesia 2013 (percent)

The figure above presents a number of all greenhouse gas emissions emitted in Indonesia from all sectors 2013. The figure illustrates that the main contributor to Indonesia's high release of greenhouse gasses is forestry, which accounts for 44.3 percent in 2013. Emissions from the energy sector in Indonesia growing at a rapid pace. The sources of greenhouse gas emissions in the energy sector are manufacturer (151,871); transportation (136,774); commertial (3,270); retidential (30,379); and non specifed (10,174). The emissions from the energy sector consis accounted for 9 percent of Indonesia's total emissions in 2007 and become 38.5 percent of total emissions in 2013 (Ministry of Environmen). As a result of economic growth and industrialization, the emissions from industry, transport, and power generation are also increasing.

Table 1. Test of the Unit Root Hypothesis

Variable	Level		First Difference		
	Sign Trend	DF Test	Trend	DF Test	
-1	-2	-3	-4	-5	-6
Ln_ CO2	0.0053 **	0.0878 *	0.2594	0.0000 **	Strong stationary in 1st difference
Ln_methane	0.0402 **	0.003 **	0.9176	0.0000 **	Strong stationary in 1st difference
Ln_GDPC	0.0361 **	0.4165	0.4975	0.0016 **	Strong stationary in 1st difference
Ln_GDPC2	0.0248 **	0.3698	0.7544	0.0017 **	Strong stationary in 1st difference
Ln_energy	0.2123	0.8432	0.4413	0.0000 **	Strong stationary in 1st difference
Coal	0.0443 **	0.5336	0.5155	0.0000 **	Strong stationary in 1st difference
Oil	0.0076	0.0587	0.4327	0.0000 **	Strong stationary in 1st difference
Openness	0.8401	0.1812	0.1146	0.0001 **	Strong stationary in 1st difference

** indicates significant at 5 percent level
* indicates significant at 10 percent level

Regression analysis based on time series data assumes that the underlying time series are stationary. It is a well-known empirical fact that many macroeconomic time series are typically nonstationary, as indicated by the serial correlation between successive observations, particularly when the sampling interval is very small. Spurious regression implies that the classical t and F-tests are not appropriate and will result in misleading conclusions (Nachrowi and Hardius Usman, 2006). Analysis of cointegration starts with the determination of the univariate properties of the time series. If the series are integrated of the same order, it is reasonable to proceed with the cointegration test. In conjunction with testing for the degree of integration, this study applies the Dickey-Fuller (DF) unit root test. Result confirm the existence unit root, and therefore non-stationary at the level, and all the variables are strong stationary at first difference. The lag length has been chosen based on the SIC for DF test.

Table 2. Estimated Long Run Coefficients

Variable	*CO_2*		*Methane*	
(1)	*(2)*	*(3)*	*(4)*	*(5)*
Constant	-18.9639	R^2	-11.92814	R^2
	(0.0002) **	= 0.982856	(0.5916)	= 0.351093
Ln_GDPC_t	3.325007	R^2adj	6.001169	R^2adj
	(0.0096) **	= 0.979917	(0.3102)	= 0.233109
$Ln\_GDPC^2_t$	-0.191014	DW	-0.387583	DW
	(0.0341) **	= 1.3414	(0.3624)	= 2.01778
Ln_energy_t	0.765649	F-Stat	0.214911	F-Stat
	(0.0031) **	= 334.4150 **	(0.08422) *	= 2.97578 **
$Coal_t$	0.102107	Prob F	0.033873	Prob F
	(0.0704) *	= 0.0000 **	(0.8974)	= 0.019578 **
Oil_t	0.0037		0.000562	
	(0.0037) **		(0.9639)	
$Openness_t$	-0.003232		-0.005258	
	(0.2213)		(0.6483)	
Pattern				

** indicates significant at 5 percent level
* indicates significant at 10 percent level

Equations are as follows:

$$\widehat{Ln\_CO}_{2t} = -18.9639^{**} + 3.325007^{**} Ln\_GDPC_t - 0.1910^{**}\, Ln\_GDPC\ {}_t^2 + 0.765649^{**}\ Ln_{energy_t} + 0.1021^{*} coal_t + 0.0037^{**}\ oil_t - 0.003232\ openness_t$$

$$\widehat{Ln\_methane}_t = -11.92814 + 6.001169_1\, Ln_{GDPC_t} - 0.3875\, Ln_{GDPC_t}^2 + 0.2149^{*}\, Ln_{energy_t} + 0.03387\ coal_t + 0.000562\ oil_t - 0.005258\ Openness_t$$

Indonesia adopted a strategic long-term development plan (2005-2025) targeting to achieve a green and everlasting Indonesia through implementing various environmental policies. Moreover, the growth paradigm of Indonesia still largely follows primary sector, for instance, agriculture, forestry, and mining contribute about 25% to Indonesia's GDP. When an economy starts moving along the growth trajectory, then at the earliest stage of economic growth, environment deteriorates rapidly due to ambient air pollution, deforestation, soil and water contamination, and several other factors (Alama, et al, 2016). With a rise in the level of income, when the economy starts to develop at a particular level of income, environmental degradation starts to come down and environmental quality improves. This phenomenon is referred to as Environmental Kuznets Curve (EKC) hypothesis in the literature of environmental economics, named after Simon Kuznets (1955), who described the inverted U-curve association between indicators of environmental degradation and income.

The basic EKC relationship is best understood as a purely long run concept (Alama et al, 2016). Estimation of the short-term dynamics may yield some interesting insights into how a country's emissions evolve over time, but the shape of the EKC must be found in the long run equation. Both pollutants follow an inverted U-shaped curve with the negative coefficient of Ln_GDPCt^2. Therefore, EKC is proven to be true in CO_2 emissions, while in methane emissions model is not statistically significant. It means environmental pollutions will increase along with the GDP growth before the turning point. This finding is consistent with Rahman et al. (2016) who found long run nexus between output growth and CO_2 emissions in Indonesia using Toda-Yamamoto approach during 1971-2011. And Alama, et al, 2016 who using ARDL method found EKC for CO_2 emissions Indonesia by the time 1971 to 2014 data. But this finding contradicts with Jafari et al, 2012 who did not find any long-run relationship between economic growth and CO_2 emissions in Indonesia using today-Yamamoto approach during 1971-2007. Shahbaz et al., 2013 who, using VECM and ARDL approach during 1975 Q1 to 2011Q4, found that both energy consumption and economic growth increase CO_2 emissions in Indonesia in both the short term and long term.

Both indicators of emissions are related negatively to the openning of the economy, implying that openness is good for the environment. This study has same found with Chen Honglei, et al., Cole, 2003; and Copeland and Taylor, 2004. Two separate views exist to present the interaction between foreign trade and environmental degradation. One of them is the Pollution Haven Hypothesis and the other one is technology transfers view. This theory posits that the existence of an EKC in developing countries is a consequence of free trade, which displaces "dirty" industries from developed to developing countries that have relatively lax environmental standards, regulations, and enforcement mechanisms (Chen Honglei, et al., 2011; Cole, 2004). On the other hand, the increase of trade openness can accelerate the capital mobility for new technologies via technology transfer and ease the facility of environment-friendly technologies. This situation can decrease the environmental deterioration in long-term as countries become more competitive and thus more efficient in use of resources (Akin, 2014).

As to the relation between energy, coal, oil and emissions, the coefficient of CO_2 model is significantly positive. Indonesia greenhouse gas emissions from coal burning by the year 2025 will be 20 times higher than in 2005 or 1.3 times higher than all energy sector emissions for the same year. The decision to rapidly expand coal - fired power generation (by 10,000 MW in Java alone) may increase these emissions even further (World Bank, 2007). The values of adjusted R^2 for CO_2 models were estimated to be above 90 percent, which indicates that the model is strongly good fitted.

While methane only accounts for 241.13 (11 percent) of greenhouse gas emissions in the Indonesia, this chemical packs a potent dose of warming 84 times more effective than CO_2 at absorbing heat. The climate system responds very quickly to changes in methane emissions. The immediate influence of methane lasts for only the 12 years it remains in the atmosphere, while carbon dioxide affects the climate for hundreds of years after emission. This study took a period for over 44 years, may it imply the model only have R^2 0.35 which means that the ability of the model to explain the dependent variable (methane) is 35 percent, the rest described other variables outside the model

Table 3. Cointegration Test

Dependent Variable	Sign Trend	Level		
		t-statistic	t-table	DF Test
-1	-2	-3	-4	-5
CO_2	0.9335	-4.64214 **	-3.51552	0.0029 **
Methane	0.978	-3.81784 **	-3.51552	0.0248 **

** indicates significant at 5 percent level

The table shows all residual regression are stationary in level in order to examine the cointegration relationship among the variables. The presence of cointegration shows that there exists a long-run relationship between the variables.

Table 4. Estimated Short Run Coefficients (Error Correction Model Result)

Variable	CO_2		Methane	
(1)	(2)	(3)	(4)	(5)
Constant	-0.034179	R^2	-0.290721	R^2
	(0.2377)	= 0.545905	(0.0477) **	= 0.58658
$d(Ln\_GDPC_t)$	9.912217	R^2adj	-7.812823	R^2adj
	(0.0064) **	= 0.43995	(0.6273)	= 0.483227
$d(Ln\_GDPC^2_t)$	-0.675442	DW	0.195154	DW
	(0.0064) **	= 1.186562	(0.8610)	= 2.13454
$d(Ln\_energy_t)$	0.594472	F-Stat	-0.784885	F-Stat
	(0.0417) **	= 5.1522 **	(0.5708)	= 5.675 **
$d(Coal_t)$	0.049209	Prob F	-0.2672	Prob F
	(0.3125)	= 0.00062 **	(0.2405)	= 0.000373 **
$d(Oil_t)$	0.003871		-0.018303	
	(0.1758)		(0.2279)	
$d(Openness_t)$	-0.003630		0.018303	
	(0.1407)		(0.1410)	
ECT	-0.663895		-0.93765	
	(0.0001) **		(0.0000) **	
Normality		(0.401785)		(0.458178)
Homoskedastic		(0.4863)		(0.1214)
Non-Autocol		(0.0668)		(0.0635)

** indicates significant at 5 percent level
* indicates significant at 10 percent level

Equations are as follows:

$$\begin{aligned} d(\widehat{Ln\_CO}_{2t}) \quad &= -0.034179 + 9.912217^{**} d(Ln\_GDPC_t) - 0.675442^{**}\, d(Ln\_GDPC_t^2) \\ &\quad + 0.594472^{**}\, d(\,Ln\_energy_t) \; + 0.049209\, d(coal_t) + 0.003871 \;\, d(oil_t) \\ &\quad - 0.003630\, d(openness_t) + \;(-0.663895)^{**}\hat{\varepsilon}_{t-1} \end{aligned}$$

$$\begin{aligned} d(Ln\_\widehat{methane}_t) \;\; &= -0.290721^{**} + 7.8128\, d(Ln\_GDPC_t) - 0.19515\, d(Ln\_GDPC_t^2) \\ &\quad + 0.7848\, d(\,Ln\_energy_t) - 0.2672\, d(coal_t) - 0.018303\, d(oil_t) \\ &\quad - 0.0183\, d(openness_t) + \;(-0.93765)^{**}\hat{\varepsilon}_{t-1} \end{aligned}$$

Unlike long run scenario, the model shows only per capita real GDP, per capita real GDP^2 and per capita energy has a positive and statistically significant influence on CO_2 emissions in short-run. The quadratic GDP has the statistically significant influence or inverted U-shape. While there are found an inverted U-shape between methane and per capita real GDP in CO_2 emissions, but it is most important to mention, the lag error correction term (ECT(-1)) shows negative significant for all of the models, which indicates both the long-run equilibrium and short-run disequilibrium. The coefficients of ECM were found to be -0.66 for CO_2 emissions model, which indicates that adjustment takes place 66 percent per year towards the long run equilibrium.

The coefficient of error-correction terms of all models has the correct sign (negative) and statistically significant at 5 percent. It suggests the validity of long-run equilibrium relationship among the variables. Meaning not only that the ECM is valid, but also that there is significant conservative force's tendency to bring the model back into equilibrium whenever it strays too far. The sizes of the coefficient of error-correction terms indicate that speed of adjustment is rather fast for the equation to return to their equilibrium level once it has been shocked at 5 percent significant level. Diagnostic tests confirm that there are found no serial correlation, no abnormality problem, and also no heteroscedasticity problem in the quadratic model for all the models as shown on the table above.

4. CONCLUSION AND POLICY IMPLICATION

Observed from empirical findings, the theory of EKC for Indonesia is statistically significant for per capita of CO_2 emissions while for methane emissions does not. While the sign of the output for both models indicates the level of per capita CO_2 emissions initially goes up but eventually goes down with the rise in GDP per capita over time. The coefficient of error-correction terms of all models have the correct sign (negative) and statistically significant at 5 percent. It suggests the validity of long-run equilibrium relationship among the variables.

There are several points that Indonesian policymakers may need to take into consideration in order to draft effective environmental policies to combat global warming while stimulating economic growth at the same time. From the policy perspective, Indonesia should adopt a green policy to achieve green growth. Most important in this respect is the use of energy. Use of fossil fuels like oil and natural gas is the main contributor to air pollution problems, need to adopt more environmentally friendly energy generation technologies (renewable) that will not only protect the environment but also will sustain increased electricity needs.

In other cases, indirect instruments policy may be more powerful and have fewer side-effects especially for industrial sector. One important example of this is "environmental taxes". By taxing activities generating negative external effects such as pollution. These green taxes may increase government revenues. Besides, implementation of Green GDP is need to gives value to the cost of environmental losses and therefore adjusts GDP to reflect the environmental costs. To get the required information for Green GDP accounting, monetary data together with physical data, as complementary data, is needed in order to reach the target of reducing emissions in 2030 by sectoral monitoring and evaluation. Furthermore, need to greater cooperation between government and rating agencies in certifying and rating the green projects.

With respect to directions for future research, it will be interesting to conduct further studies at both developed and developing countries for comparison purposes. As part of future research direction provided that data will be available, it would be interesting to analyze regions of developing and highly polluted area.

REFERENCES

Akin, Cemil Serhat. (2014). The Impact of Foreign Trade, Energy Consumption and Income on Co_2 Emissions). Department of Economics. Journal of Energy Economics and Policy. 4 (3) 465-475.

Alama, Mahmud., Wahid Murad, Abu Hanifa Md. Nomanc, Ilhan Ozturk. (2016). Relationships among Carbon Emissions, Economic Growth, Energy Consumption and Population Growth. Journal of Ecological Indicators 70 (2016) 466-479

Baltagi B. (2005). Econometric Analysis of Panel Data. John Wiley & Sons, Ltd, Inggris.

Chen Hongleia, Zou Xiaoronga, Chen Qiufengb. (2011). Export-oriented Economy & Environmental Pollution in China: the Empirical Study by Simultaneous Equation Model. Energy Procedia 5 (2011) 884–889

Cole, M.A. (2000). Air Pollution and 'Dirty Industries': How and Why does the Composition of Manufacturing Output Change with Economic Development?", Environmental and Resource Economics , 17(1) 102-123.

Copeland, B.R. and M .S. Taylor. (2004). Trade, Growth, and the Environment. Journal of Economic Literature , 42(1) 7- 71.

Daniel A. Lashof and Dilip R. Ahuja. (1990). Relative contributions of greenhouse gas emissions to global warming. Nature 344, 529 – 531.

Enders, W. (1995). Applied Econometric Time Series. New York: John Wiley and Sons, Inc

Gujarati DN, Porter DC. (2004). Basic Econometric Fifth Edition. The MC Graw-Hill Companies, NY

Jafari, Y., Othman, J., Nor, A.H.S.M., (2012). Energy Consumption, Economic Growth and Environmental Pollutants in Indonesia. Journal Policy Model. 34 (6), 879-889.

Khalil, Samina and Zeeshan Inam. (2006). Is Trade Good for Environment? A Unit Root Cointegration Analysis. The Pakistan Development Review. 45: 4 Part II (2006) 1187-1196.

Kuznets, S. (1995). Economic Growth and Income Inequality. American Economic Review 45 (1)11-28

Ministry of Environment. (2015). Status of Indonesian living Environment. KLH. Jakarta.

Nachrowi and Hardius Usman. (2006). Pendekatan Populer dan Praktis Ekonometrika untuk Analisis Ekonomi dan Keuangan. FEUI Jakarta.

Rahman, M.S., Noman, A.H.M., Shahari, F., Aslam, M., Gee, C.S., Isa, C.R., Pervin, S. (2016) . Efficient energy consumption in industrial sectors and its effect onenvironment: a comparative analysis between G8 and Southeast Asianemerging economies. Journal Energy 97 82–89

Shahbaz, M., Hye, Q.M.A., Tiwari, A.K., Leitao, N.C., (2013). Economic growth, energy consumption, financial development: international trade and CO_2 emissions in Indonesia. Renew Sustain Energy Rev. 25, 109-121.

Thomas, R. L. (1997). Modern Econometrics: An Introduction. England: Addison. Wesley Longman

U.S. Energy Information Administration (EIA). (2014). International Energy Statistics Washington, DC: U.S. Department of Energy. http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=90&pid=44&aid=8

World Bank. (2007). Executive Summary: Indonesia and Climate Change Working Paper on Current Status and Policies. http://www.wri.org/resources/data-sets/cait-country-greenhouse-gas-emissions-data

World Bank. (2016) .World Development Indicators. www.worldbank.org

World resources institut. http://cait.wri.org/profile/Indonesia

Yusuf AA and HA Francisco. (2009). Climate Change Vulnerability Mapping for Southeast Asia, IDRC-SIDA-EEPSEA- CIDA.

Growth Diagnostic: Evidence of Bali Province

CENTRAL BANK OF INDONESIA
Putriana Nurman[1], Ganis Arimurti[1], Umran Usman[1,]
Donni Fajar Anugrah[2], Robbi Nurrakhman[2], Evy Marya Deswita S[2]
December 2016

Abstract

Bali economy that is mainly driven by tourism services tends to experience stagnant growth in line with the tendency of decrease in tourism competitiveness and increasing competition among similar tourist destinations. In addition, economic growth in Bali is still yet to be enjoyed by all layers of society, as can be seen by gini ratio of Bali Province that is still in the range of 0.4 (2011-2015). In order to achieve strong and inclusive Bali economic growth, there needs to be studies of factors that hamper economic growth. Therefore, this research is conducted in order to identify key constraints to economic growth in Bali using the Growth Diagnostic concept developed by Hausmann, Rodrik and Velasco (2005), namely HRV Tree using secondary data and supported by Focus Group Discussion conducted with policy stakeholders in Bali. The study is also reinforced with a model of Computable General Equilibrium (CGE) -INDOTERM in order to observe impact of economic policy simulations.

Result of the study indicates that the main obstacle in Bali is the issue of infrastructure availability (supporting intra- and inter-connectivity and the availability of electricity supply) of Bali, the Human Resources (HR) quality, as well as the needs of the tourism product innovation. The simulation shows that highway construction, railway construction, quality of human resources improvement, increase in electricity capacity, tourism development, construction of North Bali Airport provide total impact of 2.23% in increasing economic growth and 2.31% of the increase in employment. The study recommends that the structural reforms in Bali need to be prioritized in the development strategy of infrastructure and in the development strategy of human resources quality.

JEL Classification:

Key Words: Macroeconomics, Regional Economy, *HRV Tree*, *CGE-INDOTERM*

1 Authors are analysts in Bank Indonesia Representative Office of Bali Province
2 Authors are researcher in Bank Indonesia

CPS32:
Statistical Computing & Technology

Simulation Study Multistage Clustering for Classify Stationary, Trend, and Seasonal Time Series Data Based on Autocorrelation Distance with Hierarchical Algorithm

Mohammad Alfan Alfian Riyadi*
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – m.alfan.alfian.riyadi@gmail.com
Aldho Riski Irawan
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – ir.aldhoriski@gmail.com
Dian Sukma Pratiwi
Institut Teknologi Bandung, Bandung, Indonesia – diansukmapratiwi@gmail.com
Kartika Fithriasari
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – kartika_f@statistics.its.ac.id

Abstract

Time series is a series of observational data arranged by a certain time sequence. Identification of time series model is used to determine the pattern of the data series such as, stationary model, trend, seasonal, and seasonal trend. Some forecasting methods such as decomposition, naive, and exponential smoothing requires identification on data patterns to determine the right model. But if the time series data have a large dimensions, the identification process is not easy to do one by one because it takes a long time. So it will needs certain methods to facilitate the grouping pattern of time series data. In this research, aim to cluster time series data which have the same data pattern. The data used is the simulated data to generate four types of data pattern that is stationary, trend, seasonal, and seasonal trend. The distance used is the autocorrelation distance with the complete linkage hierarchical algorithm. Cluster grouping the first stage of simulation data into three groups consisting of group 1 consists of data seasonal, group 2 consisted of data stationary, and group 3 consist of a seasonal trend data and data trend. Multistage clustering performed on the data trend and seasonal trends in group 3 which has been differenced lag 1 to form a new group that is data trends and separate seasonal trend. Based on multistage clustering has already distinguish between stationary data patterns, trend, seasonal, and seasonal trend. However, further research can be developed for data patterns based on ARIMA, or Seasonal ARIMA process and also tested on a real case.

Keywords: Autocorrelation distance, Hierarchical cluster, Multistage clustering, Time series model

1. INTRODUCTION

Time series commonly modeled by 2 methods, based on stationary and data plot. Visualization time series data can be done by plot and autocorrelation function (ACF), see Montgomery et.al (2008). It will be easy to divided time series data into 4 patterns using plot, such as stationary, trend, seasonal, and seasonal trend. Sometimes we could find a time series data in a large scale and dimension which is need more time to identify the model of the data. Some forecasting method based on the pattern of time series data like, decomposition, naive, and exponential smoothing. If we use that method, we need check the pattern of time series plot and ACF plot. For example in Exponential Smoothing method we need to determine that time series plot follow model Single Exponential Smoothing (SES), Double Exponential Smoothing (DES), or Winter's (Triple Exponential Smoothing), see Makridakis and Wheelwright (1998). So, in large scale and dimension take 1 by 1 for identify time series pattern need more time. We need some algorithm to classify the pattern of time series data without checking 1 by 1.

Clustering method can be used to make the large scale time series data into a group that identified having the same characteristics. In this research, will be used ACF as the distance for differencing clusters, and the computation is using complete linkage hierarchical algorithm. Some related research has been done before. Montero (2014) ever simulated to identify stationary and non-stationary data by generating 12 type of stationary data and 6 type of non-stationary data. The result is a distance of 19 trials conducted can be concluded that autocorrelation yield had the highest accuracy. While, the cluster algorithm used is complete linkage. Another research, Pierpaolo (2009) in his study, using stationary data and clustering for data switching time series. The distance that used in his study was ACF hierarchy and fuzzy c mean and then compared. It's conclude that fuzzy algorithm can capture the switching stationary time series.

Meanwhile, this research focuses on clustering 4 types of data pattern such as, stationary, trend, seasonal, and seasonal trend using the autocorrelation distance with the complete linkage hierarchical algorithm. Also, we would like to use multistage clustering for classifying data trend and seasonal trend or whether it can be used to classify the four type of data pattern and produce a better conclusion.

2. METHODOLOGY

2.1. Time Series Plot Pattern

Time series model can be easily identify when we see the plot pattern. An important step in selecting an appropriate forecasting method is to consider the types of data patterns, so that the methods most appropriate to those pattern can be utilized, see Makridakis and Wheelright (1998). Commonly, it can be classify into horizontal, seasonal, cyclical and trend.

2.2. Autocorrelation

For a stationary process, we have the mean $E(Y_t)=\mu$ and variance $Var(Y_t)=\sigma^2$. The correlation between and as

$$\rho_k = \frac{cov\,(Y_t\,, Y_{t+k})}{\sqrt{Var(Y_t)}\,\sqrt{Var(Y_{t+k})}} \tag{1}$$

Where we note that $Var(Y_t)=Var(Y_{t+k})$. As function of k, ρ_k is called the autocorrelation function (ACF) in time series analysis because they represent the correlation Y_t and Y_{t+k} from the same process, separated only by k time lags, see Wei (2006).

2.3. Cluster Time Series

Clustering is an unsupervised learning task aimed to partition a set of unlabeled data objects into homogeneous groups or clusters. Partition is performed in such a way that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criterion, see Pablo (2014). For time series modelling, type of cluster can be use is autocorrelation based distance.

Let $\rho_{X_T}=(\rho_{1X_T},\rho_{2X_T}\ldots,\rho_{LX_T})'$ and $\rho_{Y_T}=(\rho_{1Y_T},\rho_{2Y_T}\ldots,\rho_{LY_T})'$ be estimated autocorrelation vector of X_T and Y_T respectively, for some L such that $\rho_{iX_T}\approx 0$ and $\rho_{iY_T}\approx 0$ for $i>L$ define a distance between X_T and Y_T as follows.

$$d_{ACF}(X_T, Y_T) = \sqrt{\left(\widehat{\rho}_{X_T} - \widehat{\rho}_{Y_T}\right)'\,\Omega(\widehat{\rho}_{X_T} - \widehat{\rho}_{Y_T})} \tag{2}$$

Where,

$d_{ACF}(X_T,Y_T)$ is autocorrelation distance between and;

$\hat{\rho}_{X_T}$ is estimation of autocorrelation vector of X_T;

$\hat{\rho}_{Y_T}$ is estimation of autocorrelation vector of Y_T;

$\mathbf{\Omega}$ is weight matrices.

While ACF distance without weight so that weighted matrices be identity matrices. If weight matrices using identity matrices, so the autocorrelation distance become

$$d_{ACFU}(X_T, Y_T) = \sqrt{\left(\widehat{\rho}_{X_T} - \widehat{\rho}_{Y_T}\right)'\,(\widehat{\rho}_{X_T} - \widehat{\rho}_{Y_T})} \tag{3}$$

See

2.4. Hierarchical Cluster Algorithm

One of cluster hierarchical algorithm method is complete linkage, a method renew the similar matrix based on the maximum distance each object, see Johnson and Winchern (2007). The formula to determine the distance between cluster (i,j) and k as follows.

$$d_{(i,j)k} = max\left(d_{ik}, d_{jk}\right) \tag{4}$$

with :

d_{ik} = distance between cluster i and k

d_{jk} = distance between cluster j and k

$d_{(i,j)k}$ = distance between cluster ij and k

3. SIMULATION DATASET

This simulation study, we would like to generate 12 time series data that divided into 4 type pattern which is each dataset's pattern consist of 3 time series data. Dataset's pattern used in this simulation study is stationary, trend, seasonal, and seasonal trend. Variable that we used for each data is detailed below in Table 1 and also the plot from each data shown in Fig. 1.

Table 1. Datasets Name and Time Series Pattern for Simulation Study

Datasets Name	Time Series Pattern
data1,data2,data3	Stationary
data4,data5,data6	Trend
data7,data8,data9	Seasonal
data10,data11,data12	Seasonal Trend

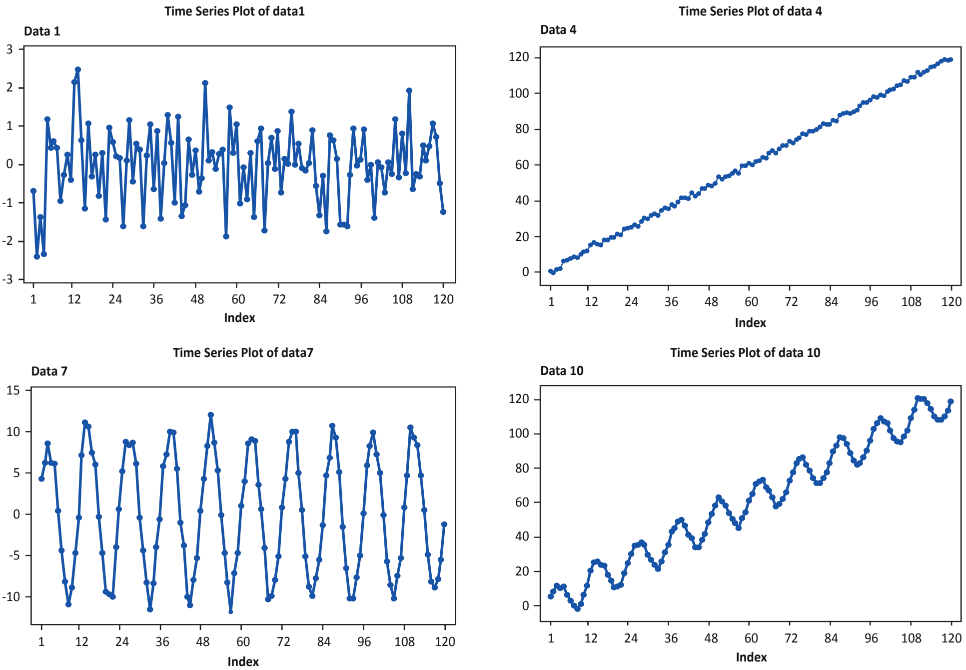


Figure 1. Time Series Pattern for Simulation Study Datasets

4. RESULT

This research, cluster time series divided into 2 stages, first stage cluster could distinguished time series data which have stationary, trend and seasonal. But, for seasonal trend time series data classified into trend time series data. Due to this case we have to use the second stage to separating the cluster between seasonal trend and trend itself, it is called a multistage cluster. Result in clustering for the first stage shown in Fig. 2.

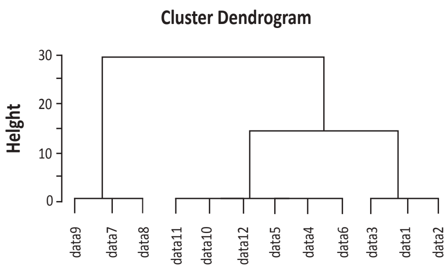


Figure 2. Dendogram Plot for 12 Datasets

The first stage it can be seen that datasets 10-12 (seasonal trend) still classify in the same group with trend's cluster. Then, it'll be need to clustering for the second stage for data 4-6 and 10-12. Before using the second clustering, time series data will be differencing to lag 1. The aim to differencing into lag 1 is for getting a different pattern from each autocorrelation function plot and eliminating trend from the data. Autocorrelation function plot after differencing into lag 1 shown in Fig. 3.

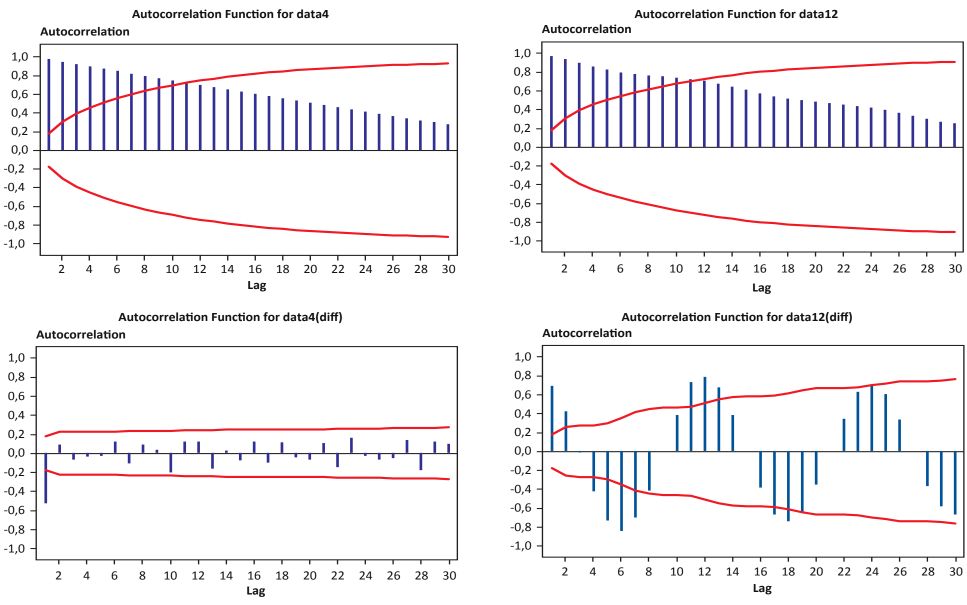


Figure 3. Autocorrelation Function Plot Before and After Differencing

Fig. 3 shown that there is a different pattern between autocorrelation plot before and after differencing into lag 1. It can be seen from the plot, that autocorrelation function plot for "data12" there is a seasonal in every 12 periods. Next, after the differencing, we can do the clustering for phase 2. Fig. 4 shown that after the differencing from cluster time series has already distinguish the pattern between trend and seasonal trend time series data.

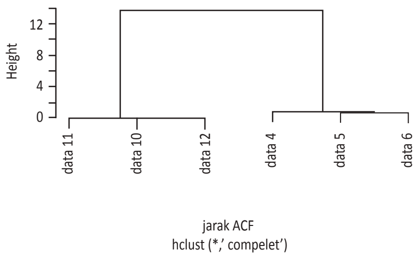


Figure 4. Dendogram Plot for Cluster Stage 2

5. CONCLUSIONS AND DISCUSSION

Based on the simulation study, multistage clustering time series well distinguished stationary, trend, seasonal and seasonal trend time series data. But, this simulation study, the generating data is not based on ARIMA nor SARIMA model. That is why if it is based on those model it might be produce a different conclusion. For the next research it can be used to generate simulation data based on ARIMA or SARIMA model and it will be good to use the real data to produce better conclusion to support this method is good to distinguish the 4 patterns time series data.

REFERENCES

Galeano, P., Pena, D. (2000). Multivariate Analysis in Vector Time Series. Resenhas IME 4, Vol 4, 383-403.

Johnson, R.A., Winchern, D.W. (2007). Applied Multivariate Analysis, 6th Edition. Prentice Hall Inc, New Jersey.

Makridakis, S., Wheelwright, Mc Gee. (1998). Forecasting Method and Application, 3rd Edition). John Wiley & Sons, USA

Montgomery, D.C., Jennings, C.L., Kulahci, M. (2008). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons, USA

Montero, P. (2013). A Package for Stationary Time Series Clustering. Master Thesis, Universidade Da Coruna.

Montero, Pablo & Jose A. Vilar. (2014). TSclust: An R Package for Time Series Clustering. Journal of Statistical Software, Vol 62, Issue 1

Pierpaolo, D., Maharaj, E.A. (2009). Autocorrelation-based Fuzzy Clustering of Time Series. Fuzzy Sets and System 160, 3565–3589.

Wei, W.W. (2006). Time Series Analysis : Univariate and Multivariate Methods, 2nd edition. Pearson, New York.

A Case Study Comparison of Three Classification Methods in Order to Identify Financial Losses from Electricity Theft

Sisa Pazi *
NMMU, Port Elizabeth, South Africa – sisa.pazi@nmmu.ac.za
Chantelle Clohessy
NMMU, Port Elizabeth, South Africa – chantelle.clohessy2@nmmu.ac.za
Gary Sharp
NMMU, Port Elizabeth, South Africa – gary.sharp@nmmu.ac.za

Abstract

Electricity theft is one of the components of non-technical losses encountered in electrical power systems. This study presents a proposed model for the detection of non-technical losses in Nelson Mandela Bay Municipality, Eastern Cape, South Africa. The proposed model was designed using the Support Vector Machines (SVM), Naive Bayes (NB) and k-Nearest Neighbour (KNN) classification algorithms. The model's ability to detect fraudulent activities in electricity consumption, which are correlated to non-technical losses (Nagi et al.,2008), was tested and evaluated using the fraud detection rate and the error rate. The SVM classifier achieved the highest fraud detection rate, followed by KNN classifier and the least performing classifier was the NB.

Keywords: Electricity theft; Support Vector Machines; Naïve Bayes; k-Nearest Neighbour.

1. INTRODUCTION

Electricity theft is on the rise in the world with electric utilities losing billions in revenue. Practical application of statistical methods can help in identifying and predicting electricity theft. Electricity theft is one of the components of non-technical losses encountered in electrical power systems (Depuru, Wang, & Devabhaktuni, 2011). Non-technical losses refer to the energy consumed but not billed (Navani, Sharma, & Sapra, 2012). These losses occur as a result of billing irregularities and electricity theft (in the form of energy meter tampering, illegal connections, etc.). This paper presents a method for detecting electricity theft based on a customer's electricity consumption pattern over a period of time.

2. LITERATURE REVIEW

Methods to detect electricity theft have been proposed by researchers. This section presents some of these methods. Monedero et al. (2012) used Pearson's correlation coefficient, Bayesian Networks and Decision Trees (DT) to detect non-technical losses in the electrical power system of Endesa Company in Spain. The proposed model identified patterns of large decreases in electricity consumption. The data set used consisted of 38 575 customers for the period of two years from July 2008 to June 2010. The three methods were used to preselect potential fraudsters for onsite inspection. Upon inspection, the percentage of real fraudsters was found to be 38%. The Support Vector Machines (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbour (KNN), Optimal-Path Forest (OPF) and Kohonen Self Organising Maps (SOM) were compared by Ramos et al. (2011) to identify electricity theft. Two data sets from a Brazilian electrical company were used to train and test these classifiers. One data set comprised of industrial consumers and the other of commercial consumers. The former was a data set consisting of 3486 consumers and the latter contained 5645 consumers. The OPF and KNN classifiers achieved similar accuracy for both data sets and outperformed all the other classifiers, where accuracy is defined as the percentage of all the consumers that are correctly classified. (Kohavi, 1995).

Depuru et al. (2011) addressed some of the factors associated with the detection of electricity theft such as; geographical location, size of customer [agricultural (small, medium), commercial (small, medium, large) and residential (very small, small, medium, large)] and, season of the year. In their study, the SVM classifier was trained and tested on data sets collected by smart meters in different geographical locations in India. Electrical energy consumption patterns under normal operations and under electricity theft were studied and customers classified into three groups. A training set consisting of 440 customers was used to train the SVM classifier. Then, the trained SVM classifier was used on a test set consisting of 220 customers. The trained SVM classifier achieved an accuracy of 98.4%.

Nagi et al. (2008) conducted a study to detect non-technical losses using the SVM in the power system of Tenaga Nasional Berhad in Malaysia. The proposed SVM classifier identified transactions of potential fraudulent customers in order to facilitate an onsite inspection. This study reported that out of the total number of candidates shortlisted by the SVM classifier, 53% were confirmed, upon inspection, to be fraudulent cases. A similar study by the same authors was conducted in 2010 (Nagi et al.,2010). The main objective of this study was to improve the hit rate achieved in the 2008 study. In addition to monthly energy consumptions, theft of electricity information, credit worthiness rating information, high risk customer information and irregularity report information, were all included in the fraud detection model. The inclusion of the additional data in the fraud detection model improved the hit rate from 53% to 60% (Nagi et al, 2010).

These studies demonstrate an extensive use of data mining techniques to identify and detect electricity theft in customers' consumption. Many research studies have been conducted internationally; however there is insufficient evidence of research studies that used data mining algorithms for the detection of non-technical losses in South Africa. Studies done in South Africa include, Davidson (2002), who proposed a method to estimate costs associated with non-technical losses. The relationship between the distribution of data and a classifier's performance was studied by Van der Walt and Barnard (2006). Doorduin et al. (2004) used simulations to discuss the feasibility of using remote check meters to measure the magnitude of electric energy losses and electricity theft in a low voltage reticulation network. The dearth of research in the South African domain, and the need to address the surge in electricity theft in the NMBM led to undertaking this study.

3. DATA ACQUISITION

The data was obtained from the Electricity and Energy Directorate of the NMBM. This consisted of customers' electricity consumptions for 24 months from March 2013 to February 2015. A portion of this data is depicted in Table 1. Each customer was classified as either clean or fraudulent, based on a random inspection undertaken by the NMBM. Clean customers were defined as customers that were paying for the energy they consumed. In contrast, fraudulent customers were those that were found to have tampered with their energy meter or had performed illegal connection. There was a total of 3156 customers, of which 2420 (77%) were categorised as clean and 736 (23%) as fraudulent.

Table 1. Empirical Data Used for Training and Testing

Customer	13-Mar	13-Apr	13-May	···	15-Feb	Class label
Customer 1	222.3	209	182.2	···	148.4	Fraudulent
Customer 2	240.5	192.5	250.2	···	334.4	Clean
Customer 3	459.3	419.1	378.8	···	711	Clean
Customer 4	128.5	128.5	128.5	···	121.5	Fraudulent
Customer 5	957.2	889.2	862.8	···	150.7	Clean
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Customer 3156	134.9	158	94.5	···	83.6	Clean

4. RESEARCH METHODOLOGY

Data mining is a form of data analysis that is used to extract useful information from large data sets (Jaya & Tamilselvi, 2013). Such analyses are used to build models or classifiers to describe important data classes. The aim of this study was to use data mining techniques to detect fraudulent usage of electricity in the NMBM. Three data mining techniques were used, that is, Support Vector Machines (SVM), Naïve Bayes (NB) and the k-Nearest Neighbour (KNN). Three classifiers were developed and compared using these techniques.

4.1. The Support Vector Machines Method

Given a training data set of d observations (in this case, customers), x_i with class labels, y_i:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2n} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dn} & y_d \end{pmatrix},$$

$$\text{where } \begin{cases} y = +1 & \text{if } x_i \epsilon \text{ Fradulent} \\ y = -1 & \text{if } x_i \epsilon \text{ Clean} \end{cases}$$

Training a SVM classifier involves minimising the error function (Boser et al.,1992):

$$\frac{1}{2} w^T w + C \sum_{i=1}^{d} \xi_i ,$$

subject to the following constraints:

$$y_i\left(w^T \Phi(x_i) + b\right) \geq 1 - \xi_i, \qquad i = 1,2, \dots, d,$$

where C is the capacity constant, w and $\Phi(x)$ are n-dimensional vectors, b is a constant and ξ_i represent parameters for handling non-separable data. After training, a customers with unknown class is predicted using the following rule:

$$x_k \,\epsilon \text{ Fradulent} \qquad \text{if } D(x_k) > 0$$

$$x_k \,\epsilon \text{ Clean} \qquad \text{otherwise}$$

where $D(x_k) = w^T \Phi(x_k) + b$.

4.2. The Naïve Bayes Method

Given a customer to be classified, represented by a vector $x=(x_1,x_2,\dots,x_n)$ representing some and measurements (independent variables), the NB classifier assign to this customer probabilities, $P(Fraudulent|x)$ and $P(Clean|x)$ (Lewis, 1998). These probabilities are computed using the Bayes Theorem. Whenever the $(Fraudulent\ |x)>P(Clean|x)$, that customer is classified as fraudulent and if $P(Fraudulent|x)< P(Clean|x)$ as clean.

4.3. The k-Nearest Neighbour Method

The KNN classifier is a highly efficient and effective technique that is easy to implement (Nitin Bhatia, 2010). This method is effective when dealing with large training data set (Mulak & Talhar, 2013). Mulak and Talhar (2013) described the method as being robust and with small error rate.

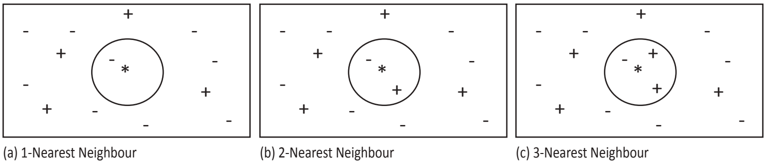


Figure 1. One, Two and Three Nearest Neighbour

In the KNN method, customer's class is predicted based on the class of its nearest neighbours. Figure 1 shows various nearest neighbours of a customer represented by the star at the centre of each circle. In

Figure 1 (a) the nearest neighbour has a negative class; therefore the negative class label is assigned to the customer. In case of a tie between the two classes, see Figure 1 (b), a random class is chosen for the customer. Three nearest neighbours are present in Figure 1 (c), one is negative and two are positive, the customer is assigned a positive class label using a majority voting.

4.4. Framework for Customer Classification

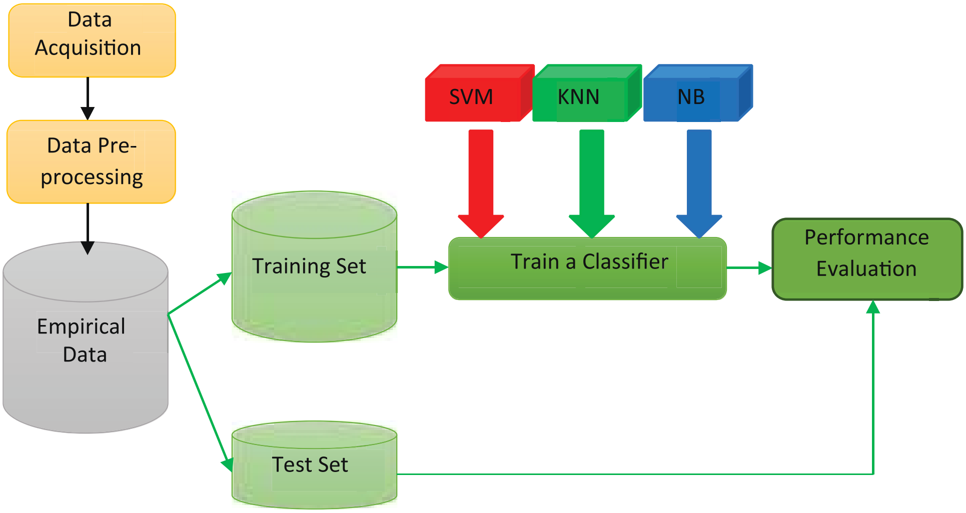


Figure 2. General Framework of the Approach

The methodology followed in proposing the fraud detection model is illustrated in Figure 2. After data acquisition, the data was pre-processed in preparation for the subsequent analysis. During the pre-processing process, the data was normalised to bring all the values on the same scale. The data was then randomly partitioned into two independent sets, i.e. training and test sets. The training set was used to derive a classifier. Three classification algorithms were used during the training phase, namely, SVM, NB and KNN. The performance of each classifier was then evaluated using the test set. The analytical routine in Figure 2 was repeated 500 times; each time performance measures were calculated. Performance measures are discussed in the following section.

4.5. Performance Evaluation

There are different measures to evaluate the performance of a classifier (Mulak & Talhar, 2013). In this study sensitivity, specificity and error rates were used to evaluate performance of each classifier. Mulak and Talhar (2013) defined true positives as positive tuples that were correctly classified as positive, true negatives were defined similarly. In this study, true positives referred to the fraudulent customers that were classified as fraudulent and true negatives as clean customers that were correctly classified as clean. Performance measures were then defined in terms of true positives and negatives as follows:

$$sensitivity = \frac{number\ of\ true\ positives}{total\ number\ of\ positives},$$

$$specificity = \frac{number\ of\ true\ negatives}{total\ number\ of\ negatives},$$

and,

$$error\ rate = \frac{number\ of\ false\ positives + number\ of\ false\ negatives}{total\ number\ of\ positives + total\ number\ of\ negatives}$$

5. RESULTS AND DISCUSSION

The experimental results are summarised in Figures 3-6. Depicted in Figure 3 is the comparison of the SVM, NB and KNN classifiers in terms of sensitivity. It is evident that the SVM classifier outperformed its competitors which had similar sensitivity. In Figure 4, the three classifiers are compared using specificity. The KNN classifier had the highest specificity, followed by the SVM classifier and the NB classifier had the lowest specificity.

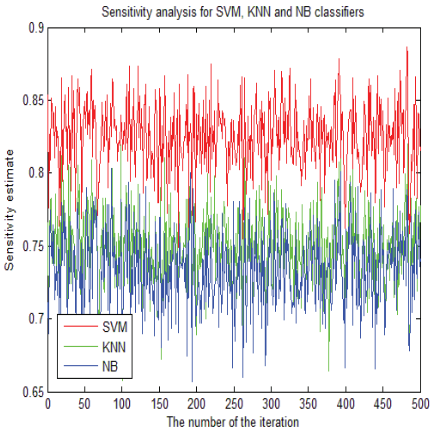


Figure 3. Comparison using Sensitivity

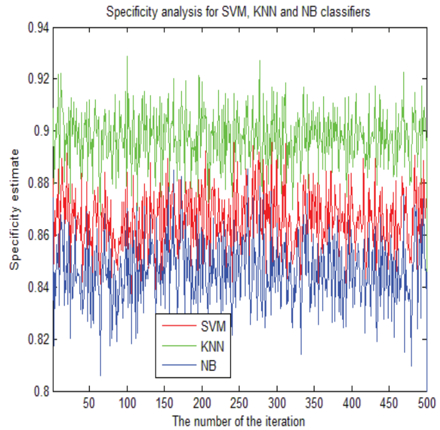


Figure 4: Comparison using Specificity

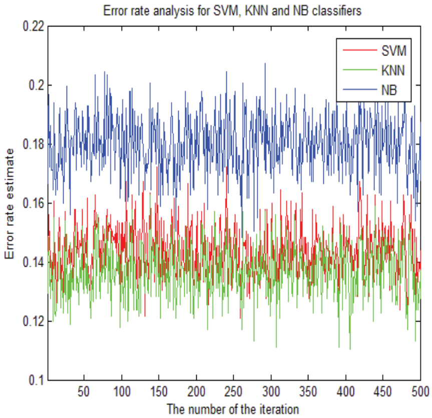
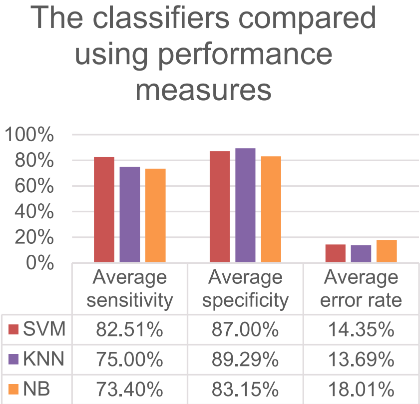


Figure 5. Comparison using Error Rates



	Average sensitivity	Average specificity	Average error rate
SVM	82.51%	87.00%	14.35%
KNN	75.00%	89.29%	13.69%
NB	73.40%	83.15%	18.01%

Figure 6. Averages for Performance Measures

The error rates for the three classifiers are shown in Figure 5, the SVM and KNN classifiers had jointly the lowest error rates and the NB classifier had the highest error rate. The averages of each performance measure, for each classifier, are shown in Figure 6.

6. CONCLUSIONS

The purpose of this research study was to find a suitable model to be applied in the detection of fraudulent electricity usage in NMBM. Such model would predict fraudulent activities in electricity consumptions by detecting any abnormal usage patterns. Results showed that the SVM classifier outperformed the other classifiers. The SVM classifier was then used for efficient detection of fraudulent electricity consumers. This proposed method proved to be reliable with a hit-rate of over 80%. The use of this model benefited the municipality in the assessment of its non-technical losses.

REFERENCES

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Fifth annual workshop on Computational learning theory* (pp. 144-152). Pittsburgh, Pennsylvania, USA : ACM.

Davidson, I. (2002). Evaluation and effective management of nontechnical losses in electrical power networks. *Africon Conference in Africa, 2002* (pp. 473-477). Pretoria, South Africa: IEEE.

Depuru, S. S., Wang, L., & Devabhaktuni, V. (2011). Support vector machine based data classification for detection of electricity theft. *Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES* (pp. 1-8). Phoenix, Arizona, USA: IEEE.

Doorduin, W. A., Mouton, H. D., Herman, R., & Beukes, H. J. (2004). Feasibility study of electricity theft detection using mobile remote check meters. *7th AFRICON Conference in Africa* (pp. 373-376). IEEE.

Jaya, B. J., & Tamilselvi, J. J. (2013). Assessment of fraud pretentious business region research articles using data mining approaches. *International Journal on Computer Science and Engineering*, 653-659.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *14th international joint conference on Artificial intelligence-Volume 2* (pp. 1137-1145). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.

Lewis, D. D. (1998). The independence assumption in information retrieval. *European conference on machine learning* (pp. 4-15). Chemnitz, Germany: Springer Berlin Heidelberg, 1998.

Monedero, I., Biscarri, F., Leon, C., Guerroro, J., Biscarri, J., & Millan, R. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson's coefficient, Bayesian network and decision trees. *Electrical Power and Energy Systems, 34*(1), 90-98.

Mulak, P., & Talhar, N. (2013). Analysis of distance neasures using k-nearest neighbour algorithm on KDD dataset. *International Journal of Science and Research, 4*(7), 2101-2104.

Nagi, J., Mohammad, A. M., Yap, K. S., Tiong, S. K., & Ahmed, S. K. (2008). Non-technical loss analysis for detection of electricity theft using support vector machines. *Power and Energy Conference, 2008* (pp. 907-912). Johor Bahru, Malaysia: IEEE.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohammad, M. A. (2010). Non-technical loss detection for metered customers in power utility using support vector machines. *IEEE transactions on Power Delivery, 25*(2), 1162-1171.

Navani, J. P., Sharma, N. K., & Sapra, S. (2012). Technical and non-technical losses in power system and its economic consequence in Indian economy. *International Journal of Electronics and Computer Science Engineering, 1*(2), 757-761.

Nitin Bhatia, V. (2010). Survey of nearest neighbour techniques. *International Journal of Computer Science and Information Security, 8*(2), 302-305.

Ramos, C., Souza, A., Chiachia, G., Falcao, A., & Papa, J. (2011). A novel algorithm for feature selection using Harmony Search and its application for non-technical losses detection. *Computers and Electrical Engineering, 37*(6), 886-894.

Van Der Walt, C., & Barnard, E. (2006). Data characteristics that determine classifier performance. *SAIEE Africa Research Journal, 98*(3), 87-93.

Support Vector Machines with Adaptive Fruit Fly Optimization Algorithm based on Velocity Variable (VFOA) for Classifying High Dimensional Data

Mukhlis
Badan Pusat Statistik, Jakarta, Indonesia – m.mukhlis@bps.go.id
Bony Parulian Josaphat*
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – bonyp@stis.ac.id

Abstract

A significant advancement of technology has many impacts, such as high dimensional data's existence and its problems. High dimensional data is the situation in which the sample size is approached or exceeded by the number of variables used in analysis. One of the classification methods successfully handling high dimensional data classification is Support Vector Machines (SVM). Its performance depends on the kernel's parameter values. The values are determined by using grid search optimization algorithm. Another impact of technology advances is metaheuristic algorithm development as an optimization algorithm, like an adaptive Fruit Fly Optimization Algorithm based on Velocity Variable (VFOA). VFOA is an enhancement algorithm of Fruit Fly Optimization Algorithm (FOA) with adding velocity variable to enhance the precision and the convergence. Therefore, the goal of this study was to improve SVM performance by applying VFOA in determining parameter values in high dimensional data to social science researchers. The result shows that the measurement tools of SVM optimized by VFOA outperforms grid search. Thus, VFOA can boost SVM performance well in classifying high dimensional data.

Keywords: SVM; FOA; VFOA; classification.

1. INTRODUCTION

Technology is rapidly developing and has many impacts. It sustains the presence of high dimensional data and its problems. High dimensional data is the situation in which the sample size is approached or exceeded by the number of variables used in analysis (Bühlmann, P. & Van De Geer, S., 2011). While in statistical method and data mining, supervised processing such as classification method of high dimensional data is still becoming a challenge due to a phenomenon, curse of dimensionality (Escalante, A. N., & Wiskott, L., 2013).

One of the best method in classifying high dimensional data is SVM. It is evidenced by Furey, T. S. et al. (2000) who asserted that SVM performed well in classifying microarray expression data. Similarly, Bing Liu (2007) conveyed that SVM was a type of machine learning having solid theoretical foundation and performing more accurately than most other algorithms in applications involving very high dimensional data. Nevertheless, SVM is sensitive to the kernel's parameter values selection (Zhang, D. et al. 2007). In order to get a better performance, an optimization algorithm can be used to select the appropriate parameters. An existing algorithm is grid search method. Yet, there are a lot of computational optimization algorithms concomitantly with a significant advancement of technology, such as FOA, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).

FOA proposed by Wen-Tsao Pan is a simple and an easy to understand optimization algorithm. Zhang, Q. et al. (2015) showed that it boosted SVM performance better than the other algorithms. Yet, it still has shortcomings, such as being easily trapped into local optima and having low accuracy. In consequence, Lu, M. et al. (2015) proposed an adaptive FOA based on velocity variables (VFOA) with adding a change in updating location of fruit flies. Hence, this research applies VFOA in selecting the parameters of SVM. It boosts SVM performance in classifying the high dimensional data in social field.

2. METHODS AND MATERIALS

2.1. Support Vector Machines

Support Vector Machines (SVM) has a solid theoretical foundation to classify accurately in many conditions of data, especially in high dimensional data (Bing Liu, 2007). Let the set of training data D be $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is an input vector in a real valued space $X \subseteq \Re^r$ and y_i is a label class with $y_i \in \{1, -1\}$. In building a classifier, SVM finds a linear function of the form

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \tag{1}$$

where $\mathbf{w}$ is called the weight vector and b is called as the bias. Hence, an input vector is classified to the positive class if $f(\mathbf{x}) \geq 0$ and to the negative one if $f(\mathbf{x}) < 0$. Basically, SVM finds a hyperplane

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0 \tag{2}$$

separating positive and negative training data that is called as the decision boundary. There is an infinite number of hyperplanes that can separate the data. SVM chooses the hyperplane maximizing the margin between the positive and the negative data points.

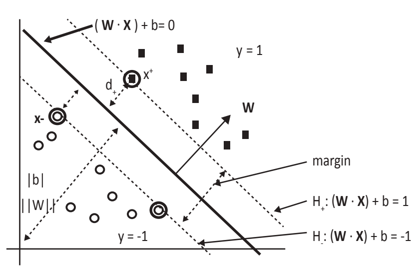


Figure 1. Separating hyperplane (Bing Liu, 2007)

In the linearly separable case, the margin of the separating hyperplane is the distance between the nearest data points called support vectors from the positive and negative class. From Fig. 1, the margin will be

$$margin = d_+ + d_- = \frac{2}{\|\mathbf{w}\|} \tag{3}$$

Maximizing the margin is the same as minimizing $\|\mathbf{w}\|^2/2$ with subject to $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$, and being solved with a Lagrange multiplier

$$L_P = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_{i=1}^{n} \alpha_i \left[y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 \right] \tag{4}$$

Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient for an optimal solution. It is used to solve the primal Lagrangian (L_p) which can be simplified into dual formulation

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \tag{5}$$

where $\alpha_i \geq 0$ and under constrain $\sum_{i=1}^{n} y_i \alpha_i = 0$.

Quadratic programming can be applied to solve the maximization problem above. The final decision boundary is

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \sum_{i \in sv} y_i \alpha_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b = 0 \tag{6}$$

In most cases, data is not linearly separable and has to be mapped into a higher-dimensional feature space. The idea is mapping the data in the input space X to a feature space F via a non linear mapping $\phi : \mathbf{x} \to \phi(\mathbf{x})$. Apllying the idea, Equation (5) will be shown as follow:

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle \tag{7}$$

with respect to $\sum_{i=1}^{n} y_i \alpha_i = 0$, and $0 \le \alpha_i \le C$.

From the equation above, there is a simple way to solve the dot products $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$ without having to know the feature vector $\phi(\mathbf{x})$ or even the mapping function ϕ itself. In SVM, this can be tricked with the use of kernel function, denoted by K,

$$K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle \tag{8}$$

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003) stated that there are kernel functions commonly found in SVM, such as linear, radial basis function (RBF), polynomial, and sigmoid kernel. Prior research from Gaspar, P., Carbonell, J., & Oliveira, L. (2012) found that the first three kernels gave the best results in SVM. In addition, RBF is the one having a better accuracy and convergence time than polynomial kernel while linear kernel giving a better result in classifying bioinformatics data which was known from the research conducted by Ben-Hur, A. & Weston, J. (2010). Thus, in this research, linear and RBF are used in order to applying optimization algorithm in SVM.

2.2. Adaptive Fruit Fly Optimization Algorithm based on Velocity Variable (VFOA)

FOA is a new optimization algorithm proposed by Wen-Tsao Pan in 2011 based on feeding behavior of fruit fly (Pan, 2011). It is simple, having a fast convergence speed, and easy to understand and also to implement. On the contrary, it also has some shortcomings, such as easily trapped into local optima and low accuracy. Lu, M. et al. (2015) proposed an improved FOA based on the flight characteristics of fruit fly, improving convergence speed and accuracy. The improved adaptive FOA based on velocity variable (VFOA) has a changing of velocity and position with the following equations:

$$V_{ix} = w * V_{ix} + c_1 \operatorname{rand}()\left(P_{ix} - X_i\right) + c_2 \operatorname{rand}()\left(G_x - X_i\right) \tag{9}$$

$$V_{iy} = w * V_{iy} + c_1 \operatorname{rand}()\left(P_{iy} - Y_i\right) + c_2 \operatorname{rand}()\left(G_y - Y_i\right) \tag{10}$$

$$X_i = X_i + V_{ix} \tag{11}$$

$$Y_i = Y_i + V_{ix} \tag{12}$$

where w is inertia weight, c is an acceleration constant, P_{ix} and P_{iy} is i-th fruit fly experienced the best position on the x and y direction. G_x and G_y is all fruit fly experienced the best position on the x and y direction.

These are the steps of the algorithm:

Random initial fruit fly swarm location.

$Init\ X_{axis}$ and $Init\ Y_{axis}$

Giving the random direction and distance for the search of food using smell by an individual fruit fly.

$$X_i = X_{axis} + Random\ Value$$

$$Y_i = Y_{axis} + Random\ Value$$

Estimating the distance (*Dist*) of the food to the origin and calculating the smell concentration judgement value (*S*).

$$Dist_i = \sqrt{X_i^2 + Y_i^2}$$

$$S_i = \frac{1}{Dist_i}$$

Substituting smell concentration judgment (*S*) into function so as finding the smell concentration of the individual location of the fruit fly.

$$Smell_i = \text{Function}(S_i)$$

The function used in selecting SVM parameters is the accuracy rates of the 10-fold cross validation or jackknife.

Finding the maximal value among the fruit fly swarm.

$$[bestSmell\ bestIndex] = \max(Smell)$$

Keeping the best smell value and x, y coordinate. The fruit fly swarm will use vision of the fruit fly which have the best smell concentration to fly towards the location.

$$Smellbest = bestSmell$$
$$X_{axis} = X(bestIndex)$$
$$Y_{axis} = Y(bestIndex)$$

Updating velocities and locations with Equation 9 to 12.

Implementing step 3 to step 5, then judging if the smell concentration is superior to the previous one. If so, implement step 6.

If the iterations attain to the maximum number, the search is stopped, if not, jump to step 7 to continue the search.

In this research, the population of the fruit fly is 20 with 100 iterations. The values of *w* and *c* are 0.76 and 3, recommended by Lu, M. et al. (2015).

3. MATERIALS

Data simulation is generated with R software based on the manipulation variables in order to know the performance of the classification method in the social high dimensional data. The plan is referred to Finch, H. (2014) that consists of 5 manipulation variables with a bit modification. First, data is generated having 5 conditions of sample size variable, that is 12, 20, 30, 40, 50. The other variable is number of predictors which is conditioned to be 14, 28, and 50 predictors or variables of the data. The data simulation has group mean separation as the other manipulation variable. There are separation between group means quantified in terms of Cohen's d effect size. The group means separation is differed by same amount, either 0.2, 0.5, and 0.8. The data is also simulated having 3 conditions of correlation among variables, which are 0, 0.5, and 0.8, and having 2 distributions, multivariate normal and skewed with skewness of 2.5. The simulation conditions created 270 datasets as soon as there were completely crossed one another.

In the classification process, the methods used in this research are SVM kernel linear optimized by grid search, (SVM GS Linear), SVM kernel RBF optimized by grid search (SVM GS RBF), SVM kernel linear optimized by VFOA (SVMVFOA Linear), and SVM kernel RBF optimized by VFOA (SVMVFOA RBF). Besides that, 10-fold cross validation was applied to get a better precision in the process. For SVM optimized by VFOA, it applied nested 10-fold cross validation for the data having 30-50 sample size, and applied jackknife – 10-fold cross validation. The first validation, jackknife or cross validation, is used as a measurement of selecting the parameter. In other word, it was used as a function of smell concentration VFOA. The classification accuracy rate were used to see the performance of SVM and the optimized one.

4. RESULTS AND DISCUSSION

The classification was executed with the conditions in the previous sections. The results of the executions show the performance of methods in classifying high dimensional data in social field. In figure 2, the accuracy classification rates are presented by the manipulation variables of correlation among variables (CR) and group mean separation (CH). It shows that all methods have a better performance in data consisting no correlation among variables. The classification accuracy rate keeps decreasing when the correlation among variables keeps increasing. On the contrary, the rates are increasing simultaneously with the group mean separation. Thus, all methods have good performance in data having a big group mean separation of the classes.

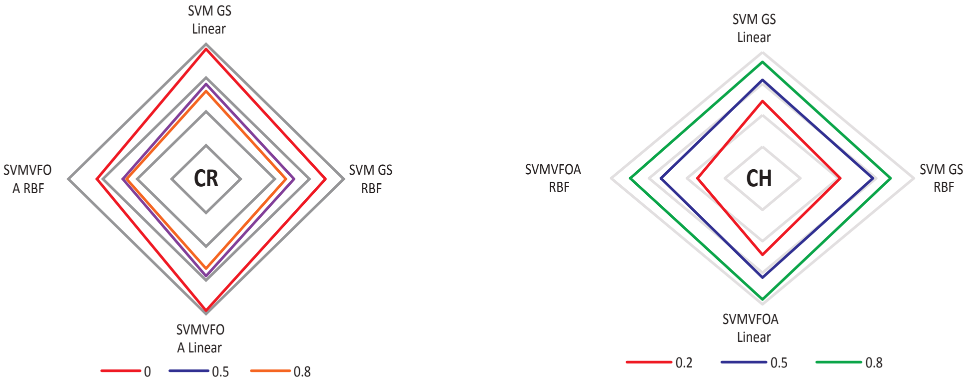


Figure 2.
Classification Accuracy Rates by Method, Corelation among the Variables (CR) and Group Mean Separation (CH)

From Table 1, overall classification accuracy rates increase synchronically with the sample size increment. All methods have lower accuracy rates in the small sample size and higher in the big one. They have the rates lying on 0.582 on the average of classifying the data simulation. The lowest accuracy belongs to SVMVFOA RBF, particularly for the smallest sample size. For the extreme one, it performs worse in the condition having 12 sample size and 50 variables with 0.12 classification accuracy rate. Besides that, the best accuracy rate is 0.715 from SVMVFOA Liner in the condition having 30 sample size and 28

Table 1.
Classification Accuracy Rates by Method, Sample Size (SZ), and Amounts of Predictors or Variables in Data (PR)

SZ	PR	SVM GS Linear	SVM GS RBF	SVMVFOA Linear	SVMVFOA RBF
12	14	0.505	0.403	0.602	0.347
	28	0.505	0.435	0.556	0.319
	50	0.597	0.5	0.602	0.12
20	14	0.569	0.592	0.578	0.571
	28	0.625	0.536	0.589	0.547
	50	0.661	0.533	0.622	0.408
30	14	0.6	0.522	0.609	0.576
	28	0.707	0.57	**0.715**	0.611
	50	0.656	0.574	0.676	0.594
40	14	0.604	0.59	0.631	0.629
	28	0.644	0.636	0.618	0.625
	50	0.643	0.606	0.647	0.594
50	14	0.641	0.586	0.627	0.641
	28	0.656	0.641	0.679	0.689
	50	0.653	0.619	0.678	0.63

Note: (The coloured cells on the table show the best accuracy in SZ and PZ. Red for RBF and blue for Linear) variables. In addition, SVMVFOA Linear performs better in classifying process than the other methods with its biggest classification accuracy rates.

Unfortunately, as mentioned above in RBF kernel, VFOA could not find a better combination of parameters than grid search optimization algorithm. It performed very bad in classifying small sample sized data simulation. VFOA sustains difficulties in assigning cost and gamma as parameters in RBF. It is caused by a dependent progress. Treatment used in the process is the same for each parameter, while they are in a different range. Besides, small sample sized data can be the other difficulties. It makes VFOA being insensitive for changes of parameters caused by a limited universal set of $Smell_i$.

Generally, the rates on Table 1 also show that VFOA enhance SVM performance. It is indicated by the best classification method in classifying the data, SVMVFOA Linear. It has the best accuracy rate which is lying on 0.628. In the process, VFOA found and set an appropriate parameter for SVM so that it can boost the performance of classifying. Besides that, VFOA outplays grid search optimization in 9/15 conditions from kernel RBF. While in the other kernel, Linear, 73% conditions are won by SVMVFOA RBF from SVM GS RBF. Therefore, in proportion, 66.67% of data conditions are classified better using VFOA than grid search optimization algorithm.

5. CONCLUSIONS

Support Vector Machines performance is affected by the kernel and its parameters. They take an important role affecting the results. In classifying high dimensional data, linear kernel is more suitable than RBF kernel. VFOA as an improvement algorithms of FOA selects a better combination of parameters. Thus, SVM optimized by VFOA have a better accuracy than SVM optimized by commonly used optimization algorithm, grid search. In conclusion, VFOA jacks up SVM performance well in classifying high dimensional data.

REFERENCES

Ben-Hur, A. & Weston, J. (2010). *A User's Guide to Support Vector Machines*. Data mining techniques for the life sciences, 223-239.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory, and Applications*. Springer Science & Business Media.

Escalante, A. N., & Wiskott, L. (2013). How to solve classification and regression problems on high-dimensional data with a supervised extension of slow feature analysis. *Journal of Machine Learning Research*, *14*(1), 3683-3719.

Finch, H. (2014). *A Comparison of Methods for Group Prediction with High Dimensional Data*. Journal of Modern Applied Statistical Methods: Vol. 13: Iss.2, Article 5.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906-914.

Gaspar, P., Carbonell, J., & Oliveira, L. (2012). *On the Parameter Optimization of Support Vector Machines for Binary Classification*. Journal of Integrative bioinformatics, 0(3): 201.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Lu, M., Zhou, Y., Luo, Q. & Huang, K. (2015). *An Adaptive Fruit Fly Optimization Algorithm Based on Velocity Variable*. International Journal of Hybrid Information Technology, Vol.8 No.3, 329-338.

Pan, W. (2011). *Fruit Fly Optimization Algorithm*. Knowledge-Based System.

Zhang, D., Huang, H., Chen, Q., & Jiang, Y. (2007). *A comparison study of credit scoring models*

Zhang, Q., Fang, L., Ma, L., & Zhao, Y. (2015). *Research on Parameters Optimization of SVM Based on Improved Fruit Fly Optimization Algorithm.* International Journal of Computer Theory and Engineering, Vol.8 No. 6, 500-505.

CPS33: Environmental & Natural Resources Statistics (2)

Statistical Modeling for Wind Direction and Velocity in Thailand

Marzukee Mayeng[1,2*]
[1]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand – yee0007@gmail.come
[2]Centre of Excellence in Mathematics, Commission on Higher Education, Ratchathewi, Bangkok, 10400 Thailand.
Nittaya McNeil[1,2]
[1]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand– nchirtki@gmail.come
[2]Centre of Excellence in Mathematics, Commission on Higher Education, Ratchathewi, Bangkok, 10400 Thailand.
Somporn Chuai-aree[1,2]
[1]Department of Mathematics and Computer Science; Faculty of Science and Technology, Prince of Songkla University, Muang, Pattani, 94000 Thailand – schuaiaree@gmail.com
[2]Centre of Excellence in Mathematics, Commission on Higher Education, Ratchathewi, Bangkok, 10400 Thailand.

Abstract

Movement of winds on the Earth's surface is both a threat that can cause enormous damage and an important renewable energy resource that could be productively harnessed on wind farms. It is thus important to understand how to analyze characteristics of winds using available data from meteorological stations. These characteristics include wind speed, direction and time. In this study we focus on five minute data collected at weather stations in Thailand and converted to hourly data. We use graphical displays and statistical models to handle time-series autocorrelation, non-normality, and speed-direction association. Conclusions suggest that applying logistic regression to averages in four-hourly periods for different speed-direction sectors provides an appropriate methodology, and areas under ROC curves are useful for comparing goodness-of-fit of models.

JEL classification: Q42
Keywords: Wind characteristics, Wind statistical modeling

1. INTRODUCTION

In meteorology, winds are often referred to their strength, and the direction from which the wind is blowing. Wind speed is affected by a number of factors and situations, operating on varying scales. These include the pressure gradient and local weather conditions. There are also links to be found between wind speed and wind direction. It is important to understand the characteristics of winds using available data from meteorological stations because wind is also an important renewable energy resource. Wind has been the fastest growing source of electricity generation in the world since the 1990s (Schreck et al., 2008). In Thailand, the energy consumption is to jump by 75% over next two decades according to the Thailand Power Development Plan of 2015 power energy demand is rapidly increasing in the strong local paper/pulp sub-sector. At the transport level, several high speed train projects shall be realized which have also not been considered in the calculations (The National Energy Policy Council of Thailand, 2015).

The rising cost and negative environmental effects of nonrenewable energy sources and the energy demand in Thailand is rising rapidly, have led to increase research of alternative energy. Wind is one of alternative renewable energy resource which benefit of being more environmentally friendly. However, most of previous studied been reviews about wind distribution and found that much of the literature on the analysis of wind data focuses on the distribution of wind ***speed***, often assumed to have a Weibull distribution.

Therefore, this study proposes an alternative approach to analyse wind characteristics include wind speed, direction and time. In this study we focus on hourly data collected at weather stations in Pattani, Thailand.

2. DATA

The data were collected from automatic weather stations of Thailand. The data was convert to available file and managed.

2.1. Displaying the Data

Since the data obtained was recorded over time. Then, to reveal wind data pattern time-series plot took part for this first step. Graphical displays of time series data are presented.

Developing a time-series data should always begin with graphical display and analysis of the available data. Many of the broad general features of a time series can be seen visually (Douglas C. Montgomery et al., 2008). The basic graphical display for time series data is the time series plot, illustrated in figure 1. The upper panel shows wind speed in metres per second (y-axis). While the x-axis of upper panel graph is time. The lower panel shows wind direction and the x-axis is time. From Direction graphs have jumps due to the restriction between 0 & 360 degrees.

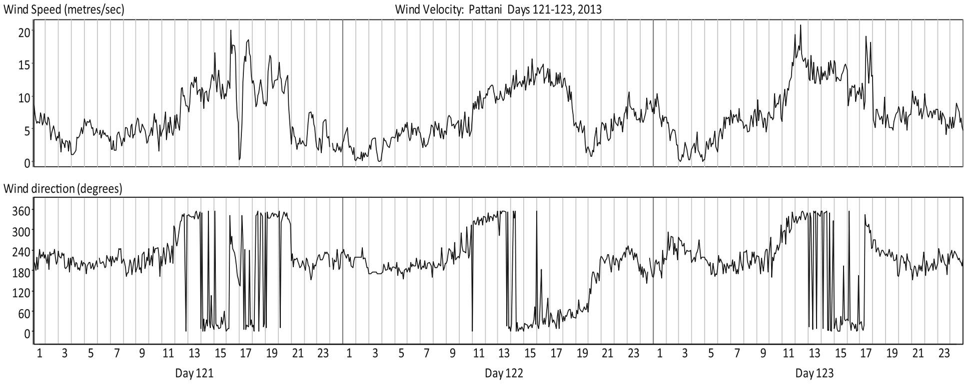


Figure 1. Time-series Plot of Hourly Data for Wind Direction and Velocity in 2013 (3 Days)

The continuous-time graphs of wind direction on an absolute scale (that is, not constrained to 0-360 degrees) using vertical bars to code speed, was improved as shown in figure 2. The graph shows wind direction with **no episodes,** normally episodes are defined as periods separated by two or more hours with no wind.

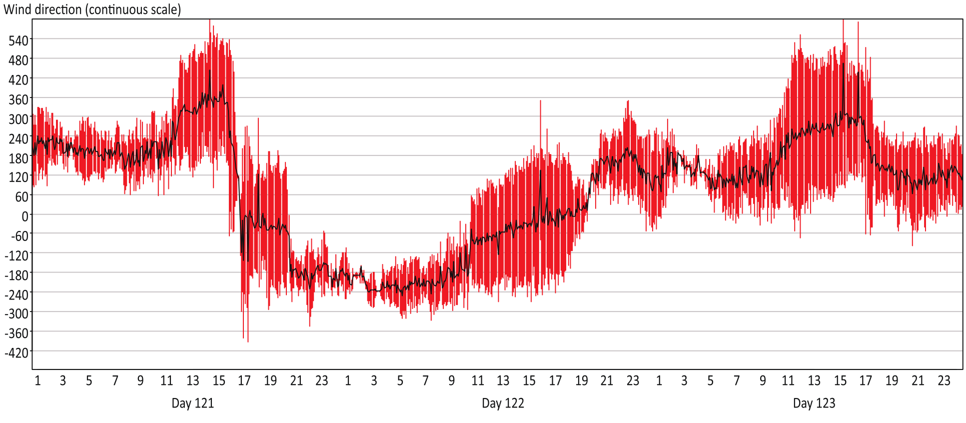


Figure 2. Wind Direction on A Continuous Scale, with no Episodes

3. MODEL FOR ANY WIND

3.1. Logistic regression model

In this study binary logistic model was applied. The outcome for ***any*** wind is a **binary** variable, coded as 1 if the wind speed in a specified hour is greater than 0 and 0 otherwise. The model is formulated as

$$\log\left(\frac{p}{1-p}\right) = factor(x_1) + factor(x_2) + \ldots + factor(x_k) \qquad (1)$$

In this formula p is the probability that there is any wind and $x_1, x_2, \ldots, x_k$ are k specified determinants each of nominal data type. (The function $\log(p)$-$\log(1-p)$ on the left side is called the **logit**.)

This is the **linear logistic model**, because it is solved to give p as the logistic function

$1 / (1+e^{-L})$, where L is a linear function of a set of indicator variables.

As a simple illustration, suppose that there is just a single determinant x_1, namely the season of the year, which could be *dry*, *wet* or *other*. In this case the linear function is $L = b_1$(if dry) + b_2(if wet) + b_3(if other), or alternatively, $L = b_0 + b_1$(if dry) + b_2(if wet). So the model has three parameters.

3.1.1. Receiver Operating Characteristic (ROC) Curve

The (adjusted) r-squared statistic is a measure of how well a model for a quantitative outcome (such as wind speed) captures the information content in the data. It ranges from 0 to 100%.

For binary outcomes, a logistic model doesn't give an r-squared, but an ROC curve assess how well the data are fitted. For a range of decision choices, it plots sensitivity (probability of finding an outcome when it is there) against the false positive error rate (probability of finding an outcome when it's not there) (Tom Fawcett, 2006).

3.2. linear regression

A linear regression model is appropriate for a quantitative outcome such as wind speed. This model takes the form

$$g(y) = factor(x_1) + factor(x_2) + \ldots + factor(x_k) \qquad (2)$$

In this formula y is the wind speed, $g(\ )$ is a transformation to satisfy statistical assumptions, and $x_1, x_2, \ldots, x_k$ are k determinants each of nominal data type.

As well as being independent, errors should be **normally distributed** and have **constant variance**. These assumptions are assessed using Q-Q plots (for normality) and by plotting residuals versus fitted values (Julian J. Faraway, 2005).

3.3. Multinomial logistic regression

In the case of the outcome can have three or more possible types such as wind directions (e.g., "North wind", "South wind", "Northeast wind" or "Southwest wind") multinomial logistic regression model was applied.

A **multinomial model** with these c = 4 outcome categories is fitted to these data. It takes the form

$$p_k = \frac{\exp(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j)}{1 + \sum_{k=1}^{c} \exp(\alpha_k + \sum_{j=1}^{m} \beta_{jk} x_j)} \qquad (3)$$

Taking 4-hour period, month and year as predictor factors, the resulting ROC curves are obtained.

3.4. Time Series Autocorrelation

Basic statistical models require that errors are independent. An ACF (Auto-Correlation Function) plot displays the correlation between errors in a model fitted to time series data (Chris Chatfield, 1996).

The correlation of hourly data are greater than 0.4 in successive hours. To make successive values approximately independent, the data was aggregated into 4-hour periods. After

re-fitting the logistic model to the aggregated data, the correlation between successive errors is reduced to less than 0.1 as shown in figure 3.

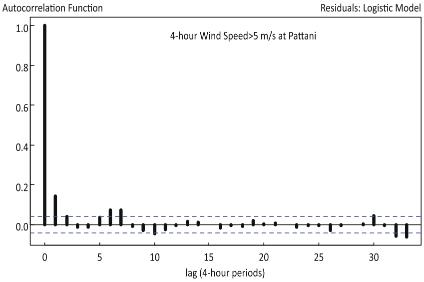


Figure 3. Time Series Autocorrelation

4. RESULTS

Following approaches described above, the result from Logistic regression model using

4-hour periods as shown in figure 4. Graphing model parameters shows that the north wind in Pattani Meteorological station occurs mainly between 8 am & 4 pm, and from May to July, with a slightly decreasing trend.

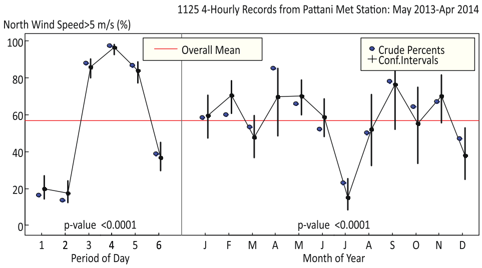


Figure 4. Wind Prevalence of 4-hourly Record from: May 2013-April 2014

The ROC curve plots shows sensitivity (probability of finding an outcome when it is there) against the false positive error rate (probability of finding an outcome when it's not there). Figure 5 shows that a model containing hour and month as factors fits the data quite well (black colour). Adding month interaction (pink colour) can improve the fit.

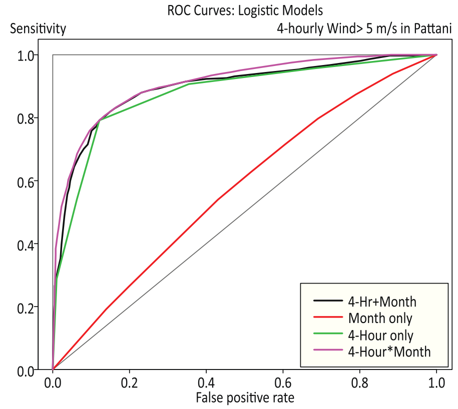


Figure 5. ROC Curve Assess Model Fitted.

A linear regression model is appropriate for a quantitative outcome such as wind speed. The errors should be normally distributed and have constant variance. After transforming with square roots, these assumptions are plausible, apart from a few high outliers, but the fit is poor, having r-squared only 28% as shown in figure 6.

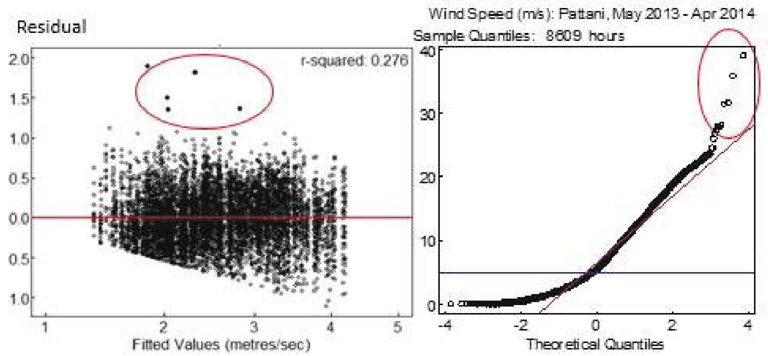


Figure 6. Residual Plot Against Fitted Values and Theoretical Quantile Plot

For wind direction detected multinomial logistic regression was applied. Wind directions was divided to be 1: Northwest wind (NW), 2: Southeast wind (SE), 3: Northeast wind (NE) and 4: Southwest wind (SW). The result from figure 7 shows Northwest and Southwest winds occur about 60% of the time, whereas Northeast winds and Southeast wind occur less than 60% of the time. Trend patterns NW winds are similar to those for NE winds and SW winds are similar to SE wind patterns.

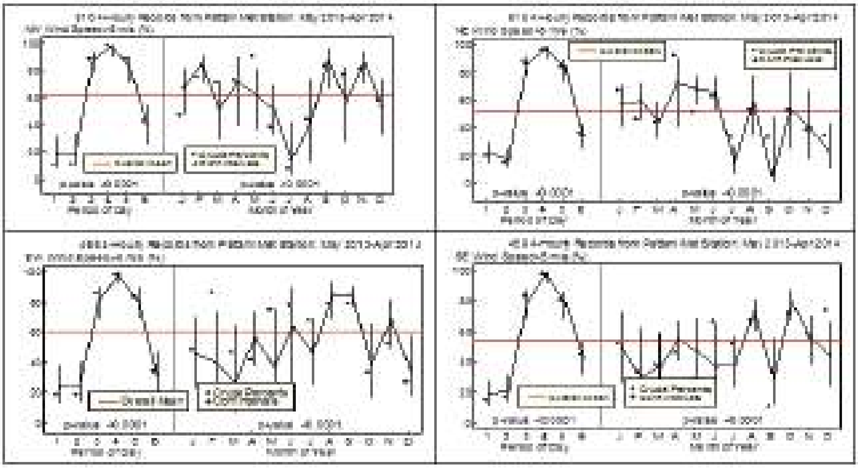


Figure 7. The Result of Wind Direction Occurs from Multinomial Model

5. CONCLUSIONS

The new methods was improved for displaying and analyzing characteristics of hourly wind velocity. These include continuous-time graphs of wind direction on an absolute scale (that is, not constrained to 0-360 degrees) using vertical bars to code speed, and methods for analysis using logistic and multinomial regression with speed-direction group as outcome and time of day, month of year factors.

Time series auto-correlations in hourly measurements are substantially reduced when outcomes were replaced by their 4-hour averages. Defining gusts as winds exceeding speeds of 5 meters/second. Methods was successfully applied to data recorded at Pattani Automatic Weather Station (AWS) in Thailand from May 2013 to April 2014.

The ROC curves was used to compare goodness-of-fit of selected models. These methods will be applied to similar data from other automatic weather stations and covering the data from meteorological stations of Thailand.

Acknowledgements

Most of all, we would like to thank emeritus professor Don McNeil for his continuous advice and encouragement.

References

Chatfield, C. (1996). The analysis of time series an introduction. Chapman & Hall.

Faraway, Julian J. (2005). Linear models with R. Chapman & Hall/CRC.

Fawcett, T. (2006). An Introduction to ROC analysis. Pattern Recognition Letters, 27: 861–874

Montgomery, Douglas C. Jennings, Cheryl L. and Kulahci M. (2008). Introduction to Time Series Analysis and Forecasting. Wiley Interscience.

Schreck S. Lundquist, J. and Shaw, W. (2008). U.S. Department of Energy workshop report: Research needs for wind resource characterization. Technical Report 500-43521, Department of Energy.

The National Energy Policy Council of Thailand. (2015). The National Energy Policy Council report: Thailand's ambitious 10-year-action plan for renewables energy. *Alternative Energy Development Plan (AEDP).*

Comparison of Temperatures between Bureau of Meteorology and Moderate Resolution Imaging Spectroradiometer

Suree Chooprateep*
Department of Statistics, Faculty of Science, Chiang Mai University
Chiang Mai, 50200 Thailand suree.choo@gmail.com
Wandee Wanishsakpong
Department of Statistics, Faculty of Science, Kasetsart University
Bangkok, 10900 Thailand w.wandee43@gmail.com

Abstract

In this study, minimum and maximum daily temperatures in Australia from Bureau of Meteorology (BOM) were compared with day and night temperatures in the Moderate Resolution Imaging Spectroradiometer (MODIS) data base. We focus on the temperatures for the period 2000-2015 of nine stations located in South-Eastern Australia. These stations in this area had high spatial correlation. Graphical displays and statistical models were used to analyse for these data. The patterns of these temperatures from MODIS and BOM were considered to compare in the average of each year, station and 8- day period.

Jel classification: **Q54**
Keywords: Australia; Climate change; Linear model

1. INTRODUCTION

The global climate change has occurred due to the environmental factor such as temperature, precipitation, biomass, solar characteristics and sea level. These factors are used as tools to predict the changes in climate and most of them are interlinked with the vegetation and temperature (Kaufmann et al., 2003; Goward et al., 2002). The variation of globally averaged surface temperature started more than 150 years ago and speeded up to 20th century (Jones et al., 1999). In general, the data have investigated the trend and pattern of temperature change from Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) data (Goward et al., 2002). The behavior of NDVI trend curve is associated with the regional scale climate change, location and also type of vegetation (Shevyrnogove et al., 2013). The LST is earth's skin surface temperature while NDVI is the most used vegetation index. NDVI and LST are commonly used Moderate Resolution Imaging Spectroradiometer (MODIS) products. MODIS is a payload scientific instrument that was launched into Earth orbit by NASA in 1999 on board the Terra (EOS AM) Satellite, and in 2002 on board the Aqua (EOS PM) satellite. Terra's orbit around the Earth is timed so that passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. Terra MODIS and Aqua MODIS are viewing the entry Earth's surface every 1 or 2 days. The instruments acquire data in 36 spectral bands ranging in wavelength from 0.4 to 14.4 µm and at varying spatial resolutions (2 bands at 250 m, 5 bands at 500 m and 29 bands at 1 km). These data related to global dynamics and processes occurring on the land, in the oceans, and in the lower atmosphere. In addition, there is a data source to use for investigating temperature change. Wanishsakpong and McNeil (2015) studied modelling of daily maximum temperatures over Australia using data from the Australian Bureau of Meteorology (BOM) for 700 weather stations. These stations have a difference of observation due to weather phenomena and changing requirement (BOM, 2015). The sample of 85 stations was analyzed using factor analysis to account for spatial correlation. In the South-Eastern Australia, the data have high correlation in the first factor. This region has experienced drought since 1973 with increased temperature during 2002 and 2003 (Nicholls, 2004). Moreover, the daily maximum temperature reaches to 10oC above average (Richman and Leslie, 2014). Therefore, it would be interesting to study 9 stations in South-Eastern region: Tumbarumba, Hay, Warracknabeal, Kerang, Point Hicks, East Sale, Mangalore, Hobart, and Flinders Island (station number: 72043, 75031, 78077, 80023, 84070, 85072, 88109, 94029 and 99005 respectively). The purpose of our study was to compare these minimum and maximum daily temperatures in the BOM database with Land Surface (day and night) Temperatures in the MODIS database for the period 2000-2015.

2. DATA MANAGEMENT AND STUDY AREA

In this study, average daily temperatures in Australia for a 16 year period (2000-2015) on the earth's surface were obtained from two databases. The first database is Australian Bureau of Meteorology (BOM) website at http://www.bom.gov.au/data. The data in BOM provided daily maximum and minimum temperatures in degree Celsius. The second database is Moderate Resolution Imaging Spectroradiometer (MODIS) from satellites, the land surface temperature data are retrieved at 1 kilometer pixels by capturing the earth's picture at the same place. The color indexes in these pixels were converted into temperatures. This source provided for every eight daytime and nighttime temperatures in degree Kelvin, the data were downloaded from the website of land subsets (http:// daacmodis.ornl. gov/cgi-bin/MODIS/GLBVIZ_1_Glb/modis_subset_order_global _col5.pl).

The temperatures were converted from degree Kevil to Celsius scale by subtracting 273.15. Therefore, this research focused on four sets of temperatures with two sets from BOM and two sets from MODIS, each set composed of 8-day period data in each station. Temperatures of 9 stations which are located on the specific longitude/latitude in the South-Eastern Australia were studied. The black dots represented 9 stations as shown in Figure 1.

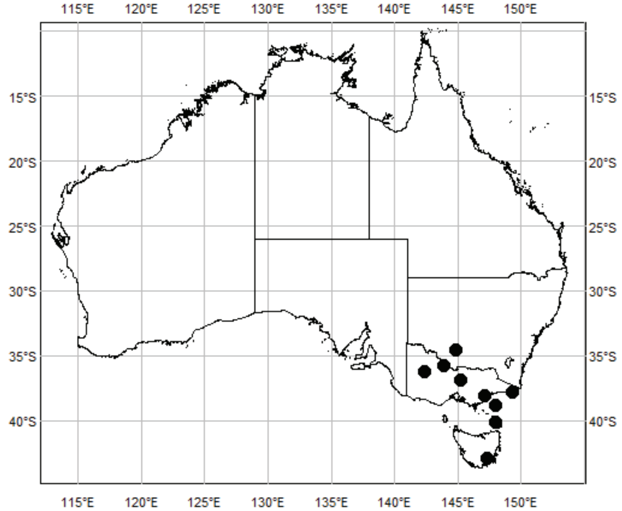


Figure 1. The 9 Locations in the Study Area

Statistical model

For each station, the data of 16 year period consist of 715 observations; 2 sets of day and night data are from MODIS and 2 sets of maximum and minimum data are from BOM. Temperatures in each area can be explained by year, day and station. The statistical model for each set of observations may

be formulated as

temperature ~ factor (year) + factor (day) + factor (station)

3. RESULTS

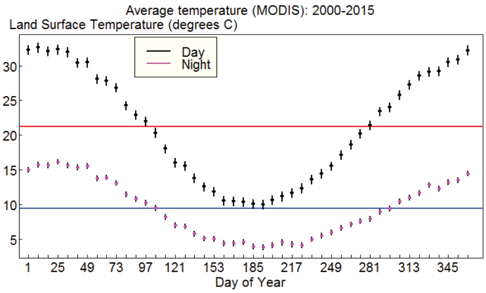


Figure 2. Land Surface Temperature of each 8-day Period

From MODIS data, the overall means of day and night temperatures were 21.2 and 9.4 degree Celsius. The ranges of day and night temperature are from 10.01 to 32.61 and 3.91 to 16.18 degree Celsius respectively. Figure 2 showed temperature means with their confidence intervals of day and night at the upper and lower curve respectively. The average temperatures on each 8-day period during 16 years or 46 periods each year were considered, both sine curves of day and night temperatures were according in the same direction. Figure 2 showed the average temperatures are decreased from February (day 33 of year) until to May (day 161 of year) and lowest in June (day 193 of year). The average temperatures are increased from July (day 201 of year) until to November (day 337 of year) and highest in December (day 361 of year).

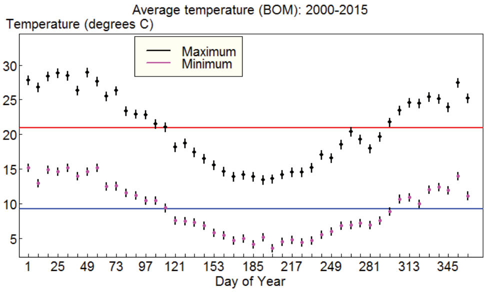


Figure 3. Maximum and Minimum Temperature of each 8-day Period

From BOM data, the overall means of maximum and minimum temperatures were 21.0 and 9.3 degree Celsius. The ranges of maximum and minimum are from 13.44 to 28.91 and 3.57 to 15.23 degree Celsius respectively. Figure 3 displayed temperature means with their confidence intervals of maximum and minimum at the upper and lower pattern respectively.

Comparison of figure 3 with figure 2, the overall means of day and night from MODIS are not different patterns of maximum and minimum from BOM respectively. The temperature of 8-day period from MODIS is higher than 30 degree Celsius started from the middle of December to the middle of February, day 353 to day 365 of year and also day 1 to day 43 of year but there is no day from BOM higher than this level (Figure 3). There are 11 values of the 8-day period, day 145 to day

225 of year, from MODIS are lower than these values from BOM.

The night temperature from MODIS is not much different minimum from BOM, only temperature at day 9 and day 361 from MODIS are much higher than minimum temperature from BOM. For the 8- day period, data from MODIS showed less variation than data from BOM. Furthermore, year and location effects were explained over this period. Figures 4 and 5 showed trends of temperature from MODIS and BOM during 16 years: 2000-2015 and display the average temperature of each station in the same figure.

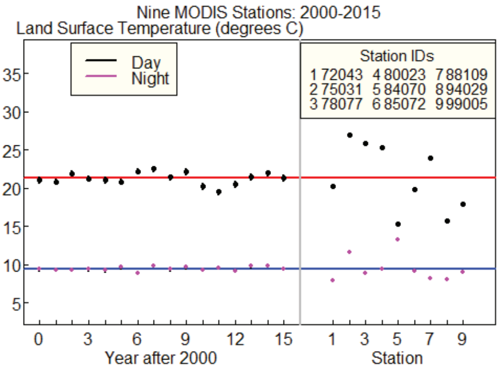


Figure 4. Land Surface Temperatures of Year and Station from MODIS

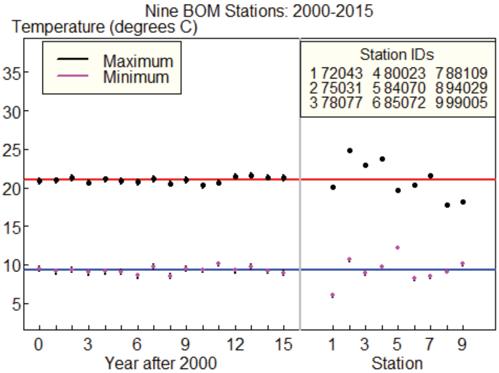


Figure 5. Maximum and Minimum Temperatures of Year and Station from BOM

Comparison temperatures among years from MODIS and BOM show on the left panel in figure 4 and 5 respectively, the average of day and night of land surface temperature from MODIS (figure 4) and also the average of maximum and minimum temperature from BOM (figure 5) during 16 years are not much different from their overall means. Day and maximum temperature curves are similar and also night and minimum curves are similar. To consider among locations (right panels in figure 4 and 5), day temperature from MODIS (figure 4) are higher variation than maximum temperature from BOM (figure 5). There are 4 stations (number 2, 3, 4 and 7) were higher than the overall day-mean and 5 stations (number 1, 5, 6, 8 and 9) were lower than the overall day- mean.

For 9 stations, the average of day temperature from MODIS correlated to the average of maximum temperature from BOM shown as small boxes in figure 6, the width of these boxes represent the means with their confidence intervals. The average of day temperature of four stations (ID: 75031,

78077, 80023 and 88109) were higher than the average of maximum temperature and there were three

stations (ID: 85072, 94029 and 84070) lower than the average of maximum temperature, where else were not different. Also, the average of night temperature from MODIS correlated to the average of minimum temperature from BOM shown as small boxes in figure 7. The average of night temperature of four stations (ID: 84070, 75031, 85072 and 72043) were higher than the average of minimum temperature and there were four stations (ID: 800223, 99005, 88109 and 94029) lower than the average of minimum temperature, where else were not different.

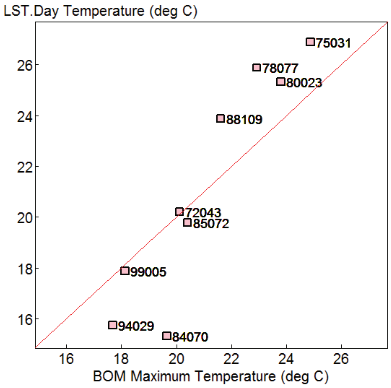


Figure 6. Day Temperature from MODIS versus Maximum Temperature from BOM

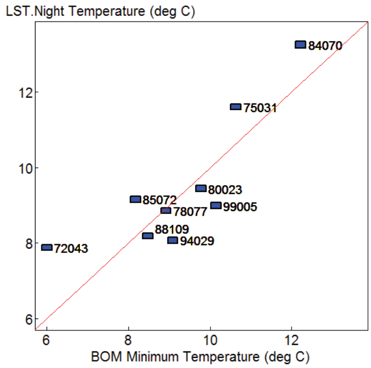


Figure 7. Night Temperature from MODIS versus Minimum Temperature from BOM

4. CONCLUSIONS

The patterns of the temperature of day and night from MODIS were similar to the patterns of temperature of maximum and minimum from BOM database. Comparison of the 8-day period from MODIS with BOM, the ranges of day and night temperature are bigger than the ranges of maximum and minimum temperature. Comparison between day and maximum temperatures for all stations, the Point Hicks station was found most different (maximum temperature higher than day temperature 4.33 degree C). In addition, comparison between night and minimum temperatures, the Tumbarumba station was found most different (night temperature higher than minimum temperature 1.88 degree C). However, there were a few differences between day and maximum temperatures and also between night and minimum temperatures in the other stations.

Acknowledgements

We are grateful to Emeritus Prof. Don McNeil for providing the valuable advice and support

References

BOM.(2015). Bureau of Meteorology, Improving Australia's climate record.http://www/bom.gov.au. Accessed 1 April 2015.

Goward, S. N., Xue, Y., &Czajkowski, K.P. (2002). Evaluating land surface moisture conditions from the remotely sensed temperature vegetation index measurements: An exploration with the simplified simple biosphere modek. Remote Sensing Environment.

Jones, P. D., New, M., Parker, D. E., Martin, S., &Rijor, I. G. (1999). Surface air temperature and its changes over the past 150 years.Reviews of Geophysics.

Kaufmann, R. K., Zhou, L., Myneni, R. B., Jucker, C.J., Slayback, D., Shabanov, N. V., &Pinzon, J. (2003). The effect of vegetation on surface temperarute : A statisticalanalysis of NDVI and climate data. GeophysicalResearchLetters.

Nicholls, N. (2004). The changing nature of Australian droughts.Climate Change.

Richman, M. B. Leslie, L. M. (2014). Attribution and prediction of maximum temperature extremes in

SE Australia.Procedia Computer Science.

Shevyrnogov, A. P., Chernetskiy, M. Y.,&Vysotskaya, G. S. (2003). Multiyear trends of normalized difference vegetation index and temperature in the South of KransnoyrskKrai.Atmospheric and Oceanic physics.

Wanishsakpong, W. &McNeil, N. (2015). Modelling of daily maximum temperatures over Australia from 1970 to 2012. Meteorological Application.

A Study of Temperature Changes and Patterns in Australia Based on Cluster Analysis

Wandee Wanishsakpong*
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand.
– w.wandee43@gmail.com
Khairil Anwar Notodiputro
Department of Statistics, Faculty of Mathematics and Science, Bogor Agricultural University,
Bogor 16668, Indonesia.– khairilnotodiputro@gmail.com

Abstract

The climate and geographical characteristics of Australia are different from one region to other regions. Signals are often unstable and varied in time and space. In this paper we discuss the temperature changes and patterns in Australia using the recorded data of 88 stations across the country during the period of 1990 to 2015. Missing value problems were existed in several stations so imputation techniques were unavoidable. Usually the temperature data consist of both seasonal and trend components so the data need to be seasonally adjusted before cluster analysis techniques were employed. The temperature changes and patterns are then discussed for each cluster.

JEL classification: **Q54**
Keywords: Cluster analysis, Imputation, Regression model.

1. INTRODUCTION

Studies on weather extremes have attracted many researchers in the area of climate change. Inferences on weather extremes can be based on all types of climatologically information (Visser, H., &Petersen, A. C., 2012). During the 20th century, increases in extreme temperatures are associated with variations of maximum temperatures in some parts of Australia (Nicholls et al., 2004).

Several studies have investigated temperature trends and patterns using different statistical methods. The methods include linear regression (Griffiths *et al.*, 2005), multivariate linear regression and spline linear models (McNeil, N. &Chooprateep, S., 2014) and polynomial regression models (Wanishsakpong, W. &McNeil, N., 2015). Wanishsakpong *et al.* (2016) fitted polynomials of order 3 on the maximum temperature data of Australia with high R-square.

In this paper we discuss the temperature changes and patterns in Australia using the recorded data of 88 stations across the country during the period of 1990 to 2016. Missing value problems were existed in several stations so imputation techniques were unavoidable. Many statistical techniques relevant for imputing missing data as well as clustering stations such as principle component analysis (Jones, 1999), factor analysis (Wanishsakpong &McNeil, 2015; Wanishsakpong *et al.*, 2016) and cluster analysis (Mahlstein & Knutti, 2009). Before cluster analysis techniques were employed, the data was seasonally adjusted. The temperature changes and patterns are then discussed for each cluster.

2. DATA AND METHOD

Data

Mean maximum monthly temperatures were obtained from the website http://www.bom.gov.au of the Australia Bureau of Meteorology (BOM) for over 700 weather stations. These stations have different having incomplete records due to the weather phenomena and changing requirements (BOM, 2015). An area-based sample of 112 stations was selected in accordance to the Australia Climate Observations Reference Network Surface Air Temperature (ARORN-SAT) as shown in Figure 1.

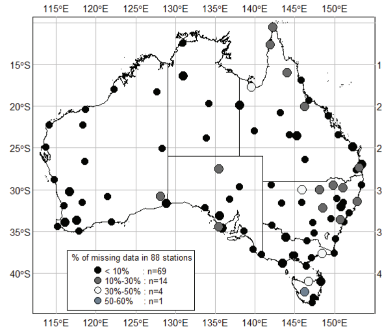


Figure 1. Positions of 88 Stations Included in this Study with Indication of Percentage Missing Data for each Station

Methods

The first step is to impute missing values. The missing values were found in the 88 observed stations (See Figure 1). A random sample of 28 stations was chosen and a linear regression model was fitted:

$$y_{kt} = \tau_0 + \tau_1 y_{nt} + \tau_2 y_{st} + \tau_3 y_{k(t-1)} + \tau_4 y_{k(t+1)} + \tau_5 m + \in_{tk}$$
$$(t = 1,2,\ldots,312, \text{ and } = 1,2,\ldots,28\ )$$

where y_{kt} denotes the mean maximum temperature at month t of station k;

y_{nt} is the mean maximum temperature at month t for the nearest station from station k;

y_{st} is the mean maximum temperature at month t for the second nearest station from station k;

$y_{k(t-1)}$ is the mean maximum temperature at month $(t - 1)$for station k;

$y_{k(t+1)}$ is the mean maximum temperature at month $(t + 1)$for station k;

$\tau_0, \tau_1, \tau_2, \tau_3$ and τ_4 are the corresponding regression parameters of the models;

m is a factor of month and $\in_{tk}$ are the error terms.

After the missing values have been imputed, the complete series of data were then deseasonalized as described in Montgomery *et al.* (2008).

The second step is to form clusters of stations. The clustering technique starts with the individual stations. The most similar stations are first grouped, and these initial groups are merged according to the temperature similarities (Johnson, 1998).

All data analysis and graphical displays were carried out using the R language R Development Core Team

3. RESULTS

The mean maximum monthly temperature data for a random sample of 28 stations was fitted with the linear regression model and the resulting equation is as follows:

$$\hat{y}_{kt} = \widehat{\beta}_{0m} + 0.0844 y_{nt} + 0.04263 y_{st} + 0.44965 y_{k(t-1)} + 0.45151 y_{k(t+1)}$$

where $\widehat{\beta}_{0m} = -0.116, -0.570, -0.390, -0.431, -0.989, -1.6635, -1.4257,$
$-1.1475, -0.4373, -0.565, -0.517, -0.809$ for $m = 1,2,\ldots,12$, respectively.

Note that all coefficients of the regression were significant (P-value < 0.0.5) and adjusted $R^2 = 0.96$.

After the missing values were imputed and the data were seasonally adjusted then we form cluster analysis of stations using the K-means method with complete linkage. By setting the similarity level of 80% we obtain six clusters described in Table 1.

Table1. Description of Resulting Clusters

Cluster	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance
1	24	15614	24.4477	40.9188
2	15	9084.5	24.2255	31.9906
3	19	13898	26.5428	36.0285
4	12	11617.3	29.067	60.2
5	16	13934.9	28.2018	42.988
6	2	577	16.9849	16.9849

4. CONCLUSIONS

In this study, mean maximum monthly temperature from 1990 to 2015 were analyzed. A linear regression model was used to estimate missing value in all 88 stations by using the information of neighboring stations and previous as well as following year temperatures of the same month. Since the data is time series data which have seasonal component. Therefore, the data was deseasonalized to remove seasonal variation. Cluster analysis was used to group 88 stations into 6 clusters corresponding to the part of geographical regions. The mean maximum monthly temperature ranged from 18°C to 36°C. The highest average temperature is about 36°C in desert region which locate at the northern and northwestern part of the country. The patterns of temperature change still have periodic patterns and trends of temperature were increasing in all clusters.

References

BOM. (2015). Bureau of Meteorology, Improving Australia's climate record. http://www/bom.gov.au. Accessed 1 April 2015.

Griffiths, G. M., Chambers, L. E. Haylock, M. R. Manton, M. J. Nicholls, N. &Baek, H. J. (2015). Change in mean temperature as a predictor of extreme temperature change in the Asia-Pacific region. International Journal of Climatology.

Johnson, D. E. (1998). Applied Multivariate Methods for Data Analysis. Duxbury Press: Pacific Grove, CA.

Mahlstein, I., &Knutti, R. (2009). Regional climate change patterns identified by cluster analysis. Climate Dynamics.

McNeil, N., &Chooprateep, S. (2014). Modeling sea surface temperatures of the North Atlantic Ocean. Theoretical and Applied Climatology.

Montgomery, C. M., Jennings, C. L., &Kulahci, M. (2008). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.

Nicholls, N., Della-Marta, D., &Collins, D. (2004). 20th century changes in temperature and rainfall in New South Wales. Australian Meteorological Magazine.

Visser, H., &Petersen, A. C. (2012). Inferences on weather extremes and weather-relates disaster: a review of statistical methods. Climate of the past.

Wanishsakpong, W., &McNeil, N. (2015). Modelling of daily maximum temperatures over Australia from 1970 to 2012. Meteorological Applications.

Wanishsakpong, W., Notodiputro, K. A., &McNeil, N. (2016). Clustering and Forecasting Maximum Temperature of Australia. Chiang Mai Journal of Science.

Modeling of Temperature Patterns in Kathmandu Valley of Nepal from 2000 to 2016

Ira Sharma*
Prince of Songkla University, Pattani campus, Thailand. Email: irasg123@gmail.com
Phattrawan Tongkumchum
Prince of Songkla University, Pattani campus, Thailand. Email: phattrawan.t@psu.ac.th
Attachai Ueranantasun
Prince of Songkla University, Pattani campus, Thailand. Email: attachai.u@psu.ac.th

Abstract

The aim of this study was to determine a pattern of Land Surface Temperature (LST) in Kathmandu, Nepal from 2000 to 2016 using appropriate statistical methods. The data, recorded every eight day with 1 km2 spatial resolutions, were obtained from Moderate Resolutions Imaging Spectro-radiometer. The study area had 9 regions of 21×21 km2 area, each of which was further divided into 9 sub regions. The data for all sub regions were separately used for time series pattern analysis. First, the data were seasonally adjusted and auto correlation was detected by autoregressive method. Then, the data were filtered to eliminate correlation effects. Finally, polynomial regression model of second degree was applied to find the temperature patterns. The results showed that 27% of the area were found to have the temperature patterns of steep rising, while the patterns in another 27% had increasing then decreasing shape. Moreover, 25% of the area had decreasing then increasing pattern and the remaining 21% of the area showed no change appreciably. This method can be applied in other climatic factors that have influence on ecosystem.

Keywords: LST, Auto-regression method, Polynomial regression model
JEL classification: **Q54**

1. INTRODUCTION

The changes in climate parameters due to global warming have been creating negative impacts on human society and the natural environment. The local, regional or global warming of air is connected to the destruction of ecosystem. It has impacts on different spheres of public life as well. In Asia, hot days and hot nights are predictedly increasing as compared to the cold days and the cold nights in the period from 2011 to 2099 (e.g. Mahmood and Babel, 2014). The corn yield, in USA, increases as temperature rises but only up to 29°C and after that, the production decreases (e.g. Schlenker and Roberts, 2009). A similar result was seen for rice and wheat production in Nepal (e.g. Malla, 2008). Climate change can be particularly hard-hitting for the developing countries, including Nepal, which rely on natural resources for the economy and livelihoods.

A difficult task is to detect the pattern of temperature change in an area using appropriate statistical method. Climate scientists have found temperature change by using different methods such as observation and computer simulation modeling (e.g. Johannessen *et al*., 2003), annual average method (Jones *et al*., 1999), empirical orthogonal functions (e.g. Semenov, 2007), factor analysis (e.g. Chooprateep and McNeil, 2015) and Pearson correlation analysis (e.g. Griffiths *et al.*, 2005). In addition, majority of the studies rely on the linear regression model to find the variation of temperature in different parts of the world (e.g. Lean and Rind, 2009; Chooprateep and McNeil, 2016; Wanishsakpong and McNeil, 2016; Hughes *et al*., 2006). However, the linear regression can show only the change between the beginning and end of study period. Therefore, in this study, the temperature patterns were obtained by using polynomial regression model after adjusting the seasonal effects and auto correlation of the data to show the pattern of change varying within the study period. Additionally, the trend of temperature is analysed using linear regression model to see the net temperature change in 15 years period.

2. METHODOLOGY

The study was carried out around Kathmandu valley of Nepal in an area of 3969 km^2. The area extends to a total of 8 districts: Kathmandu, Bhaktapur, Lalitpur, Sindhupalchok, Ramechhap, Makwanpur, Nuwakot and Dhading. The first three districts make up the Kathmandu valley, and other five surround the valley. The average maximum temperature of the Valley is more than 30˚C in summer and the minimum is less than 1˚C in winter. There is a heavy monsoon period in the mid of the year.

The data

Land Surface Temperature (LST) is a remote sensing data from Moderate Imaging Spectro-radiometer (MODIS) sensor fitted in Terra satellite of National Aeronautics and Space Administration (NASA). It can monitor the various environmental changes like temperature, rainfall, vegetation, draught, fire, flood (see NASA, 2015). First, the LST data of 1 km^2 pixel resolutions with an area of 21×21 km^2 for 9 different regions were ordered for a period of March 2000 to June 2016. The data can be retrieved from MODIS website (see ORNL DAAC, 2016). Nine different regions were selected at coordinates 27.761E/ 85.206N,27.761E/ 85.394N,27.761E/ 85.582N, 27.595E/ 85.207N, 27.595E/ 85.394N, 27.595E/ 85.581N, 27.429E/ 85.207N, 27.429E/ 85.394N and 27.429E/ 85.581N. Figure 1 shows that eight regions surrounding the centrally located region, the number 5.

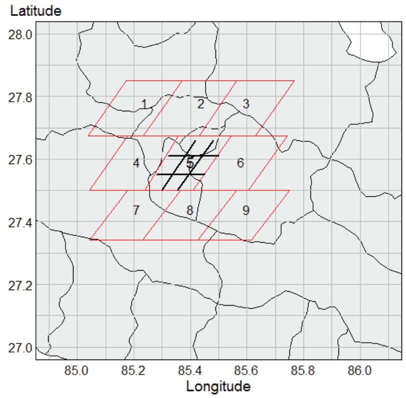


Figure 1 Study Area Around the Valley, Showing 9 Regions and the Sub Regions

The coordinates for making polygons around each region were obtained from Modland Tile Calculator by forward and backward mapping procedure and the tool is available freely at its website (see Modland tile calculator, 2016). For each region, a time series of 742 successive observations at 8 days time intervals were available. The data were truncated to 690 by eliminating some from the beginning and the end period of the data set to adjust it for just the 15 years time frame. Then the data of each region were aggregated within a smaller sub regions, approximately 7 x 7 km^2 in area as shown for region 5 in Figure 2. These sub regions were named as North-East (NE), North (N), North-West (NW), West (W), Central (C), East (E), South-West (SW), South (S) and South-East (SE). There are total 81 sub regions in study area.

The coordinates for the plots of slopes within 81 sub regions were obtained from the same Modland Calculator. In each region, eight co-ordinates surrounded the central one in a pattern of 3 x 3 matrix. The process was repeated for all 9 regions.

Methods for analysis

The statistical methods and plots are created for each of the 81 sub regions separately. First of all, the data are seasonally adjusted with the use of spline function, that helps to stabilize the mean of the data at each sub region. The autocorrelation of these data is detected by an autoregressive process of the first order lag (AR1). The model takes the form,

$$y_t = \phi_1 y_{(t-1)} + z_t \tag{3}$$

Here, y_t is the seasonal adjusted temperatures at observation t, ϕ_1 is the constant and $y_{(t-1)}$ is the first order lag phase $(t-1)$ of seasonally adjusted data, z_t the value not explained by the past values.

Finally, the correlation free data were obtained by moving auto correlation component by convolution method of filtration, from equation (3), and the form of equation is,

$$z_t = y_t - \phi_1 y_{(t-1)} \tag{4}$$

Now, z_t is the filtered temperature at observation t, and all other components are explained as above. Furthermore, simple linear regression model was fitted to the filtered data. The model takes the form,

$$z_t = \beta_0 + \beta_1 y_t + \varepsilon \tag{5}$$

Here, z_t is the filtered data, β_0 is the intercept and β_1 is the coefficient, while y_t is seasonal adjusted temperature and ε is the error term. Finally, the predicted temperature from this linear model was fitted to polynomial regression model of second degree (quadratic model). The form of model is,

$$x = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e \tag{6}$$

Here, x is the fitted temperature derived from the linear model, α_0 is the intercept and α_1, α_2 are the coefficients, t is the observation day of each sub region and e is the error term.

All data analysis and graphical displays were carried out using R Statistical Programming (see R Core Team 2015).

3. RESULTS

The data, seasonally adjusted and auto correlation filtered, were fitted with linear regression model. Figure 2 shows a total 10 panels for Region 5. Nine of them show the linear trends of the temperature at nine sub-regions. The plots for all the regions indicated that, 46% of the sub region had decreasing trend, 53% had increasing and one sub region showed no temperature change during the study period. Moreover, none of them had significant p-value. During 15 years period, the net linear rise of vegetation ranged from 0.009 to 0.430 and the fall from -1.047 to -0.010, along with overall auto correlations (a1) below 0.30, in the study area.

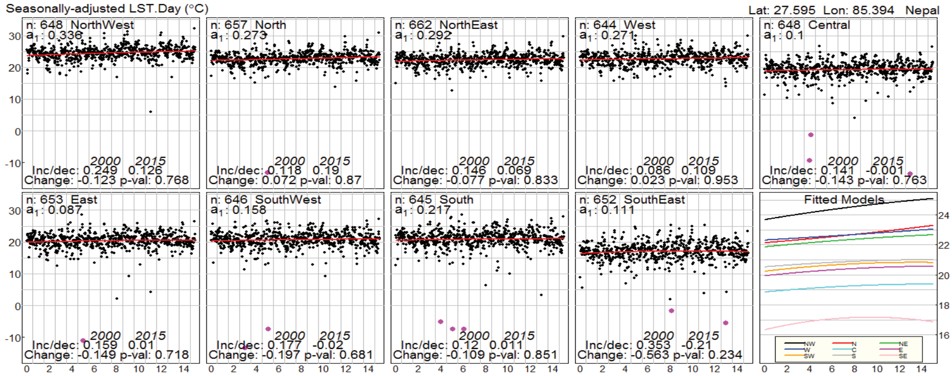


Figure 2. Linear Trends of Nine Sub Regions of Region 5 and the Respective Quadratic Slopes

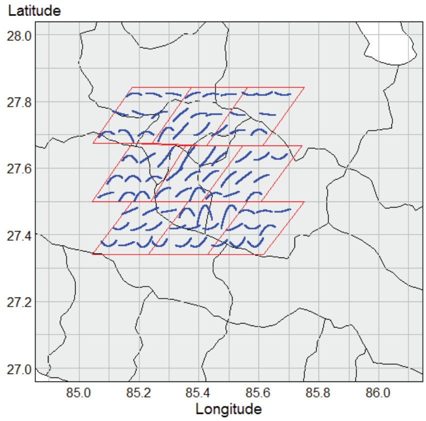


Figure 3. Temperature Patterns in Nine Regions

The fitted values were obtained from liner model and the polynomial model was fitted to them. The quadratic slopes of the temperature were obtained for each sub region. The last panel in Figure 2 shows the quadratic slopes of all those nine sub regions, of region 5, in a single panel. These slopes, plotted separately for 81 sub regions in Figure 3, showed various patterns which were categorized into 4 different groups as per their shapes- 'steep rising' (27%), 'increasing then decreasing' (27%), 'decreasing then increasing' (25%) and finally the 'no change' (21%) group.

4. DISCUSSIONS

This study analyses the linear trend of vegetation in 81 sub regions and that does not show statistically significant results. Therefore the quadratic slopes were used, that could serve the purpose to show the patterned changes of temperature. The study detects four different patterns, steep rising, increase then decrease, decrease then increasing and the final group having no change in 15 years period. The linear model has been used by most of the previous research works to find temperature change (e.g. Hughes *et al.*, 2006; Griffiths *et al.*, 2005; Choprateep and McNeil, 2015) in a specified period of time. They have shown, just the net change from initial to end time of study period. No pattern could be seen in the period between the two time points. It can be explained that the net temperature change between beginning and end of the study time cannot describe temperature patterns on how it has been changing during the specified period.

This study has applied a combination of linear regression followed by polynomial regression models, which is an effective technique for modeling the temperature. This method also adjusts for spatial correlation and seasonal variations of the data. Therefore, the results are more accurate as compared to the other previous methods.

5. CONCLUSIONS

The combination of autoregressive process, linear regression and polynomial regression model can be successfully applied to find time series patterns and the trends of temperature. The time series patterns show that there are a lot of variations of temperature change, even in adjacent sub regions, during 15 years period. The methodological approach used in this study can be applied to similar studies at local, regional and global scale. Also, it is applicable to model the other factors, for example vegetation, precipitation and wind in future and the method can be applicable to find the analysis of temperature pattern of any other areas, both regionally and globally.

6. ACKNOWLEDGMENTS

This work was supported by the Higher Education Research Promotion and the Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission. We would acknowledge the Department of Mathematics and Computer Science of Prince of Songkla University for providing the platform for this study. We are grateful to the Professor Don McNeil for his immense guidance during the study.

References

Chooprateep, S. and McNeil, N. (2015). Surface air temperature changes from 1909 to 2008 in Southeast Asia assessed by factor analysis. *Theoretical and Applied Climatology*. Vol. 123 (1). Pp: 361-368.

Griffiths, G.M., Chambers, L.E., Haylock, M.R., Manton, M.J., Nicholls, N., Baek, H.J., Choi, Y., Della-Marta, P.M., Gosai, A., Iga, N., Lata, R., Laurent, V., Maitrepierre, L., Maamigawa, H., Ouprasitwong, N., Solofa, D., Tahani, L., Thuy, D.T., Tibig, L., Trewin, B., Vediapan, K. and Zhai, P. (2005). Change in mean temperature as a predictor of extreme temperature change in the Asia-Pacific region. *J. Climatol*. Vol. 25. Pp: 1301-1330

Hughes, G.L., Rao S.S., and Rao, T.S. (2006). Statistical analysis and time-series models for minimum/ maximum temperatures in the Antarctic Peninsula. *The Royal Society*. Vol. 463. Pp: 241-259. DOI: 10/1098/rspa.2006.1766.

Johannessen, O.M., Bengtsson, L., Miles, M.W., Kuzmina, S.I., Semenov, V.A., Alekseev, G.V., et al. (2004). Arctic climate change: observe and modeled temperature and sea-ice variability. *Tellus*. Vol. 56. Pp:328-341

Jones, P.D., New, M., Parker, D.E., Martin, S. and Rijor, I.G. (1999). Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics*. Vol. 37. Pp: 173-199.

Lean, J.L. and Rind, D.H. (2009). How will Earth's surface temperature change in future decades?. *Geophysical Research Letters*. Vol. 36. L5708. DOI: 10.1029/ 2009 GL o38932.

Mahmood, R. and Babel, S.M. (2014). Future changes in extreme temperature events using the statistical downscaling model (SDSM) in the trans-boundary region of the Jhelum river basin. *Weather and Climate Extremes*. Vol.5-6(1), Pp: 56-66

Malla, G. (2008). Climate change and its impact on nepalese agriculture. *The Journal of Agriculture and Environment*. Vol.9, Jun.2008.Pp: 62-71

Modland tile calculator. (2016). Available from http://landweb.nascom.nasa.gov/cgi-bin/developer/ tilemap .cgi. Accessed on 5 September 2016.

NASA. (2015). Global Climate Change Vital Signs of the Planet. Available from: http://climate. nasa.gov/effects. Accessed on 5 December 31, 2015.

ORNL DAAC. (2016). MODIS subset of NASA Earth Data. Available from http://daacmodis. ornl.gov/cgi-bin/MODIS/GLBVIZ_1_Glb/modis_subset_order_ global_col5.pl. Accessed 5 October 2016.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of National Academy of Sciences of United States of America (PNAS)*. vol. 106 no. 37. Pp.:15594–15598, doi: 10.1073/pnas.0906865106

Semenov, V.A. (2007). Structure of temperature variability in the high latitudes of northern hemisphere. *Atmos. Oceanic Phys*. Vol. 43. Pp: 744-753.

Wanishsakpong, W. and McNeil, N. (2016). Modelling of daily maximum temperature over Australia from 1970-2012. *Meteorological applications*. Vol. 23(1). Pp: 115-122.

Combined Wavelet Fuzzy Logic (WFL) to Predict Drought Events in Indonesia Using Reanalysis Dataset

Heri Kuwanto*
Department of Statistics-Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – heri_k@statistika.its.ac.id
Dinni Ari Rizky Taufanie
Department of Statistics-Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – dinnitaufanie@gmail.com
Dedy Dwi Prasetyo
Department of Statistics-Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia – dedy_dp@statistika.its.ac.id

Abstract

Drought is characerized by lack of precipitation level, which may lead to some significant effect in many vital activities. Indonesia has been listed as one of the countries with relatively high risk of drought event. The Standardize Precipitation Index (SPI) derived from the observed rainfall intensity is one of the tools to rate the drought level. However, collecting reliable observation is a great challenge in Indonesia. To deal with this issue, this research uses reanalysis precipitation data at the 1000hPa atmospheric level. The Wavelet Fuzzy Logic (WFL) will be applied to develop the predition rules or models. The WFL is able to reduce the bias and minimized human error and consistency. Through WFL, this research develop Indonesia's drought vulnerability maps in 2016. Direct comparison with SPI based mapping has been investigated as well. The results indicates also that reanalysis data is a good proxy of observed rainfall.

Keywords: Drought; Fuzzy; Wavelet ; Reanalysis.

1. INTRODUCTION

Drought is characterized with low precipitation rate than normal condition. If drought happens. Drought may lead to a disaster if it happens for long duration, which thus influences the economic and social live. The intensity, duration and space distribution of drought event will vary [1].

Indonesia is one of countries which is vulnerable to drought events. Within the period of 1844-1970, there were 16 events, and this number increases during periods of 2001-2005 to 37 events. Furthermore, statistic reported that there were 58 forest and land fire. The intensity tends to increase overtime [2]. This phenomena gave significant impact to many sectors, especially rain dependent activities.

One of the important components of national strategy dealing with drought is comprehensive drought monitoring system which can provide an alert toward the beginning and the end of drought events, the severity level as a vital information in many sectors such as agriculture, clean water, energy and health. The severity of drought level in a certain location can be detected through drough analysis index. One of the analyses is by performing the Standardized Precipitation Index (SPI) and one of the methods that can be used to predict drought events is Wavelet Fuzzy Logic (WFL).

Predicting drought in Indonesia is difficult to be conducted due to lack of reliable and complete dataset especially observation data. To overcome this problem, one of the strategies is to utilize satelite based data, which are available in grid points bases with certain level of resolution.

2. LITERATURE REVIEW

Standardized Precipitation Index (SPI)

In 1993 McKee developed a method to calculate drought index using SPI [5] with the purpose to monitor drought events. Analysis of meteorological drought using SPI can be done on monthly, three month, six month and twelve month basis. Table 1 listed the index and its corresponding weather condition.

Tabel 1. SPI and Its Corresponding Class

Drought Index	Weather condition
SPI ≥ 2,326	Exceptionally wet
1,645<SPI≤2,326	Extremely wet
1,282<SPI≤1,645	Considerably increased moisture
0,935<SPI≤1,282	Moderately increased moisture
0,524<SPI≤0,935	Slightly increased moisture
-0,524<SPI<0,524	Near normal
-0,935<SPI≤-0,524	Minor drought
-1,282<SPI≤-0,935	Moderate drought
-1,645<SPI≤-1,282	Severe drought
-2,326<SPI≤-0,1645	Extreme drought
SPI≤ -2,326	Exceptional drought

(Source : National Drought Mitigation Center, 2010)

Wavelet

The first step of doing wavelet is local analysis, which will be able to show the signal information which does not owned by other signal analyses such as tendency, cut off point, and similarity. Due to its ability to see data from many views, wavelet is able to simplify and reduce noise without reducing the quality of the data. Wavelet transform can be divided into two categories i.e. continuous wavelet transform which is useful for time series data defined on all real values, and discrete wavelet transform for integer time series data [6]

The Continuous Wavelet Transformation calculates the signal convolution by modular window every time with the specified scale. This transformation uses scale and translation, where the scale is associated with frequency while translation is associated with location of modular window on the signal range, associated with time. Mathematically CWT can be defined as:

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi * \left[\frac{(n' - n)\delta t}{s}\right], \tag{1}$$

with :

$W_n(s)$ is wavelet transform, s is scale, δt is *time spacing*, n = 1,2,...,N. n' = 0,1,...,N-1. x is time series data before transformation, ψ^* is *wavelet*, with * shows complex conjugate, and N is number of data.

The basic function of wavelet can be designed based on the need in order to obtain best transformation results, which also becomes main difference with Fourier transform which uses only sine function as the modular window. Wavelet transform uses mother wavelet which fits the data in the decomposition process. This research uses Marlet mother wavelet as follow:

$$\psi(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2}, \tag{2}$$

After decomposition, the wavelet reconstruction can be done by the follwoing function:

$$x_n = \frac{\delta j \delta t^{1/2}}{C_\delta \psi_0(0)} \sum_{j=0}^{J} \frac{\Re\{W_n(s_j)\}}{s_j^{1/2}}, \tag{3}$$

Fuzzy Logic

Lotfi Zadeh in 1965 introduced fuzzy logic with the basic idea of *inclusion, union, intersection, complement, relation* and *convexity*. With *fuzzy logic*, human expert system can be implemented in machine language system easily and efficiently. Fuzzy logic uses et of rules to describe the behavior. Fuzzy set has characteristic of function covering real number in the interval [0,1] [7].

In classical set, small change may lead to significant different so that the use of classical fuzzy set is considered as unfair. Fuzzy set is used to overcome this problem. On fuzzy set, a profile can enter two different sets such as normal and dry, based on how large the existence in the set, which can be see from the membership value. Illustration about to use of fuzzy set is given in Figure 1.

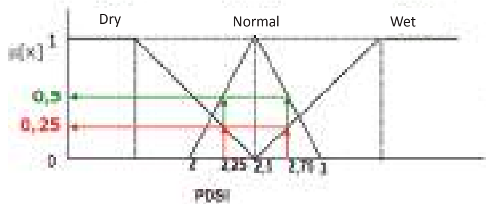


Figure 1. Dry, Normal, and Wet Sets

Using the above fuzzy set, we can see that if the SPI in January is 2,25, the condition can be dry or normal.

Fuzzy Time Series

The basic concept of fuzzy time series is by tranforming the observation into fuzzy set using membership function because the observation has feature to explain the fuzzy set. Fuzzy time series is a method to join linguistic variable wih th process of applying fuzzy ligic in time series to solve the data uncertainty [8].

There ae several definitions required in fuzzy time series method [9]. Supose that U is sample set, where U={u_1,u_2,...,u_n} and U=[D_{min}-D1, D_{max}+D2]= [start,end]. Fuzzy set A is part of U which can be stated as :

$$A = \frac{f_A(u_1)}{u_1} + \frac{f_A(u_2)}{u_2} + \cdots + \frac{f_A(u_n)}{u_n} \tag{4}$$

Where f_A is membership function of fuzzy set, and $f_A : U \rightarrow [0,1]$, $f_A(u_1)$ is membership value of u_i in fuzzy set A and 1≤ i ≤ n. Some required definition about fuzzy time series can be found at [9].

3. RESEARCH METHODOLOGY

Variable

Data used in this research is secondary data of precipitation rate on the period of 1948-2015 which can be obtained from *National Center for Environmental Prediction* – NOAA and available on http://www.esrl.noaa.gov. The dataset is available on grid point basis on the coordinate 6°LU-11°LS dan 95°BT-145°BT. The data structure can be seen in Table 2.

Table 2. Structure of Data

Coordinate	6°LU			...	11°LS		
Month	95°BT	...	145°BT	...	95°BT	...	145°BT
01 - 1948	X 1 1 1	...	X 1 50 1	...	X 1 1 17	...	X 1 50 17
02 - 1948	X 2 1 1	...	X 2 50 1	...	X 2 1 17	...	X 2 50 17
...	...	...	...	...	...	...	...
12 - 2015	X 816 1 1	...	X 816 50 1	...	X 816 1 17	...	X 816 50 17

Steps of the analysis

The steps of the analysis are given as follows:

- Specify the latitude and longitude covering Indonesia region with 1° grid point representing 111,32 km. There are 850 grids in total
- Pre-processing the precipitation dataset
- Calculate the SPI
- Describe the SPI in Indonesia

- Decomposition and reconstruction of wavelet.
- Apply fuzzy logic for prediction with the steps as follows
  - i Define D_{min}, D_{max} as the range of set U.
  - ii Do partition on set U into several interval lengths with the same length $u_1, u_2, u_3, \dots, u_n$.
  - iii Define the fuzzy set on U.
  - iv Fuzzy the historical data :if x ϵ ui then x is changed to fuzzy set Ai where ui has the maximum membership function
  - v Form the Fuzzy relation from the result of step (iv) and determine the fuzzy relation group
- Predict the output on fuzzy set and de-fuzzy the predicted output

4. RESULTS AND DISCUSSION

Standardized Precipitation Index (SPI)

The following figure depicts average of monthly SPI in Indonesia taken from a sample of grid point.

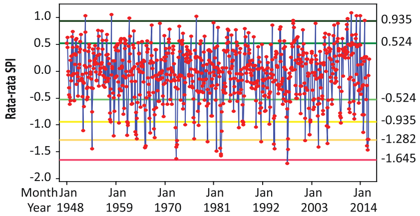


Figure 2. Monthly Average of SPI

Based on Figure 2, we see that Indonesia has been suffered from severe drought in August 1997, where the SPI reached -1,72179. This SPI value lies on the interval of extreme drought, where all Indonesia regions experienced negative SPI. Moreover, severe but not extreme drought happened in July and September 2015. The following figures presents the drought characteristic of 3 regions in Indonesia ie. Padang, Bojonegoro, and Sorong.

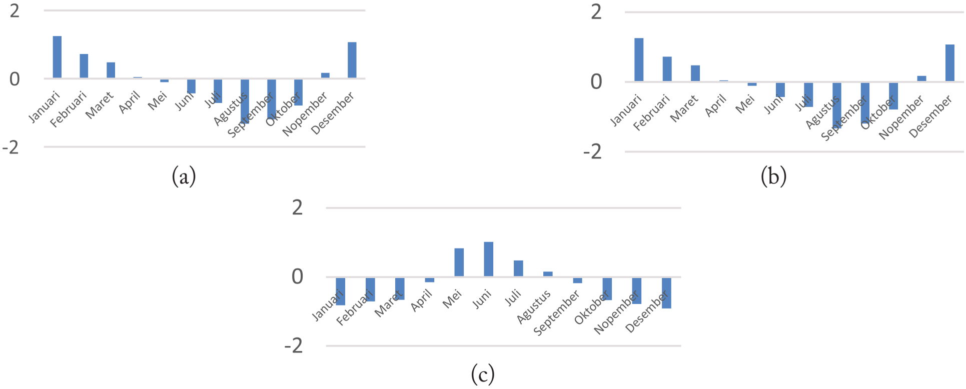


Figure 3. (a) SPI in Padang (b) SPI in Bojonegoro (c) SPI in Sorong

Figure 3 shows that on January to March, the precipitation rate shown by SPI indicates a normal condition. May is the beginning of minor drought and reach the peak on June. This minor drought condition happens for 4 months consecutively to August. The condition back to normal on September and October. November and December have positive SPI indicating wet condition. The drought cycle in Bojonegoro (as shown by Figure 3(b)) is started on June with minor drought and reach the peak to moderate drought on August. September and October start to a minor drought. November is the beginning of rainy season and it has moderate moisture in January.

The drought cycle in Sorong (Figure 3(c)) is different with those two regions located in west part of Indonesia. October is the beginning of minor drought event and reach the peak on December and January. The SPI in February steadily increases over time and reach normal condition in April. May to June is rainy season.

Wavelet

Wavelet used in this research is *Wavelet Fuzzy Logic*, denoted hereafter as WFL.

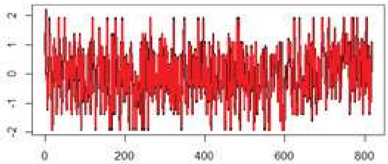


Figure 4. Result of Wavelet (-1°,100°)

The wavelet is applied, and the output of wavelet is used as an input for fuzzy time series. The process is continued with decomposition and reconstruction as the wavelet output. Figure 4 is an illustration of wavelet transform output for dataset at longitude 100° and latitude -1° , where black line is the raw data, and red line is the reconstructed data.

Fuzzy Time Series

The first step is to define the range and do partition with the same interval length. Mt is the median of the interval.

Table 3. Set of U

Range	Interval		Mt
u1	-3	-2,75	-2,875
u2	-2,875	-2,625	-2,75
…	…	…	…
u48	2,75	3	2,875

After getting the interval and Mt for every u_i, the next step is fuzzification of the historical data based on the membership degree with triangle membership function. Afterwards, the Fuzzy Logical Relationship (FLR) and Fuzzy Logical Relationship G*roup* (FLRG) can be calculated, and they are listed in Table 4.

Table 4. FLRG, M(t), and Defuzzification

F(t - 1)	*F(t)*	*M(t)*	$\hat{y}$(t)
A_1	*→ 0*	*-2,875*	*-*
…	*…*	*…*	*…*
A_8	*→ $2A_9$*		*-1,875*
A_9	*→ $2A_8$,$2A_9$,$2A_{13}$,$1A_{14}$,$1A_{15}$,$1A_{16}$ $2A_{17}$,$1A_{18}$,$2A_{20}$,$3A_{21}$,$2A_{25}$,$2A_{26}$, $1A_{28}$,$1A_{29}$,$1A_{34}$*	*-2 -1,875*	*-0,7163*
A_{47}	*→ 0*	*2,875*	*2,875*

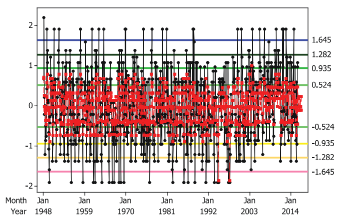
F(t-1) is past data, F(t) is data at time t (current), M(t) is the median of fuzzy set and y(t) is defuzzification i.e. the data of prediction. The way to interpret the values in the table is as follow : If SPI yesterday (F(t-1)) A_8 thus there will be several possibilities for today's prediction (F(t)) i.e. A_9 twice. To obtain the prediction value for today (F(t)) if yesterday is A8 thus the prediction is calculated as:

$$\hat{y}(t) = \frac{2 \times (M(t)A_9)}{2} = \frac{2 \times (-1.875)}{2} = -1{,}875$$

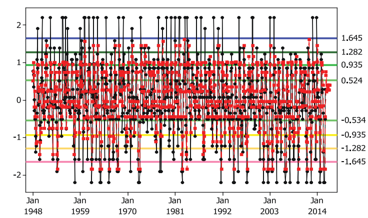
From the above calculation, the prediction for today if yesterday A_8 is -1,875.

4. PREDICTION RESULT

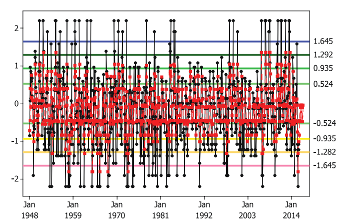
This section performs the SPI prediction in 2016 over Indonesia region. As an example, we perform again the SPI prediction for the above three regions:



(a)



(b)



(c)

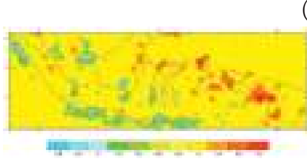
Figure 5. (a) Prediction in Padang (b) Prediction in Bojonegoro (c) Prediction in Sorong

For Padang (Figure (5a)), the classification accuracy is 35,17% and during 2016 is projected to be safe from drought. Bojonegoro (Figure 5b) is predicted to have positive SPI. It is predicted that this region will have slightly moisture on January and February 2016. Sorong (Figure 5c) has classification accuracy of 34,8% and predicted to have minor drought. The classification accuracy for those three regions is very low. However, other regions shows better results with accuracy above 60%. The following part discusses the prediction results in Quarter I and Quarter II 2016.

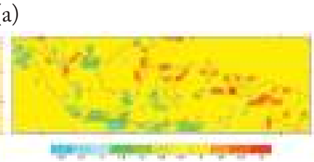
Prediction in Quarter I 2016

The condition over Indonesia region is predicted to be severe drought in January i.e. in Okabe Merauke, shown by the red color on Figure 6a.

(a)



(b)



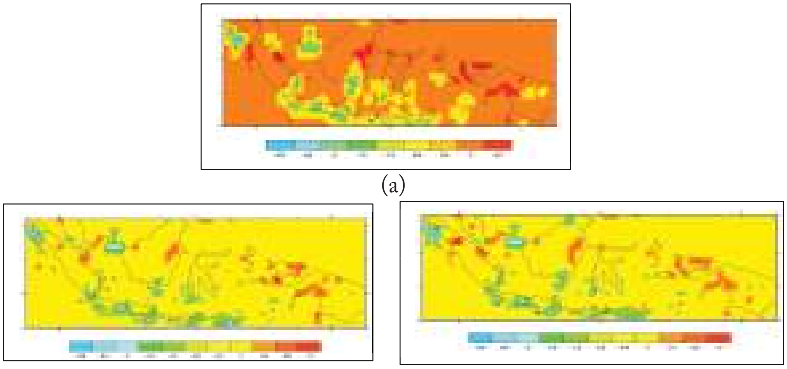
(c)

Figure 6.
(a) Drought Prediction in January 2016 (b) Drought Prediction in February 2016 (c) Drought Prediction in March 2016

The severity level of the drought based on SPI reach about -1,294 which is severe drought. The pattern from January to March show that drought is in which it is only in Halmahera and Papua Barat, spread out to North Sulawesi, East Kalimantan and Riau. Beside the drought level, we can also know the pattern of wet region or unvulnerable regions toward drought. On the first quarter, the area of wet region becomes narrower.

Prediction in Quarter II

The beginning of second Quarter is predicted to be normal in all regions. There are several regions such as Aceh, South Sumatra, Java, Bali, NTT, NTB and South Kalimantan which are predicted as wet regions. Drought is predicted to begin on May as shown in Figure 7, where moderate drought will happen in Riau, North part of Nort Sumatra, East Kalimantan and Papua. June will have similar condition as May, but with larger area of moderate drought such as North Sumatra.



(a)

(b) (c)

Figure 7. Drought Prediction in April, May and June 2016 Consecutively

5. CONCLUSION

Based on the analysis, we conclude that Indonesia will be suffered from severe drought in January 2016 in the following regions : Okaba and Merauke. Papua and Kalimantan are two regions which are vulnerable toward drought event in 2016. Halmahera is predicted to be vulnerable to drought in the first quarter. North Sulawesi is vulnerable to drought on the first and third quarter. These results have been validated with the real condition in Indonesia and shows a good reliability, which means also that using reanalysis data to predict drought events with SPI is a good approach.

References

Balai Hidrologi. (2003). *Permasalahan Kekeringan dan Cara Mengatasinya*. Puslitbang Sidoarjo.

BAPPENAS. (2014). *Rencana Aksi Nasional Adaptasi Perubahan Iklim*. Jakarta.

Amberkahi, T., Anggraeni, T. K., Pribadi, Y. H., Mardanis, D., Febryanti, D., Kusairi. (2014). *Buku Informasi Peta Kekeringan dengan Metope SPI Eds April 2014*. Stasiun Klimatologi Pondok Betung.

National Drought Mitigation Center. (2015). *What is Darought*. USA.

McKee, T.B., Doesken, N. J., & Kleist, J. 1993. The Relationship of Drought Frequency and Duration to Time Scales, Procedings of the 8th Conference on Applied Climatology.

Percival, D. B., & Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press. New York.

Kusumadewi, S., Hartati, S., Harjoko, A., & Wardoyo, R. (2006). *Fuzzy Multi-Attribute Decision Making (Fuzzy MADM).* Yogyakarta: Graha Ilmu

Nguyen, H.T & Wu, B. (2006). *Fundamental of Statistics with Fuzzy Data*, StudFuzz. Springer-Verlag Berlin Heidelberg.

Song, Q., & Chisso, B.S. (1993). Fuzzy Time Series and Its Model. *Fuzzy Sets and System*, 54, 269-277.

CPS34:
Demography & Social Welfare (6)

Pattern of Utilization of Antenatal Care in Nepal (2001-2015)

Jonu Pakhrin Tamang*
Department of Mathematics and Computer Science; Faculty of Science and Technology,
Prince of Songkla University, Muang, Pattani, 94000 Thailand- jpakhrin3@gmail.com

Nittaya McNeil*
Department of Mathematics and Computer Science; Faculty of Science and Technology,
Prince of Songkla University, Muang, Pattani, 94000 Thailand- nittaya.ch@psu.ac.th

Phattrawan Tongkumchum
Department of Mathematics and Computer Science; Faculty of Science and Technology,
Prince of Songkla University, Muang, Pattani, 94000 Thailand- phattrawan.t @psu.ac.th

Sampurna Kakchapati
Department of Mathematics and Computer Science; Faculty of Science and Technology,
Prince of Songkla University, Muang, Pattani, 94000 Thailand- kck_sampurna@yahoo.com

Abstract

Nepal is one of the few countries in the world where the rates of maternal and neonatal mortality remain high. Antenatal care (ANC) plays a vital role in reducing maternal and neonatal mortality and morbidity. In 2001, Nepal adopted the WHO recommendation of at least four ANC visits for all women. Since 2009, ANC service have been provided free of charge. The aim of this study is to investigate the pattern of utilization of ANC in Nepal in order to identify trends and possible high-risk areas. The study is based on a retrospective analysis of data from annual reports between 2001 and 2015 published by the Department of Health Services. A negative binomial model was used to determine the utilization of ANC services. The overall utilization of at least four ANC visits was 51.4% with a steady increase in the utilization over time. There was pronounced geographic variation in ANC utilization with low utilization in the mountainous and Terai regions and low utilization in rural areas. These findings could help public health officials improve the utilization of ANC services in Nepal.

JEL classification: I0
Keywords: Antenatal care; Negative binomial; Mountain region.

1. INTRODUCTION

Antenatal care is the care provided to the pregnant women from the time of conception to the child birth. It is important component of the reproductive health (UN, 2014; USAID, 2015). ANC serves as a platform for the delivery of highly effective health interventions that can reduce preventable maternal and newborn deaths (Lincettto, 2006). It is an opportunity to inform women about the danger signs and symptoms for which assistance should be sought from a health care provider without delay. It offers women appropriate advice and information about the place of delivery based upon her particular circumstances and the health status (WHO, 2003).

World Health organization (WHO, 2016) recommended that each women receives at least four goal-oriented or focused ANC visits. Only half of the women worldwide receive the recommended number of care during pregnancy. Around 85% of women receive at least one ANC visits while the coverage of women who attended the recommended at least four ANC visits is only 58%. However, the coverage of women in the least developed countries who attended ANC visits at least once is 77% followed by 42% of women attended at least four ANC visits (UNICEF, 2016). Among these developing countries, Nepal had the coverage of 58.30% of ANC visits (MoHP and New ERA, 2011). Since, last three years the national average of first ANC visits as percentage of expected pregnancy has increased from 86 % to 96 % in last 2015. The proportion of pregnant women attending at least four ANC visits as per the protocol among women who had first antenatal visits has declined in 2015 in comparison to previous years in all regions. The district level performance on coverage of at least four ANC visits accounts that only four districts have the coverage of more than 80% and a total of 50 districts with the coverage between 40% to 80 % while 21 districts hold below 40% (DoHS, 2015). Government of Nepal implemented safe motherhood programme since 1991. However still there is vast disparities in the utilization of antenatal care all over the country which accounts the maternal mortality as 258 per 100,000 live birth (Dahal, 2013). Globally several studies were conducted on determinants of antenatal care utilization (Ha *et al.,* 2015; Joshi *et al.*, 2014; Sohag *et al.,* 2013; Tran *et al.,* 2012). Statistical model commonly used in the very field of interest are mulitiple linear model and logistic model (Lund *et al.,* 2014; Haddad *et al.,* 2016).

However, the use of negative binomial model has been rarely implemented in determining the pattern of ANC utilization. Thus, this study investigate the pattern variation of utilization of at least four antenatal care visits from 2001 to 2015 in Nepal. This findings could be used by the public health officials to improve the utilization of antenatal care in Nepal.

2. METHODS

Data of antenatal care utilization were compiled from annual reports of Department of Health Services for the 15 years period from 2001 to 2015. The residential area (by district) and year were selected as the determinants in studying the utilization of four antenatal care visits.

The country is divided horizontally into three ecological zones (mountain, hill and terai). It is further divided into 14 zones and 75 districts. Based up on the population density of those areas, from the 75 districts 19 districts were combined together forming 8 aggregated districts and 56 non-aggregated groups, following the criteria as defined by Kakchapati *et al.,* (2013). We thus obtained 64 districts including both the non-aggregated and aggregated districts as shown in figure 1.

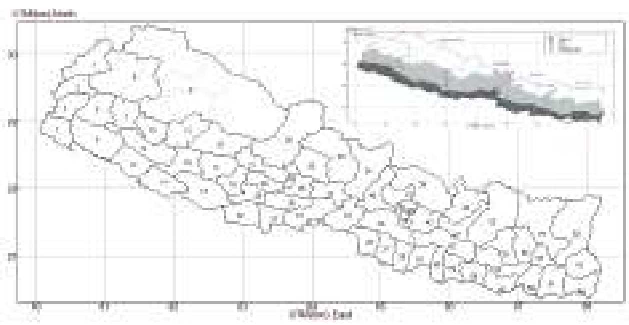


Figure 1. Map of Nepal Showing 64 Districts

Statistical methods

The negative binomial model is an extension of the Poisson regression model that allows for over-dispersion. If λ_{it} denotes the mean ANC utilization in district *i* and year *t*, an additive model with this distribution is expressed as

$$\ln(\lambda_{it}) = \ln(P_{it}) + \alpha_i + \beta_t \quad (1)$$

The terms α_i and β_t represent districts and year effects and P_{it} is the corresponding population at risks in 1000s, respectively, and are centered at 0. The variance of this distribution is $\lambda_{it}(1 + \lambda_{it}/\theta)$ with the Poisson model arising in the limit as $\theta \rightarrow \infty$.

Model fitting is performed by the linearity in the plot of deviance residuals against normal quintiles. Confidence intervals for comparing the adjusted of four ANC utilization within each factor with the overall ANC utilization were obtained using sum contrasts (Tongkumchum & McNeil, 2009). From the model, the confidence intervals for factor-specific of four ANC utilization divided into three groups according to their space entirely above, around and below the over percentage of at least four ANC visits. R program was used for all statistical and graphical analysis.

3. RESULTS

Figure 2 shows the trend in percentage of women who attended at least four ANC visits between 2001 and 2015. Utilization decreased for the first three years and then steadily increased for the next 12 years.

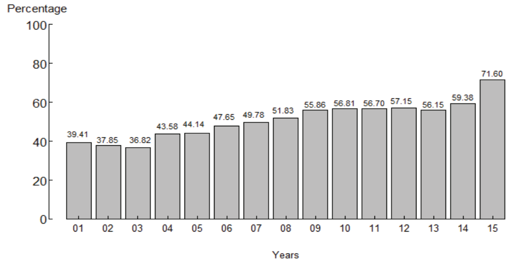


Figure 2. Percentage of Women who Attended at Least Four ANC Visits between 2001 and 2015 A Negative Binomial Model was Fitted to the Data. The Normal Q-Q Shows the Residuals from Fitting the Model are Normal.

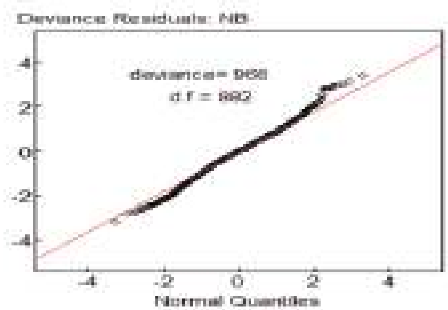
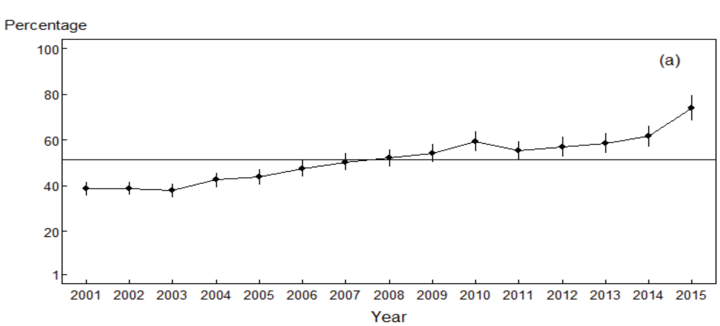


Figure 3. Normal Q-Q plot

The estimated percentage of women who attended at least four ANC visits for each year and each district from the model are displayed separately with 95% confidence intervals in Figure 4. The horizontal line denotes the overall percentage (51.4%). Figure 4(a) shows the trend of ANC utilization over time. There was a steady increase between 2001 and 2015. Figure 4(b) shows the utilization by district. There were 14 districts that had a significantly higher utilization than the overall percentage and 19 that had a significantly lower utilization.



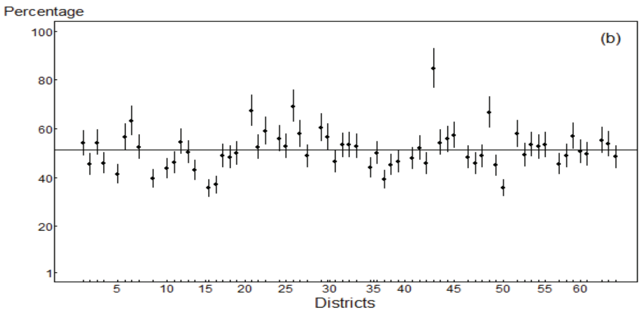


Figure 4. 95% confidence intervals of at least four ANC visits adjusted by year for districts. The utilization of ANC was categorized into three groups: higher than, lower than and similar to the overall percentage and displayed using a thematic map as shown in Figure 5. A low ANC utilization was found in some districts in the mountainous and Terai regions.

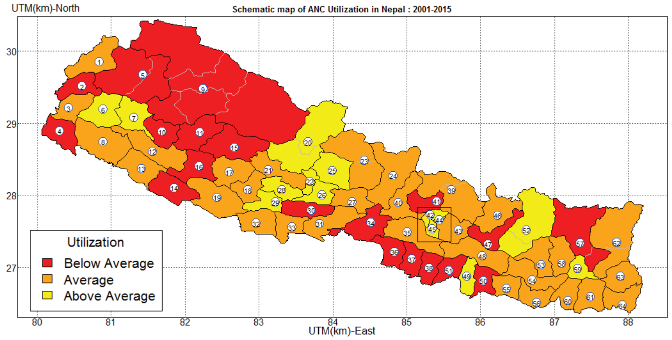


Figure 5. Map of Nepal Showing the ANC Utilization

4. DISCUSSION

This study present at least four ANC utilization among pregnant women from 2001 to 2015 and whether the pattern varies according to the districts and years. Approximately 51% of the respondents completed at least four ANC visits.

The study shows the slow but steady increase in at least four ANC utilization from 2001 to 2010 in Nepal. There was slight drop in 2011 and increment from 2012 to 2015. In the past 15 years there has been a five-fold increase in the percent of women with four or more antenatal visits during their pregnancy (from 9% in 1996 to 50% in 2011) (DoHS, 2011). There are several factors which contribute in the increase in the ANC utilization in Nepal. One of the factor is the scheme of hospital delivery system. Compared to the previous years, there is increase in the number of women delivering at health institution. There has been increase in the percent of institutional delivery which accounts (35%) (MOHP and New ERA, 2011). The institutional delivery has promoted awareness among the pregnant women which bring positive impact encouraging women for having ANC visits. Similarly, the maternity incentive scheme named as "Aama Surakshya Karyakram" has contributed in the motivation among women in utilization of ANC. On the other hand, peer report and exposure to mass medias have also accounted and played a role in the increase in the ANC utilization among Nepalese women by Acharya *et al.* (2015). The study conducted by Sanjel *et al.* (2011) in Nepal revealed that (61.4%) of women had completed at least four ANC visits. The women in Hilly areas of Eastern region of Nepal were more concerned about their pregnancies and aware of complications that arise during this period. As, the study shows (82%) of respondents completed at least four ANC visits by Sah *et al.* (2013).

The result shows low utilization of at least four ANC service in some districts of mountain and terai regions. In the mountain regions lack of basic access to maternal health services, difficult geographical terrain, poorly developed transportation and communication system, poverty, illiteracy, women's low status in the society, political conflict, and shortage of health care professionals and under-utilization of currently available health care services are the major challenges and may affect the low ANC utilization. However the uptake of four or more antenatal care visits is higher in urban areas (75%) than in rural areas (48%) (MoHP and New ERA, 2011).

In developing countries like Nepal most of the people are not familiar with the needs of antenatal visit. Even those who know go to service center at a late stage. Pregnant women do not follow the recommended number of four antenatal visits. Maternity services are still inaccessible to most of rural communities of Nepal.

The study has some limitations. Study used secondary data so most of the informations were limited. Further study could be conducted on assessing the quality of ANC utilization.

5. CONCLUSION

This study identified that there is variation in the utilization of at least four ANC services in different eco development regions. Most of the women from the hilly region are accessible to utilization of ANC services. However there seems to be lack of utilization of four antenatal service in the mountain and terai region of Nepal. Meanwhile, the trend of utilization of at least four ANC service is increasing slowly throughout the year 2001 to 2010. There was slight drop in the year 2011 and progressing to slow increment up to 2015. Thus, the unequal distribution of at least four service utilization reveals that still there are so many pregnant women who are devoid antenatal care services. Therefore, this study aids information to the public health officials to improve the utilization of ANC services in all the regions of Nepal.

REFERENCES

Acharya, D., Khanal, V., Singh, J.K., Adhikari, M., & Gautam, S. (2015). Impact of mass media on the utilization of antenatal care services among women of rural community in Nepal. BMC Research Notes, 8, 345-350.

Dahal, R. K. (2013). Utilization of Antenatal Care Services in Rural Area of Nepal. International Journal of Collaborative Research on on Internal Medicine and Public Health, *5*(2), 120–131.

DoHS. (2015). Annual Report. Kathmandu, Nepal.

Ha, B. T., Tac, P. V., Duc, D. M., Duong, D. T., & Thi, L. M. (2015). Factors associated with four or more antenatal care services among pregnant women: a cross-sectional survey in eight south central coast provinces of Vietnam International Journal of Women's Health, 7, 699-706.

Haddad, S.G., DeJong, J., Terrei, N., Restrepo, C.M., Perin, J., Vaz, L., Newby, H., Amouzou, A., Barros, J.D.A., & Bryce, J. (2016). Pattern and determinants of antenatal care utilization. Journal of Global Health, 6(1), 1-18.

Joshi, C., Torvaldsen, S., Hodgson, R., & Hayen, A. (2014). Factors associated with the use and quality of antenatal care in Nepal: a population-based study using the demographic and health survey data. BMC Pregnancy and Childbirth, 14(94), 1-11.

Kakchapati, S., Yotthannoo., S., & Choonpradub., C. (2010). Modelling tuberculosis incidence in Nepal. Journal of Asian Biomedicine, 4(2), 355-360.

Lincetto, O., Mothebesoane-Ano, S., Gomez, P., Munjanja, S. & Munjanja, S. (2006). Chapter 2: Antenatal care, Opportunities for Africa's Newborn: practical data, policy and programmatic support for newborn care in Africa. Geneva: World Health Organistion, p. 51-62.

Lund, S., Nielsen, B.B., Hemed, M., Boas, M.I., Said, A., Said, K., Makungu, H.M., & Rasch, V. (2014). Mobile phone improve antenatal care attendance in Zanzibar. BMC Pregnancy and Childbbirth, 14:29.

MoHP, New ERA, and ICF International Inc. (2011). Nepal Demographic and Health Survey, Kathmandu, Nepal.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna Austria. 2016; 0 :{ ISBN} 3-900051 - 07-0. doi:10.1038/sj.hdy.6800737.

Sah, R.B., Gaurav, K., Baral. D.D., Jha, N., & Pokharel, P.K. (2013). Antenatal practices in in hilly area of eastern region of Nepal. Journal of Universal College of Medical Sciences, 3(4), 12-15.

Sanjel, S., Ghimire, R.H., & Pun, K. (2011). Antenatal practices in Tamang community of hilly area in Central Nepal. Kathmandu University Medical Journal, 9(2), 57-61.

Sohag, A., Memon, S., Bhatti, M., & Azeem, M. A. (2013). Factors affecting Utilization of Antenatal Care the Opinion of pregnant Women. Pakistan Journal of Physiology, 9(1), 17–19.

Tongkumchum, P., & McNeil, D. (2009). Confidence intervals using contrasts for regression model. Songklanakarin Journal of Science and Technology, 31; 151-156.

Tran, T. K., Gottvall, K., Nguyen, H. D., Ascher, H., & Petzold, M. (2012). Factors associated with antenatal care adequacy in rural and urban contexts-results from two health and demographic surveillance sites in Vietnam. BMC Health Services Research, 12(1), 30-40. http://doi.org/10.1186/1472-6963-12-40.

WHO. (2016). Pregnant women must be able to access the right care at the right time. Available from: http://www.who.int/mediacentre/news/releases/2016/antenatal-care-guidelines/en/.pdf [Jan 8, 2017].

WHO. (2003). Antenatal Care in Developing Countries: Promises, Achievements and Missed Opportunities.

UN. (2014). Framework of actions for the follow-up to the Programme of Action of the International conference on Population and Development Beyond 2014.

UNICEF. (2016). Only half of women worldwide receive the recommended amount of care during pregnancy. The United Nation's Children Fund:2016.http://data.unicef.org/maternal-health/antenatal-care.

USAID. (2015). Ending preventable maternal mortality: USAID maternal health vision for action-Evidence for strategic approaches. Washington DC. Geneva: Available from: http://www.who.int/gho/maternal_health/reproductive_health/antenatal_care_text/en/.pdf [October 11, 2015].

Direct and Indirect Effect of Urbanization, Unemployment, Poverty, and Absorption of Non-Agricultural Labor Force on the City Development Performance

Indryanty Yakub*
BPS-Statistic Sulawesi Tenggara, Indonesia – indryanty@bps.go.id

Titik Harsanti
Institute of Statistics, Jakarta, Indonesia – titik@stis.ac.id

Abstract

Cities in many developing countries like Indonesia often do not work and participate fully as the concentration of economic activities, social and public administration. The gains of the city performance can be seen by City Development Index (CDI). Base on many literature, there are four issues effect on city performance such as urbanization, unemployment, poverty and the absorption of non-agricultural labor force. CDI has been counted in Indonesia in 2009 and 2010, but did not see the effect of these four issues. This study will describe city performance and examines the direct and indirect effects of the four issues on the city development index in Indonesia 2013.

The data comes from publication such as Human Development Index, survey of the environmental behavior, Regional Review of county and city Gross Regional Domestic Product, labor market indicators, etc. and household data of national socio economic survey Indonesia 2013. The study employed path analysis to examine the significances effect of the four issues on the city development index in Indonesia 2013. CDI is built of infrastructure, waste, education, health and cities product dimension.

Descriptive analysis shows that the average of 98 city CDI in Indonesia is 68.10. The highest CDI achievement is South Jakarta city at 77.63 and the lowest is Subulussalam city, Aceh province amounted to 59.58. Base on CDI dimension, waste level still in the medium low (under 40%). Path analysis shows that among four issues, urbanization, unemployment and poverty have direct effect on the CDI. Absorption of non-agricultural labor force and Urbanization affects indirectly on CDI. Urbanization influences 56,2%, the highest effect both direct and indirect to CDI.

It can be concluded that Indonesia has Medium High CDI, indicates a good city perform on average. Among five dimension of CDI, waste management has the lowest performance and need to be more improved. Urbanization is the most influence issues on the city performance, direct and indirect.

Keywords: City Development Index (CDI), Urbanization, Unemployment, Poverty and absorption of non-agricultural labor force.

1. INTRODUCTION

Cities in many developing countries like Indonesia has not made its function properly as the main concentration of place where people stay, and do not participate fully also as the concentration of economic activities, sosial, and public administration. Its happened because the low of city performance of the city itself (Adisasmita, 2014). City management should be on modern way, high capacity, professional, dynamic, and should have the right anticipation programme on future (Amins, 2009). However in the fact, the supplying of infrastructure facility and public services which is needed is available on limited number. It means that the city management still on the low performance, and it has made an unequality in many dimensions of social human life, therefore the serious problems of the city appears. Base on many literature, there are four issues effect on city performance such as urbanization, unemployment, poverty, and the absorption of non-agricultural labor force.

The problems is caused by a rising of cities people with high growth speed. United Nations (UN) (2001) notes that number of world people whos lived in the city with afraid feeling, affected with poverty and trapped to live on the slam area, is nearly one billions and it will always increasing up to 1.6 billions in 2030. In addition its also note if that would be more people, up to 2.5 billions of people who will not get the proper sanitation facilities. UN has predicted if in 2050 that would be around 70 percent of world population who will stay in the city, and also its been predicted if on that year the air of the city would covered more than a half of the emition of the global greenhouse gases and 75 percent of the global energy consumption.

In Indonesia, it has been projected if that would be an increase of number of cities people sharply from 49.8 percent in 2010 to 66.6 percent of people in 2035 (BPS, 2010). The density of people is also shows the increasing trend with the cities in Java island as the populous area which is increase in range 727 people/km^2 to 12.592 people/km^2 (DKI Jakarta) on 2000 become in range 803 people/km^2 to 15.015 people/km^2 (DKI Jakarta) on 2013 (Badan Pusat Statistik (BPS/Statistics Indonesia), 2014), and it notes the migrants as the majority people who has filled the city area.

The new comer whos came from rural area or other small city are need a house as a place to lived and also need a proper work to do. The government of those city be expected to make it available in sufficient quantities of it (Adisasmita, 2014). If it doesnt happened the emergence of various urban problems, especially unemployment and poverty will appear. That happened in Indonesia, Sjafrizal (2012) whos made a research to the results of 2010 population census has examined that the flow of migration of people from rural area to urban area has been higher than the increase of amount of job which has to made in the urban area. Consequently the unemployment on the urban area are sloping increase so it triggered an increase number of poor people.

Many problems are made by the new comer. As the research of Firman (2007) said that the city area has been very tightly and losses nearly all its green openspaces. Zahnd (2008) also examines the appearing of transportation problem such as jamming traffic which caused by the increasing of number of car. In Jakarta at that time (2008) the number of car is fivefold higher than 20 years before. In 1980 the number of car owner was increased 12 percent per year until 1997, with the addition of the road was only been increased two percent per year.

As a related of the job supplying especially for the new comer, its relies heavily on the potential economic sector of each city, such as an agricultural and non-agricultural potential economic sector (Adisasmita, 2014). The higher job opportunies will increase an absorption of labor force, especially in non-agricultural sector. Consequently its also make income of them being increased. If its steadily happened the poverty will fall into the lowest point.

In this case, the most important thing to being controlled is the number of cities people especially the migrants. As stated in Karl Fox and Brian Berry (1983) concept that isn't big or breadth of the city affected the growth of the city, but it is the population of the city itself (Adisasmita, 2014). The government stand as the city controller has a big responsibilty to make a higher of city performance.

The gains of the city performance can be seen by City Development Index (CDI). CDI has been counted in Indonesia in 2009 and 2010, but did not see the effect of these four issues. This study will describe city performance and examine the direct and indirect effects of the four issues on the city development index in Indonesia 2013.

Section 2

According to Indonesian language dictionary, city is the concentration area where people stayed, with high density of people. It has the modern facilities and also a majority of worker in a non-agricultural sector. The CDI is being introduced for the first time by United Nation Centre for Human Settlement (UNCHS/UN-HABITAT) in 1998 and has made a publication report such as Global Indicator Database and State of World Cities Report that has been released every year. Asian Development Bank (ADB) in 2001 has made a research too with the CDI and has made a publication report also named Cities Data Book 2001. The World Bank also has made a programme to comparing the city using the CDI named Global City Indicator Facility in 2009. The CDI is expansion of the Human Development Index (HDI). The CDI is more better than HDI (UNCHS/UN-HABITAT, 1998). The CDI can comparing the city performance over time which can not be done by HDI. The HDI can not see the real changing effect of the investment from the government in a short time such as one or each two years although on the main dimension of the HDI itself. CDI is also has nearly all dimension as the indicate the city performance which can not be seen in HDI (BPS : 2010). Based on the research of BPS, there are other indicator that can also examine the gains of city performance named Regional Development Index (RDI) but it has a numerous dimension to review and there was a limited data to get as a component of its index, therefore its not a suitable measurement to have to measure the gains of city performance (BPS, 2010).

The technique used to construct the CDI is similar to that used by UNDP for their Human Development Index. Separate sub-indices are constructed and combined to create a composite index. Thus, the CDI is based on five sub-indices such as City Product, Infrastructure,Waste, Health and Education. The values of which range from 0 to 100 (UN HABITAT, 1998).

The City Development Index is calculated according to the formula in the table below. It has separate sub-indices for Infrastructure, Waste Management, Health, Education, and City Product, which are averaged to form the CDI. Each sub-index is a combination of several indicators that have been normalized to give a value between 0 and 1. Because the variables used to make up the CDI are strongly related to each other, there are a number of ways to calculate the CDI that give almost identical results. The weightings given to each indicator have been initially calculated by a statistical process called Principal Components Analysis and then simplified.This formulation of the index by and large uses the same formula as in UNDP Human Development Report (1999) for the Health, Education and City Product sub-indices. For meaningful ranking of cities, the index requires data that are essentially complete, robust and precise, so not many variables are suitable. All the underlying data had to be checked for accuracy and completeness. Where there were missing data or based on very inaccurate estimates, they were either replaced by data from another national city of similar size, by country-wide figures (or national urban data, if available or by figures for a near city or place at a similar level of development (but only if absolutely necessary). Also, Formal waste disposal or Wastewater treated is taken as zero if not provided. Where City Product was not provided, it was calculated so that City Product x Household size = 0.45 x Mean Household Income (which is similar to the main estimation formula). For most transition countries 0.35 x Household Income is used since, in transition economies, much GDP goes into indirect services and subsidies (UNCHS/UN HABITAT, 1998). The CDI formula is describing below.

Table 1. The City Development Index Formula

Indeks	Formula
(1)	*(2)*
Infrastructure	25 x (Water connections+ Sewerage + Electricity + Telephone)
Waste	50 x (Waste water treated + Formal solid water disposal)
Health	(Life expectancy – 25) x 50/60 + (32 – Child Mortality) x 50/31,92
Education	25 x (Literacy + Combined enrollment)
City Product	(log city product – 4,61) x 100/5,99
City Development	(Infrastructure Index+Waste Index+Health Index+Education Index+City Product Index) / 5

Sumber : UNCHS/UN-HABITAT (1998)

ADB (2001) has made a research and found that CDI is correlate well with the city level. The higher CDI following the higher city level. CDI is also has a high correlation with the poverty and urbanization. Adisasmita (2014) said that city performance has been influenced by expantion of the job opportunities, trade, and industry sector.

Sovani N.V (1988) has founded that the excess of urbanization is the first cause which make unemployment appear. (Sjafrizal, 2014). Tindaon (2010) said that the absorption of the non-agricultural labor force in the city is influence the unemployment. Tindaon (2010) also had founded from his research that the growth of the people caused by the urbanization is significantly has influenced the absorption of the non-agricultural labor force such as industry, electricity, gasses and clean water also the trade. Sjafrizal (2014) also said that if there were a higher job opportunities in the city especially in non-agricultural sector such as industry, trade, and public services the absorbtion of non-agricultural labor force will increase. Directly it can make income of the people in those city increase also, therefore the number of poor will sicnificantly decrease.

Section 3

This study is using the data which comes from publication such as HDI, survey of the environmental behaviour, regional review of the country and city Gross Regional Domestic Product, labor market indicator, etc and household data of national socio economic survey Indonesia in 2013. This study used 98 cities as a subject. The research variables used is as follows the percentation of household who has the access to get a clean water, proper sanitation facilities, electricity, and telephone. Its used to count the infrastructure index. The second is the percentation of household with waste water treated such in drainage, open sea, and river, also formal solid waste disposal in 2013. Its used to count the waste index. The third is life expectancy and child mortality as a component of health index. The forth is literacy and combined enrollment as a component of education index. The fifth is the Gross Regional Domestic Product as a componen of city product index. The urbanization variabel is counted by percentation of the urban people in 2013. Unemployment is counted by unemployment rate. The poverty is counted by percentation of poor people in 2013, and the last is the absorbtion of non-agricultural labor force variable that has counted by the percentation of non-agricultural labor force in 2013.

Caused by the limited data of child mortality, so that this indicator is had been counted with interpolated estimation procedure with west model on Coale and Demeny model with Trussel Equation. The base comparing model is made by using the comparing with the life expectancy in same period of time. The counting procedure is used the Microsoft Excel 2010 programme, SPSS 20.00 and MORTPAk 4.1.

The study employed path analysis to examine the significances effect of the four issues on the CDI in Indonesia 2013. The framework of this study is describing below.

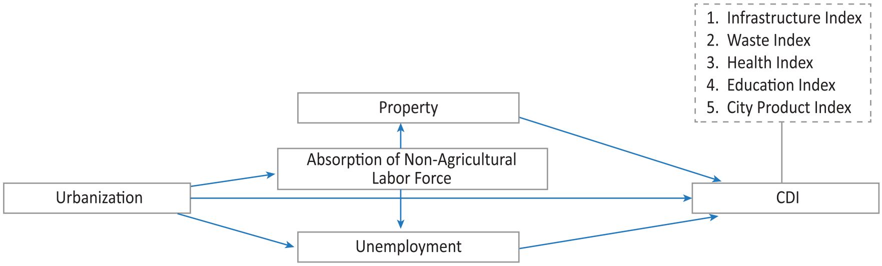


Figure 1. The Framework of Study

The hypothesis of this study is there is a direct correlation of urbanization, unemployment, and the poverty to CDI. Absorbtion of non-agricultural labor force has indirect effect to CDI pass through unemployment or poverty lines.

Section 4

CDI is count the city performance in index form base on development scale of each city. Descriptive analysis shows that the average of 98 city CDI in Indonesia is 68.10. It means that the city in Indonesia is The highest CDI achievement is South Jakarta city at 77.63 as the metropolitan city and the lowest is Subulussalam city, Aceh province amounted to 59.58 as the small city. CDI of the metropolitan city in Indonesia is as follow.

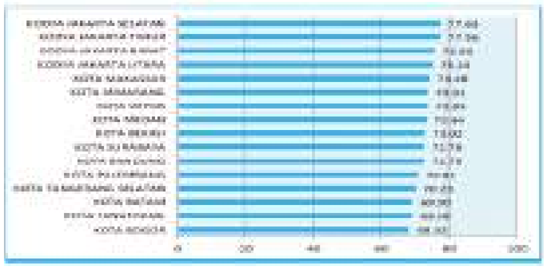
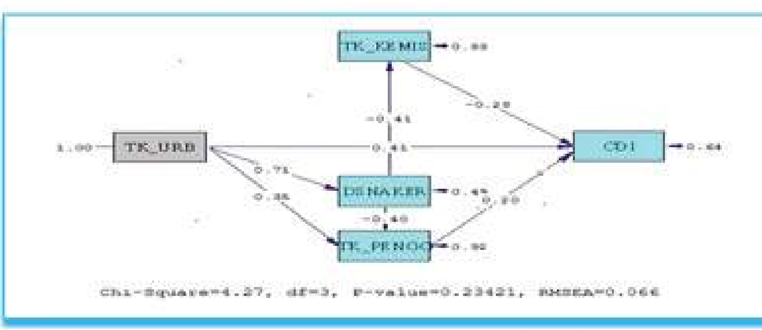


Figure 2. The CDI of the Metropolitan City in Indonesia 2013

The descriptive analysis shows that the increasing CDI in the minimum gains is followed by increasing of city level. Its also cause the increasing of urbanization number and the absorption of non-agricultural labor force. On the other hand unemployment and poverty are steady when the CDI increasing. As a comparing the average CDI of each city level with sub-indices, the descriptive analysis shows that the average of 98 cities CDI in Indonesia is 68.10. The gains of CDI in each city still at the Q5 (in range 60 – 80). It means that in average the city performance in Indonesia is in medium high performance and its still need to increase more. According to distribution of those city, there were 48 cities or 48,97 percent which has CDI under the average of gains of national CDI. The waste index is been the lowest gains index of all (under 40 percent). It means that in every city in Indonesia, the waste sector is been the bad sector that has to be more improved by the government.

Path analysis shows the affect correlation direct and indirectly of the urbanization, unemployment, poverty, and the absorption of non-agricultural labor force. For data processing used the LISREL 8.70.

The results of path analysis is show below on the path diagram as follow.



*TK_KEMIS is refers to poverty, TK_URB is refers to Urbanization, DSNAKER is refers to the absorption of non-agricultural labor force, TK_PENGG is refers to Unemployment

Figure 3. The Path Diagram

The path structural equation is as follows

CDI = 0.41*TK_URB + 0.20*TK_PENGG - 0.28*TK_KEMIS; R^2 = 0.36

TK_KEMIS = - 0.41*DSNAKER; R^2 = 0.17

DSNAKER = 0.71*TK_URB; R^2 = 0.51

TK_PENGG = - 0.40*DSNAKER + 0.35*TK_URB; R^2 = 0.08

After the path coefficient known, the path significant test should do to examine the path. The significantly of each path is been seen based of t value with α five percent is 1.96. The path coefficient isn't being significant if the t value in the range -1,96 to 1,96. The test with LISREL 8.70 shows that none of path coefficient are being in the range -1,96 to 1,96, it means all coefficient is significant.

The f- test should be done too, according to the path diagram that showing there were two equation that being formed by some exogenous variable, such as in a first and in a fourth equation. The result shows that all exogenous variables are significantly affect endogen variable.

The decomposition model that been formed is as follows

Table 2. The Decomposition Model

Variable	Direct Effect	Indirect Effect			Total Effect
		Pass through TK_PENGG	Pass through TK_KEMIS	Pass through DSNAKER & TK_KEMIS	
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
TK_URB → CDI	0,41	0.07		0.082	0.562
TK_PENGG → CDI	0.2				0.2
TK_KEMIS → CDI	-0.28				-0.28
DSNAKER→ CDI		-0.08	0.115		0.035

The table shows that :

The urbanization has a total effect to CDI in the amount of 56,2 percent. Directly urbanization also affect the CDI in the amount of 41 percent, and indirectly affect the CDI seven percent pass through unemployment, and simultaneously affect the CDI in the amount of eight percent through the absorbtion of non-agricultural labor force and poverty.

Unemployment has a total effect two percent to CDI directly. The path coefficient is positive and significant. It means that the higher unemployment as same as the higher CDI in the city.

Poverty has a total effect in amount of 28 percent directly. The coefficient is negative and significant. It means that the higher poverty will decrease the CDI.

The absorption of non-agricultural labor force has a total effect in amount of 3.5 percent to CDI. The coefficient is positive and significant, it means that the higher absorption of non-agricultural labor force will also increase the CDI. The absorption of non-agricultural labor force doesn't have a direct effect to CDI, but it has indirect effect eight percent through unemployment and 11,5 percent through poverty.

The goodness of fit test is has to do according to examine the model founded. The test using the 17 spesific criteria. According to the test, the model founded is a fit model.

2. CONCLUSIONS

It can be concluded that Indonesia has Medium High Performance as an average CDI in Indonesia that gains 68.10. Its indicates a good city perform on average. Among five dimension of CDI, waste management has the lowest performance and need to be more improved. Urbanization, unemployment, poverty and the absorption of non-agricultural labor force have a significant effect to city performance. The urbanization is the most influence issues on the city performance, direct and indirect.

REFERENCES

Adisasmita, Rahardjo.H. (2014). Teori Pertumbuhan Kota. Yogyakarta: Graha Ilmu.

Amins, Achmad. (2009). Manajemen Kinerja Pemerintah Daerah. Yogyakarta: Laksbang Presindo.

Arifianto, Eko. (2010). *Mengukur Kinerja Kota – kota di Indonesia dengan Pendekatan City Development Index (CDI) Kajian Studi pada 32 kota di Pulau Jawa Tahun 2008* [Tesis]. Jakarta : FE-UI

Asian Development Bank (ADB). (2001). *Urban Indicators for Managing Cities.* ADB.

Badan Pusat Statistik. (2010). *Penyempurnaan Penyusunan Indeks Pembangunan Regional*. Jakarta : BPS

Badan Pusat Statistik. (2012). *Analisis Data Kemiskinan berdasarkan Data Pendataan Program perlindungan Sosial (PPLS) 2011*. Jakarta : BPS

Barrientos, Armando. (2010). *Social Protection and Poverty, Social Policy and Development Programme Paper Number No. 42.* E-Paper. United Nations Research Institute for Social Development.

Budihardjo, Eko. (2003). *Kota dan Lingkungan Pendekatan Baru Masyarakat Berwawasan Ekologi*. Jakarta. LP3ES Indonesia.

Firman, Tommy. (2007). *The Patterns of Indonesian Urbanization 1980-2007* [Tesis]. Bandung:ITB.

Ghozali dan Fuad. (2008). *SEM Teori dan Konsep dengan Program LISREL 8.80*. Semarang : BP UNDIP.

Ilmar, Aminuddin. (2014). *Hukum Tata Pemerintahan*. Jakarta : Prenadamedia Grup.

Johnson, Richard A. Dkk. (2007). *Applied Multivariate Statistical Analysis.* USA. Person Education, Inc

Joreskog, Sorbom. (2004). *Interactive LISREL SSI Scientific Software International*. LISREL Homepage.

Latan, Hengky. (2013). *Structural Equation Modeling Knsep dan Aplikasi Menggunakan Program LISREL 8.80.* Bandung : Alfabeta.

Mulyandari, Hestin. (2011). *Pengantar Arsitektur Kota*. Yogyakarta : PT. Andi Offset.

Riadi, Edi. (2013). *Aplikasi LISREL untuk Penelitian Analisis Jalur.* Yogyakarta : PT. Andi Offset.

Riduwan, Kuncoro. (2013). *Carea Menggunakan dan Memaknai Path Analysis (Analisis Jalur)*. Bandung: Alfabeta.

Romdiati, Haning, dkk. (2011). *Pertumbuhan Penduduk dan Kesejahteraan*. Jakarta : LIPI Press Anggota IKAPI.

Sjafrizal. (2012). *Perencanaan Pembangunan Daerah dalam Era Otonomi*. Jakarta : PT. Grafindo Persada Sjafrizal. (2014). *Ekonomi Wilayah dan Perkotaan*. Jakarta. Raja Grafindo Persada.

Sjafrizal. (2014). *Ekonomi Wilayah dan Perkotaan*. Jakarta. Raja Grafindo Persada

Tindaon. (2010). Analisis penyerapan Tenaga Kerja Sektoral di Jawa Tengah (Pendekatan Demometrik) [Skripsi]. Universitas Sumatera Utara

Todaro, Michael P.dkk. (2003). *Pembangunan Ekonomi Di Dunia Ketiga.* Jakarta. PT. Erlangga

UNCHS/UN-HABITAT. (1998). Epilogue The City Development Index (CDI). Diakses pada 23 Desember 2014. http://www.un.org/ga/Istanbul+5/116.pdf

Wibowo, (2014).*Manajemen Kinerja.* Jakarta. Raja Grafindo Persada.

Wijanto, Setyo Hari. (2007). *Structural Equation Modeling dengan LISREL 8.8 Konsep dan Tutorial.* Graha Ilmu. Jakarta

World Bank Institute. 2005. *Introduction to Poverty Analysis : poverty Manual.* World Bank Institute.

World Wealth Report (2015). *World Wealth Report 2015*. Diakses 23 Januari 2015 melalui http://www.worldwealthreport.com

Zahnd, Markus. 2008. *Model Baru Perancangan Kota yang Kontekstual*. Semarang. Kanisius

A Deep Study of Education Quality Mapping of the Districts in Java

Setia Pramana*
Institute of Statistics, Jakarta Indonesia – setia.pramana@stis.ac.id

Achmad Fauzi Bagus Firmansyah, Imam Habib Pamungkas, and Muhammad Arief Eko Pratama
BPS Indonesia

Abstract

Education development is one of the main priorities in the national development agenda. In order to determine which regions, cities or districts which attain good development and distribution of education, grouping districts/cities based on indicators of education is needed. In this article we would like to map the district based on the education indicators, study the spatial effect, and investigate cluster specific indicators.
For this purpose, eleven education indicators such as student-teacher ratio, Illiteracy rate, Net enrollment ratio, and dropout rate from BPS Statistic Indonesia are used and several methods are implemented and compared. First for simple clustering the Fuzzy C-Means Gustafson Kessel is implemented. To take into account for spatial effect the Fuzzy Geographically Weighted Clustering is carried out. The previous method use all indicators in clustering the district, though in the reality some districts may share similar condition on only a few indicators. Thus a simultaneous clustering method of district and indicators is applied. In this case the Cheng and Church biclustering algorithm is used.
The result reveals that distinct clusters are observed across Java indicating the education attainment still remain disperse, though spatial effect is seen as the closer to the city the better the indicators. Furthermore some districts are affected with their neighboring conditions. Moreover, cluster specific indicators is obtained. It provides better insight on the cluster obtained and can be used for focusing the education improvement program.

Keywords: Biclustering, Spatial clustering, Fuzzy clustering, Education indicators.

1. INTRODUCTION

As one of the major economies of Southeast Asia, Indonesia through its government focusses on its social and economic development. To reach this goal, development of human capital is crucial. Education and skills are central to Indonesia's growth prospects in the next decade. PISA shows that Indonesian students performs three years behind as compared to the OECD average. Furthermore, Indonesia's Human Development Index (HDI) was ranked 111th out of 188 countries.

This situation encourage the Indonesia government to improve the quality of basic education as top development priority. It is now implemented in the regulation that 20% of government expenditure is targeted for the education development. The OECD/ADB suggested that education development in Indonesia can be focused on quality, participation, and efficiency. As the region in Indonesia is so many and diverse, understanding the strengths and also weaknesses of every region is important for efficiency of the education development program by providing the right policy for the right district/region.

Despite the fact that Java is the most populated island and largest economy in Indonesia, education development still the main problem. There are significant educational gap between cities area and district in Central and East Java. Some districts in Eastern region and Madura Island are the area with higher inequality than others. Mapping of districts based on their education indicators is important for local and central government policy.

Currently there are several studies on clustering or mapping of education in the district level in In Java.

However, these studies ignored the spatial effect of the education quality and use less indicators. Moreover, there is few investigation on region's specific indicators which is very useful in region policy related education.

This article is aimed to not only cluster the districts in Java based on the education indicators and study the spatial effect of the education, but also investigate cluster specific indicators.

2. METHODS

Data

For the study, 11 indicators obtained from The National Socioeconomic Survey (Susenas) 2014 and Region in Figure 2014 of 118 Districts in Java were used (see Table 1).

Table 1. Education Quality Indicators

Dimension	Indicators
Quality	Adult Literacy Rate
	Expected Year of Schooling
	Average Years of Schooling
Participation	Net Enrolment Rate in Primary School (PS)
	Net Enrollment Ratio in Junior High School (JHS)
	Dropout Rate in PS
	Dropout Rate in JHS
Facilities	Ratio of Students per Schools PS
	Ratio of Students per Schools JHS
	Ratio of Students per Teachers PS
	Ratio of Students per Teachers JHS

Illiteracy, Expected Year School, and Mean Year School are variables that used for reflecting quality of education. Participation in education seen through Net Enrollment Rate and Drop Out Rate. For measuring educational facilities, Ratio of Student–Teacher and Ratio of Student-School will be used.

Statistical Approach

For clustering the district the following approaches are implemented:

Fuzzy C-Means (Bezdek, J. C., Robert, E., & William, F. (1984).

This is an improvement method of K-Means algorithm. This clustering technique permits a data item to belong to some clusters with a defined fuzzy membership grade. It is robust to extreme observation. The FCM algorithm minimize the following equation

$$J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\|x_j-v_i\|^2,$$

Where u_{ij} is cluster membership and $v_i=\frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m},\quad 1\leqslant i\leqslant c,$

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|^2}{\|x_j-v_k\|^2}\right)^{1/(m-1)}\right]^{-1},\quad 1\leqslant i\leqslant c,\ 1\leqslant j\leqslant n.$$

Fuzzy Geographically Weighted Clustering (Mason and Jacobson, 2007)

The education attainment of a region is affected by the neighboring region. To consider the basic spatial interaction model (distance between districts and population of each district) the FGWC is implemented. It adjusts the cluster membership for the fuzzy clustering algorithm discussed before by the weight based on the distance of districts.

The adjusted cluster membership for the fuzzy geographically weighted clustering algorithm is calculated in each iteration and is as follows

$$\mu'_i = \alpha \mu_i + \beta . \frac{1}{A} \sum_{j=1}^{n} w_{ij} \mu_j$$

where $w_{ij} = \frac{(m_i m_j)^b}{d_{ij}^a}$ $\quad \alpha + \beta = 1$

β = proportion of geographically effect

m= population area *i*

d=distance between areas

Biclustering

The clustering technique usually only cluster either the subjects or the variables. Biclustering is a method for simultaneous clustering of the rows and columns of a matrix which can provide a subset of subjects and variables with similar pattern. Madeira and Oliveira discussed several structures and approaches of biclustering.

In this study we use one of the first biclustering approaches proposed by Cheng and Church (2000). In this approach, a bicluster is defined as a subset of rows and a subset of columns with a high similarity score called mean squared residue (H). It was used as a measure of the coherence of the rows and columns in the bicluster. A submatrix is considered a bicluster if $H<\delta$ for some $\delta>0$. This is why the CC approach is also sometimes called delta (δ) biclustering. The algorithm CC biclustering is as follows:

Find biclusters with mean squared residue (H) < δ

Remove the row/col that reduces H the most

Add rows/cols that do not increase H

Stop when H < δ

Mask bicluster with random values

Repeat to find next bicluster

The biggest challenge in clustering is to check how good the clustering result is. The number of cluster need to define in advance, here obtaining the right number of cluster is crucial. For Cluster Validation Index Wang & Zhang (2007) proposed a Xie Beni Index which is a ratio of compactness and separation as defined below:

$$V_{XB} = \frac{J_m(u, v)/n}{Sep(v)} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \|x_j - v_i\|^2}{n \min_{i,j} \|v_i - v_j\|^2}$$

To perform and implement all methods mentioned before, the following R packages are used

advclust and RcmdrPlugin.FuzzyClust (Firmansyah, 2016)

spatialClust (Pamungkas, 2016)

Web based: Forum Analisis Statistik FAST (Pratama, 2016 and Dalimunthe et al 2014)

3. RESULT

The three approaches were performed for different number of clusters. The result from Xie Beni Index shows that three clusters is the best for FCM and FGWC.

Fuzzy C-Means

The result is visualized by mapping/chloropleth shown in Figure 1a. From that figure, it can be said that cluster 1 is an area that is located on Madura Island and some areas on East Java. Meanwhile cluster 2 is for districts in Southern part of West Java, and some district in Central and East Java. Cluster 3 consist district of DKI Jakarta, Yogyakarta and Banten.

Radar plot presented in Figure 1b indicated that Cluster 1 has low value on educational achievement, Cluster 3 is a cluster with the median and high quality education.

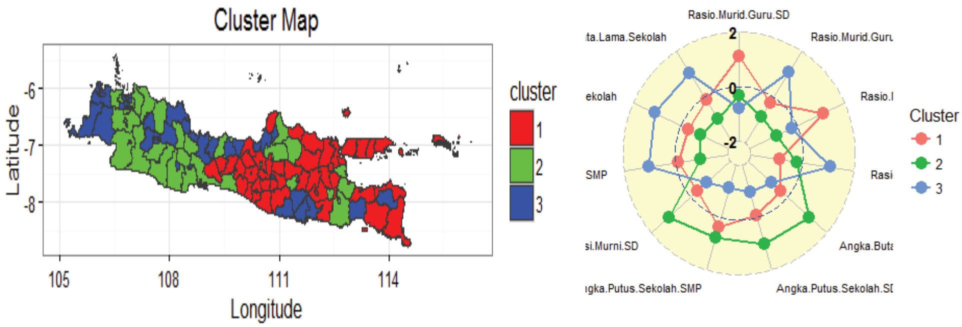


Figure 1. Result of Fuzzy C-Means Clustering

Fuzzy Geographically Weighted Clustering

Considering the basic spatial interaction model (distance between districts and population of each district) seem to change the clustering membership. As presented in Figure 2, some districts move from cluster 2 two cluster 1 such as Pati district. The interpretation of the cluster is the same as FCM approach.

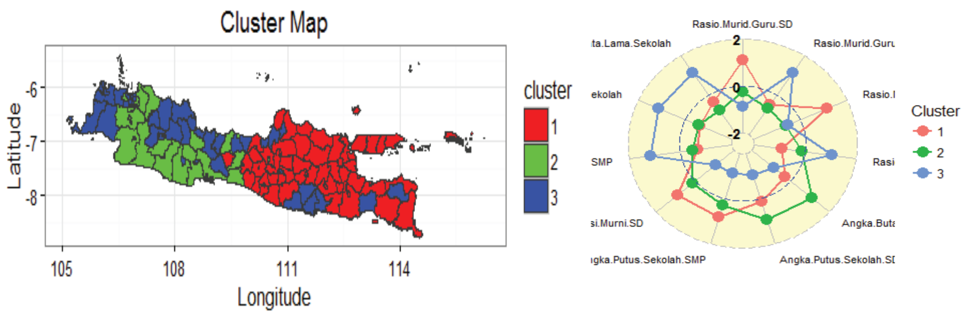


Figure 2. Result of Fuzzy Geographically Weighted Clustering

The results of the cluster along with the name of the areas that are included in each cluster are presented in Table 2. We observed that cluster 3 consists of capital cities which expected to have better education attainment.

Cheng and Church Biclustering

For CC Bicluetring, we obtained 5 biclusters which characteristics can be seen from the visualization of the results of analysis using bar chart presented in Figure 3. The variables that are relevant in each bicluster represented by the bar. The average of each variable from the data is shown by the red point.

Table 2. Highest indicators and the cluster memberships on each cluster

Cluster	High Indicators	Districts
1	Ratio of Students per Schools PS (222/119) Ratio of Students per Teachers PS (400/340) Dropout Rate in JHS (5.2 / 3.9) Net Enrollment Ratio Junior PS (97/96)	Blitar (city) Malang (city) Probolinggo (city) Pasuruan (city) Mojokerto (city) Madiun (city) Surabaya (city) Batu (city)
2	Dropout Rate in PS (20.2/9.6) illiteracy Rate (0.16/0.10)	Kediri (city) Cilacap Kebumen Pemalang Magelang (city) Salatiga (city) Pekalongan (city) Tegal (city)
3	Ratio of Students per Teachers JHS (17.8/14.8) Ratio of Students per Schools JHS (256/117) Expected Years of Schooling (13.1/12.4) Net Enrollment Ratio Junior High School (85/82) Average Years of Schooling (9.0/7.1)	Sukabumi (city) Bandung (city) Tasikmalaya (city) Banjar (city) DKI Surakarta (city) Semarang (city)

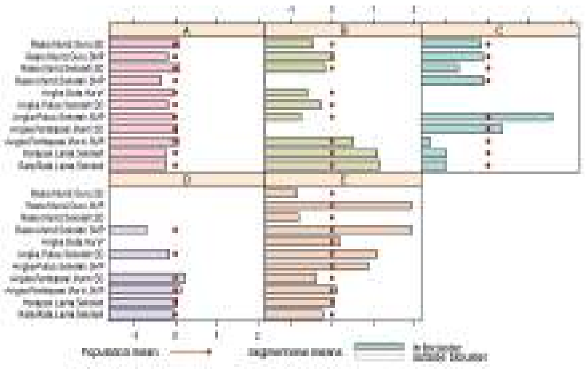


Figure 3. Result of Cheng and Church Biclustering

Bicluster A is a bicluster with a low educational level shown by all variables less than the overall median, though the drop out and illiteracy is small as well. Bicluster B is a cluster with high average years and expected years of schooling index and low drop out rate. Bicluster C is a cluster with very low educational level with highest dropout rate and the lowest in the participation on education. Bicluster E is a region with less number of teachers, high drop out rate and low participation index.

4. CONCLUSIONS

The revealed that there are at least three clusters observed in Java representing the education quality. Some districts are affected with their neighboring conditions. In general, district in "horse hooves" area and Madura Island are clustered in the cluster with low education attainment. The reason of this should be investigated further. Compared with cluster 1, cluster 2 is an area of the district / city with a better level of education. Therefore, the development of education can be focused on regions that grouped on cluster 1.

Cluster 3 is the best quality of education and dominated by urban region. Cluster 2 and 3 is cluster with the quality of education being but government in cluster 2 seriously to improve educational quality by allocating the highest budget for education. Cluster specific indicators can be used for focusing the education improvement program

REFERENCES

Madeira, S.C. and Oliveira, A.L. (2004). Biclustering algorithms for biological data analysis: a survey. *Journal IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* Volume 1 Issue 1.

Mason G.A., and Jacobson, R.D. (2007). Fuzzy Geographically Weighted Clustering. *Proceedings of the 9th International Conference on Geocomputation*, Maynooth, Ireland Cheng, Y. and Church. G.M. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology* (ISMB'00), pages 93–103, 2000

Bezdek, J. C., Robert, E., & William, F. (1984). FCM: The *Fuzzy cmeans*CLustering

Algorithm. *Computers & Geosciences vol.10*, 191-203

Wang, W., and Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems* 158, 2095-2117.

Dalimunthe, et.al. (2014). FAST: a Web-Based Statistical Analysis Forum. *Proceeding Islamic Countries Conference on Statistical Science 13,*, (pp. 185-200).

Pamungkas, I.H., (2016). Development of Fuzzy Geographically Weighted Clustering With Optimization Gravitational Search Algorithm on Forum Analisis Statistik (FAST). Skripsi. Sekolah Tinggi Ilmu Statistik. Jakarta.

Firmansyah, A.F.B., (2016). Application Cluster Ensemble and Analysis Of Fuzzy C-Means And On The R Package Gustafson Kessel. Skripsi. Sekolah Tinggi Ilmu Statistik. Jakarta.

Pratama, A.E., (2016). Development of Cheng and Church Biclustering on Forum Analisis Statistik (FAST). Skripsi. Sekolah Tinggi Ilmu Statistik. Jakarta.

The Comparison of Respondent Confidence in the Stimulation Card Evaluation on Three Methods in the Conjoint Analysis

Fitri Catur Lestari*
Sekolah Tinggi Ilmu Statistik, Jakarta, Indonesia – fitricaturlestari@stis.ac.id

Abstract

There were many types of method in conjoint analysis such as fullprofile design with ranking response (method A), fullprofile design with rating response (method B) and pairwise comparison with ranking response (method C). It was necessary to evaluate and compare them based on the result and the respondent time to determine their choice.

The purposive sampling method had been used to select 125 students as research samples. The consideration of samples selection was based on the segments that supposed as representative the population such as GPA, level/grade, high school origin, gender, and economic status (income) of the parents.

There were different result of three methods of conjoint analysis: respondent's preference of lecturer quality in method A and C were mastery of course material, ability to motivate students, character, communication and teaching method. While the method B, the sequence attributes that prefer were mastery of lecture, character, communication, ability to motivate students, and learning method.

By 5% significance level paired t test, the result were: there were significant duration (time) differences between three methods (p value= 0.000), the method A was a method with the longest time for respondents to finished the stimulation card evaluation and in contrast the method C was a method in the shortest time for respondents to finished it.

Keywords: Fullprofile design; Ranking response; Rating response; Pairwise comparison.

1. INTRODUCTION

The conjoint analysis is a part of multivariate analysis. The conjoint analysis concern in preference study. Chaplin defined preference as the attitude that prefer something to other things (Chaplin, 2002). In the conjoint analysis there are varied methods. It rely on the attributes combination design (stimuli), data collection and analysis procedure. Many types of methods in conjoint analysis are such as fullprofile design with ranking response (method A), fullprofile design with rating response (method B) and pairwise comparison with ranking response (method C). It is necessary to evaluate and compare them based on the result and the respondent time to determine their choice (represented by the stimuli cards. The stimuli cards differ conjoint analysis with the other analysis).

The conjoint analysis can be applied in education field. The lecturer that is a main component in the university have an important role in determining the quality of the university (Sudiana, 2003). The lecturer quality significantly (sig. 0,01) influences the students learning motivation and has the highest correlation compared with other factors (Pujadi, 2007). The students that were directly involved with the lecturer in the learnig proccess surely have preference on the lecturer quality. Therefore, the students preference on the lecturer quality is neccessary to be investigated by conjoint analysis.

The objectives of this research are to know student preference of the lecturer quality on STIS by three methods of conjoint analysis and to know respondent time/duration to evaluate the stimuli cards by three methods of conjoint analysis.

2. METHODS

Three methods of conjoint analysis are used to know student preference of the lecturer quality on STIS. While to compare the respondent time/duration to evaluate the stimuli cards, this research used t test.

Consecutively five steps in the conjoint analysis are:

A. Formulate The Problem

The first step in conjoint analysis is formulate the problem (Aaker DA, 1980). Formulating the problem start from defining the product as a set of attributes that consist of several level. The information about attributes that represent the consumer preference can be gained from expert discussion, secondary data exploration or pre-survey (Rosada, 2002).

The lecturer quality is formed in attributes that consist of several levels. The attributes of lecturer quality in this research are determined by pre-survey. The result of the pre-survey are:

Table 1. The Attributes that Influence the Student Preference of Lecturer Quality

No.	Attribute	Percentage
1	lecture delivery	24
2	learning method	20
3	lecture mastery	19
4	personality	11.5
5	ability to motivate the student	9
6	other (8 attributes)	16.5

Based on table 1, there are 5 (five) attributes that have large percentage between 9,00% sampai 24,00%. Each attribute has 2 levels: lecture delivery (two-way communication and one-way communication), learning method (creative and not creative), lecture mastery (mastery and not mastery), personality (humorist and serious), and ability to motivate the student (motivate the students and do not motivate the students).

B. Design The Stimuli

The combination design used in this research is full profile or large factor evaluation (Sawtooth, 2001). Because there are 5 (five) attributes, there are 25 or 32 stimulation cards (that contain many characteristics of the lecturer specially exact field lecturer). Method A dan B are defined as conjoint analysis with fullprofile design.

Beside full profile design, pairwise design (Sawtooth, 2001) also used in this research and defined as method C. Because there are 5 attributes, the simulation cards that evaluated by the students or respondents are 5(5-1)/2 or 10 cards (that also contain many characteristics of the lecturer specially exact field lecturer). In this step, the questionnaire was made in which questions that are relevant with demography variable to know the characteristic of the students.

C. Determine the Method of Data Collection

The survey used primary data by interview and simulation cards evaluation that involved 125 students as respondents. They drawn by purposive sampling method with sample selection was based on the equalization in segments that allegedly there are different preference of the students on the quality of lecturers ie GPA, grade, high school origin, gender, and parent economic capacity (parent income). The evaluation response is rank response. The purpossive sampling method is one of sampling methods that is included in non-probability sampling by collecting the information from appropriate sources that are supposed have ability to give information that is needed or they are the only sources who can give the information that is needed by us (Wibisono, 2003). In the next writing, mentioning "students" means "respondents" because of the nonprobability sampling do not conclude for the population. The research was held in STIS on Wednesday-Thusrday, July 20th-21th, 2011.

D. Choose the Procedure in Conjoint Analysis

Analysis procedure used in this research is regression method with dummy variables. Generally the base model of conjoint analysis (Kuhfeld, 2000) with that procedure:

$$Y_{ij} = \beta_0 + \sum_{i=1}^{m}\sum_{j=1}^{k} \beta_{ij} X_{ij} + \varepsilon_{ik}$$

Y_{ij} = Rank all respondents
β_o = Intercept
k = Number of levels in i-th attribute
m = Number of attributes
X_{ij} = Dummy variable in i-th attribute j-th level
β_{ij} = Part worth i-th attribute j-th level
ε_{ij} = Error

By regression model, part worth of the levels in each attribute (NKT) can be determined and are used to determine importance value of the level compared by other level in an attribute. After this proccess, Important Relative Value (NRP) can be counted by formula:

$$NRP = \frac{UT_i - UR_i}{\sum_{i=1}^{k} (UT_i - UR_i)}$$

NRPi = NRP of i-th attribute
UTi = the highest NKT of i-th attribute
URi = the lowest NKT of i-th attribute
k = Number of attributes

Thurstone procedure used to describe about pairwise of the attributes by pairwise comparison (Rosada, 2002). The steps are:

1. Count the frequency matrices by adding the score of all observations by this rules:

   $$F_{ij} = \begin{cases} 1, \text{ if attribute i>attribute j} \\ 0, \text{ if attribute i<attribute j} \\ 0.5, \text{ if attribute i=attribute j} \end{cases}$$

   Fij =frequency of i-th row and j-th column (column more important than row)
2. Count the proportion matrices (P_{ij}) by dividing each component of frequency matrices with the number of respondent.

   $$P_{ij} = \frac{F_{ij}}{n}$$

3. Transform the proportion matrices component to standard normal (Zij).
4. Order the column of Z matric from column with smallest mean to greatest one.
5. Count the nearest columns difference.
6. Count scale value with first cale value is zero and the next scale value is the cummulative of the former scale value.
7. Conclude the important factors

The statistic method that is used to compare the respondent time/duration to evaluate the stimuli cards is t test (Walpole, 1992).

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

d = average of samples data difference in 2 different treatment
μ_d = average of population data difference in 2 different treatment
s_d = standard deviation of samples data difference in 2 different treatment
n = number of samples/observation

The data preparation and analysis were proccessed by software: SPSS 13.0, Microsoft Excel 2003, dan SAS 9.1.

3. RESULT

By method A and C, the student preference on the lecturer quality based on important relative value (NRP) and scale value (NS) consecutively from prefered attribute until unprefered attribute are:

1. lecture mastery (NRP=31.05%, NS=0.00)
2. ability to motivate the students (NRP=21.01%, NS=0.32)
3. personality (NRP=20.29%, NS=0.60)
4. lecture delivery (NRP=17.44%, NS=0.72)
5. learning method (NRP=10.21%, NS=0.86)

By method B, the attribute that have the highest influence to student preference is lecture mastery with NRP = 32.79%. The next attribute is personality, lecture delivery, ability to motivate the students and learning method with NRP in sequence 17.82%, 17.20%, 16.75% dan 15.44%. The lecturer who is liked most by the students is the smart lecturer (mastery in his/her field) with highest NKT 2.15. Consecutively the type of lecturer: humorist (NKT=1.17), communicative or two-way communication (NKT=1.13), able to motivate the students (NKT=1.10), and creative in learning method (NKT=1.01).

The student preference of the lecture quality is varied on the segments: GPA, grade, high school origin,

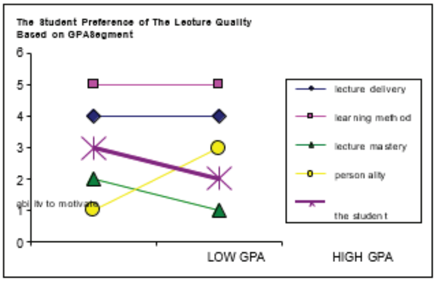


Figure 1. The Student Preference of The Lecturer Quality Based on GPA Segment

By Figure 1, the high GPA students prefer smart lecturer (mastery in his/her field). While students who have the lower GPA like the humorist lecturer. The high GPA student, after mastery attribute, like motivating lecturer and then humorist lecturer. While the low GPA student, after personality attribute, like smart lecturer and then motivating lecturer. Both of them place the lecture delivery attribute as 4th important attribute and learning method as the last important attribute.

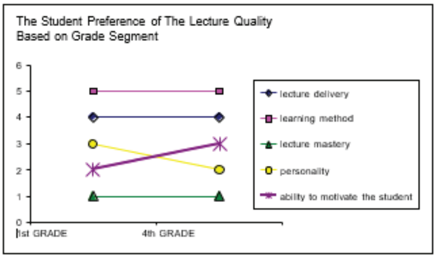


Figure 2. The Student Preference of The Lecturer Quality Based on Grade Segment

The student preference on lecturer quality is different in the 2nd and 3rd attribute (Figure 2). The first level students prefer lecturer who motivates the students than humorist lecturer. In contrast, the fourth level students prefer the humorist lecturer rather than the motivating lecturer. It is assumed because the first grade students need to be motivated at the first learning process in the university and the fourth grade students need a relaxation in the class by humorist lecturer on their last learning process in the university.

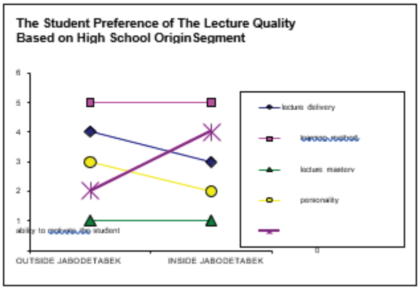


Figure 3. The Student Preference of The Lecturer Quality Based on High School Origin Segment

From figure 3, it could be describe that the students from outside Jabodetabek (Jakarta Bogor Tangerang Bekasi) prefer motivating lecturer, humorist lecturer and communicative lecturer but the students from Jabodetabek prefer humorist lecturer, communicative lecturer and motivating lecturer.

By presupposition, the students from outside Jabodetabek need the motivation because they live far from their family. The students from Jabodetabek live in urban city so they need the humorist lecturer in the class or learning proccess to minimalize the pressure of their environment.

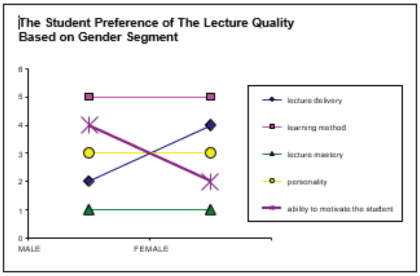


Figure 4. The Student Preference of The Lecturer Quality Based on Gender Segment

The female students prefer motivating lecturer to communicative lecturer while the male ones prefer lecturer who deliver the lecture by two way communication or communicative lecturer to motivating lecturer (Figure 4). It might caused that female students are dominated by their feeling and need motivation from the lecturer than the male one that are dominated by their mind or logic.

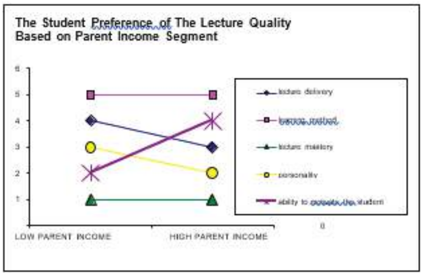


Figure 5. The Student Preference of The Lecturer Quality Based on Parent Income Segment

There are different preference by parent income segment like the previous segments. The students whose parent have the low income prefer consecutively motivating, humorist and communicative lecturer. Contrarily the students whose parent have the high income prefer consecutively humorist, communicative and motivating lecturer. It assumed that in the low parent economic status, students need motivation from the lecturer rather than in the high one. In the high parent economic status, students want to fullfill their tertiary need such as joke, entertainment, comfort situation, etc.

The next result is the comparison of three methods conjoint analysis based on respondent time to determine the choice of stimuli cards.

Table 2. Description of Time in Three Methods (minutes)

Method	Mean	Number of Samples	Standard Deviation	Standard Error Mean
A	2.71	125	3.7219	0.3329
B	2.04	125	1.4629	0.1308
C	1.25	125	0.7715	0.069

Table 2 shows that the method A is a method with the longest time (average= 2.71 minutes) for students to finished the stimuli card evaluation and in contrast, the method C is a method in the shortest time (average= 1.25 minutes) for students to finished it. In method A, the evaluation time has the largest variance (square of standard deviation) while in method C, it has the smallest one.

Table 3. Time Difference Test

Paired Method	t-value	Degree of freedom	p-value
A and B	14.556	124	0.000
A and C	24.893	124	0.000
B and C	19.346	124	0.000

Based on table 3, the conclusion is there is time (for students to finished the stimuli card evaluation) difference in three methods in conjoint analysis because all p-values are lower than alpha 5%. By 5% significance level, there is significantly duration (time) differences between three methods, method A is a method with the longest time for students to finished the stimuli card evaluation, in contrast, the method C is a method in the shortest time for students to finished it.

4. CONCLUSIONS

There were different result of three methods of conjoint analysis: respondent's preference of lecturer quality in method A and C were mastery of course material, ability to motivate students, character, communication and teaching method. While the method B, the sequence attributes that prefer were mastery of lecture, character, communication, ability to motivate students, and learning method. By 5% significance level paired t test, the result were: there were significant duration (time) differences between three methods (p value= 0.000), the method A was a method with the longest time for respondents to finished the stimulation card evaluation and in contrast the method C was a method in the shortest time for respondents to finished it.

REFERENCES

Aaker DA, Day GS. (1980). Marketing Research. New York : John Willey&Son, Inc.

Chaplin, JP. (2002). Kamus Lengkap Psikologi. Edisi kelima. Terjemahan: dr. Kartini Kartono.

Jakarta: PT. Raja Grafindo Persada. http://www.sawtooth.com, Conjoint Analysis, (2001).

Kuhfeld, WF. (2000). "Conjoint Analysis Examples", SAS Institut, Inc. http://www. sawtooth software. com. [28 Januari 2004]

Pujadi, A. (2007). Faktor-Faktor yang Mempengaruhi Motivasi Belajar Mahasiswa: Studi Kasus pada

Fakultas Ekonomi Universitas Bunda Mulia. Business & Management Journal Bunda Mulia, Vol: 3, No. 2.

Rosada, R. (2002). Perbandingan Metode Pairwise Comparison dan Full profile dalam Pengumpulan Data untuk Analisis Konjoin. Skripsi. Jurusan Statistika IPB. Bogor.

Sudiana, IN. (2003). Peranan Profesionalisme Tenaga Pengajar (Dosen) terhadap Proses Pembelajaran di Perguruan Tinggi. Jurnal Pendidikan dan Pengajaran IKIP Negeri Singaraja, No. 4 TH. XXXVI.

Walpole, RE. (1992). Pengantar Statistika. Edisi ke-3. Jakarta: PT. Gramedia Pustaka Utama. Wibisono, D. (2003). Riset Bisnis. Jakarta : PT. Gramedia Pustaka Utama.

This page intentionally left blank

Plenary Seasson 03 (PS03):
Emerging Challenges In Data Collection, Survey Methodology And Implications For Official Statistics

This page is intentionally left blank

Modernised Business Process and Some Challenges in BPS Statistics Indonesia

Heru Margono
Director of Survey and Census Methodology Development, Statistics Indonesia

STATISTICS INDONESIA

SENSUS EKONOMI

The Agent of Trustworthy Statistical Data for All

Modernised Business Process and Some Challenges in BPS Statistics Indonesia

Emerging Challenges in Data Collection, Survey Methodology, and Implication for Official Statistics

March 2017

Heru Margono, BPS - Statistics Indonesia

Outline

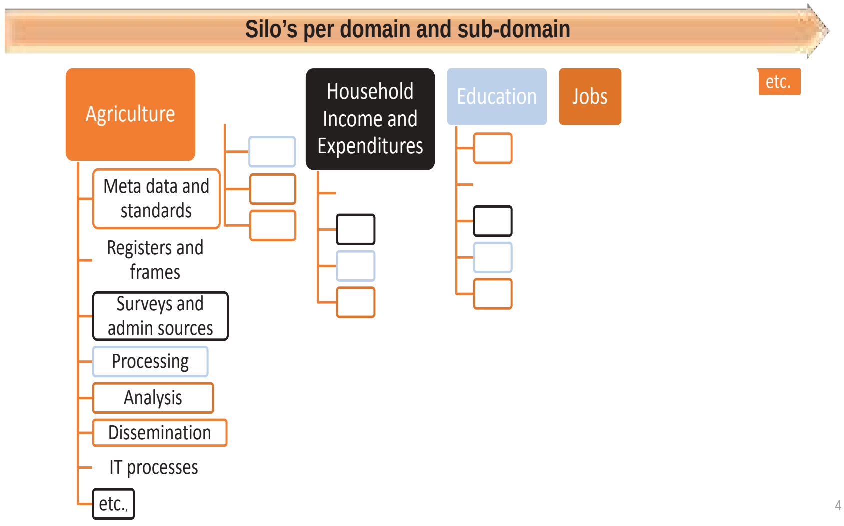
- ❑ BPS Commitments on Modernization
- ❑ Generic Statistical Business Process Model
- ❑ BPS Statistical Business Framework and Architecture
- ❑ Some Challenges
- ❑ Integrated Statistics Approach

2

Drivers of Modernization Programme

- Data needs had been increasing both in frequency and variety
- Statistical production has been traditionally organised by topic, e.g. transport, trade, industry, social, etc → organisational silos
- Financial pressures are encouraging new ways of thinking
- International best practices → Some statistical organisations are moving towards a process-based approach
- Wealth of information, including unstructured information (innovative sources, e.g. Big Data)
- Availability of new methodological and technological tools
- Drawbacks of traditional data collection systems (high costs, response burden, lower response rates)

Traditional approach



4

Pros & Cons

- Specialist approach
- Few dependencies
- Sub optimal allocation of capacities
- Specific IT-applications
- Lack of coordination
- Lack of flexibility
- Human factor

5

BPS Commitments on Modernization

- ❑ BPS has committed to conduct a fundamental, comprehensive, and cross-sectoral change and reform program since 2008.
- ❑ As one of government agency, BPS has also committed to follow National Bureaucratic Reform Program.
- ❑ STATCAP-CERDAS (Statistical Capacity Building-Change and Reform of The Development of Statistics) and Bureaucratic Reform Program is one concept of change and reform agenda, whereas the main goal of Bureaucratic Reform Program is to achieve Good Governance, while Statcap-Cerdas is the boosters for Bureaucratic Reform Program.
- ❑ STATCAP-CERDAS integrate :
  - ❑ 1. Internal : employees, subject matters
  - ❑ 2. External : stakeholder and data users towards 'One Data' and 'Big Data'.

6

BPS Will Achieve Modernized Process By:

- ❑ Integration of statistical production
- ❑ International Standards
- ❑ Better acquisition and use of technology
- ❑ Innovative data collection and capture modes
- ❑ Common approaches to data management
- ❑ Using administrative data
- ❑ Rationalising the survey programme
- ❑ Strengthened competencies of BPS staff
- ❑ Provincial and District Offices will remain strong but well-coordinated from the centre
- ❑ Strong relationships with users

01/02/2018

How to modernized?

Defining and modelling statistical processes based on

Generic Statistical Business Process Model (GSBPM)

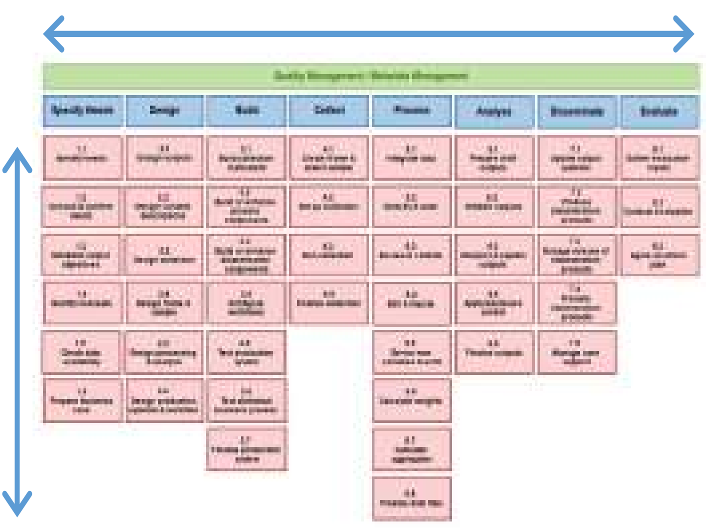
Why do we need a model?

- To define and describe statistical processes in a coherent way
- To standardize process terminology
- To compare / benchmark processes within and between organisations
- To identify synergies between processes
- To inform decisions on systems architectures and organisation of resources

Structure of the GSBPM

Process Phases (1st level)

Sub-processes (2nd level)



The Generic Statistical Business Process Model (GSBPM)

- The GSBPM is a flexible tool to describe and define the set of business processes needed to produce official statistics.
- The GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs.
- It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources

01/02/2018

BPS – Business Process Model

4. COLLECT							
4.1 Create frame and select sample		4.2 Set up collection		4.3 Run collection		4.4 Finalise collection	
4.1.1 Prepare sampling frame	S, P	4.2.1 Arrange the collection field force	S, A, N, P	4.3.1 Send out pre-collection information	S, A, P	4.4.1 Data Entry/ Record the data	S, A, N, P
4.1.2 Select sample	S, P	4.2.2 Train collection staff	S, A, N, P	4.3.2 Collect/receive data	S, A, N, P	4.4.2 Analyze the process metadata	S, A, N, P
		4.2.3 Prepare collection resources and instrument	S, A, N, Z, P	4.3.3 Monitor and report data collection progress	S, A, N, P	4.4.3 Raw Data approval release for further processing	S, A, N, P
		4.2.5 Prepare the quality assurance resources	S, A, N, Z, P	4.3.4 Non Responds Follow Up	S, P	4.4.4 Archive collection instruments	S, A, N, P
		4.2.6 Prepare the data management protocols	S, A, N, Z, P				

1st level

2nd level

3rd level

Legend :
S : Survey
A : Administrative Data
N : National Account
Z : Analyze
P : Price

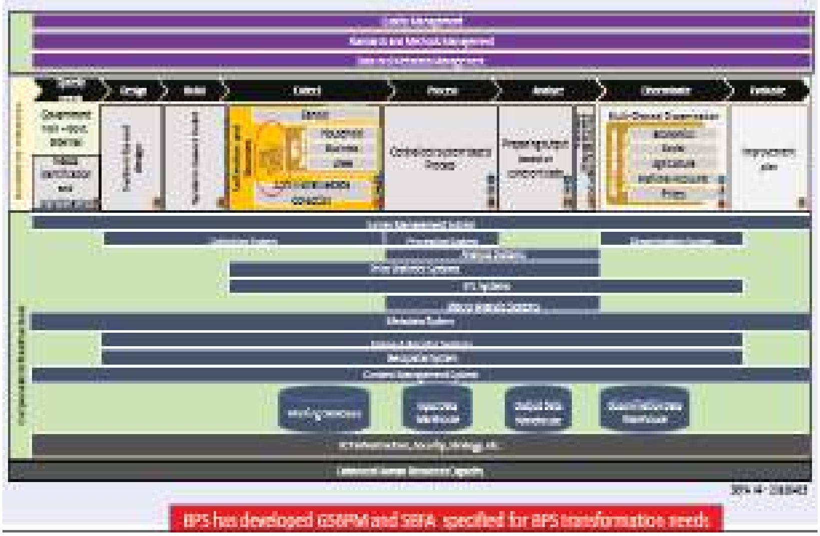
The Statistical Business Framework and Architecture (SBFA)

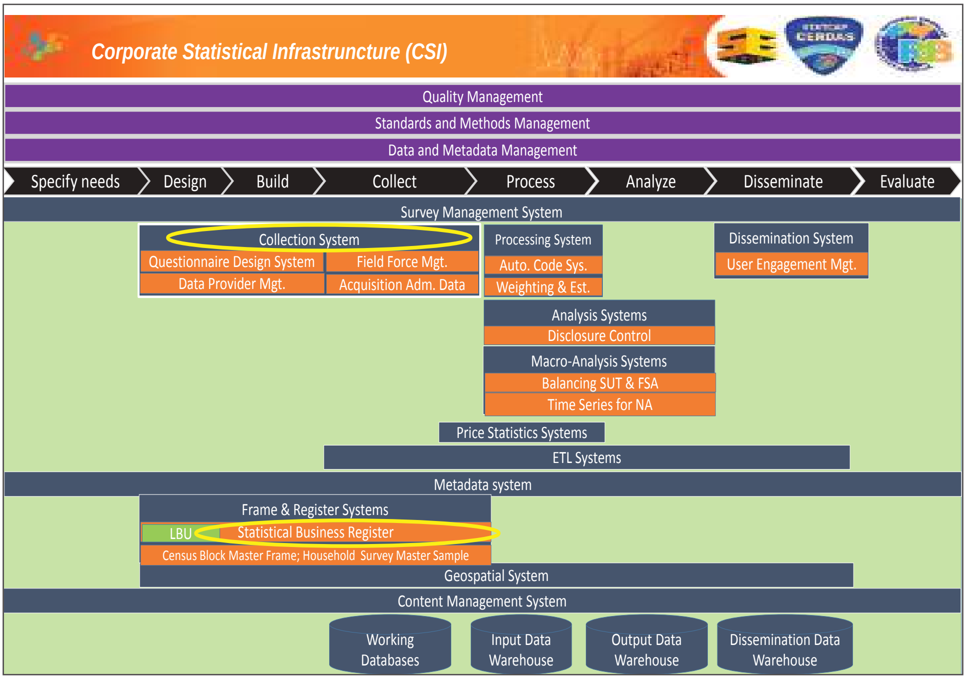
SBFA provides a blueprint as to how to structure BPS's processes and ICT infrastructure

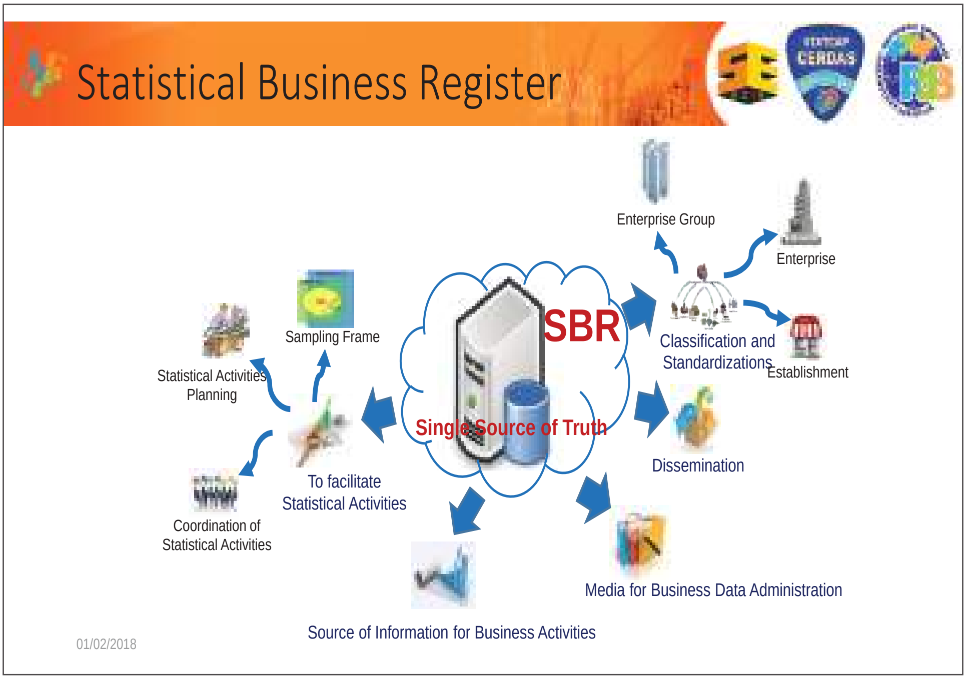
- ❑ SBFA plays as the main instrument in BPS modernisation programme
- ❑ It covers both statistical activities and strategic organisational tasks and capabilities
- ❑ It provides the underlying structure for the modernisation programme, guidance to statistical design and redesign projects, and assurance that a consistent approach will be used now and in future.
- ❑ In BPS's case, it provides the underlying structure for the BPS modernisation programme, with the overall aim of improving quality of BPS statistical products.
- ❑ The SBFA builds on the BPS Strategic Plan and the BPS mission statement, and formulates a vision for BPS until 2025
- ❑ It is composed by a GSBPM, a business activity process flow, a set of Principles and common and shared Infrastructures.

01/02/2018

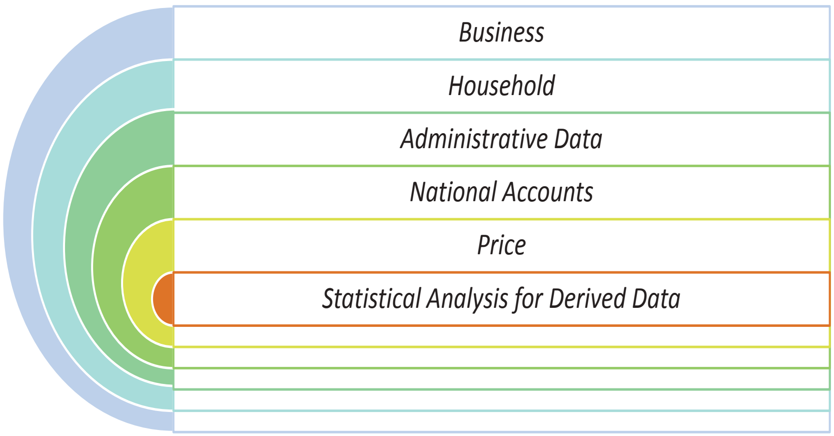
Statistical Business Framework Architecture (SBFA) of BPS




Corporate Statistical Infrastruncture (CSI)
Quality Management
Standards and Methods Management
Data and Metadata Management
Specify needs
Design
Build
Collect
Process
Analyze
Disseminate
Evaluate
Survey Management System
Collection System
Questionnaire Design System
Field Force Mgt.
Data Provider Mgt.
Acquisition Adm. Data
Processing System
Auto. Code Sys.
Weighting & Est.
Dissemination System
User Engagement Mgt.
Analysis Systems
Disclosure Control
Macro-Analysis Systems
Balancing SUT & FSA
Time Series for NA
Price Statistics Systems
ETL Systems
Metadata system
Frame & Register Systems
LBU
Statistical Business Register
Census Block Master Frame; Household Survey Master Sample
Geospatial System
Content Management System
Working Databases
Input Data Warehouse
Output Data Warehouse
Dissemination Data Warehouse


Statistical Business Register
Enterprise Group
Enterprise
SBR
Classification and Standardizations
Establishment
Sampling Frame
Statistical Activities Planning
Single Source of Truth
Dissemination
To facilitate Statistical Activities
Coordination of Statistical Activities
Media for Business Data Administration
Source of Information for Business Activities
01/02/2018

Business Process by Product Classes



17

Organisation and Governance at BPS

Eight Key Principles

1. National Accounts and Macro Statistics are the primary beneficiaries of statistical integration
2. Each survey serves the needs of multiple SMAs
3. Preference for registration and administrative data over surveys
4. Integrated survey operations from identify needs to dissemination and evaluation
5. Integration of economic statistic infrastructure through Statistical Business Register
6. Building an integrated IT landscape
7. Use technology based/paperless data collection
8. Use continuous surveying to spread the workload over a year

Centralized Corporate Support Services

Consolidation of the BPS cross-cutting Support Services :

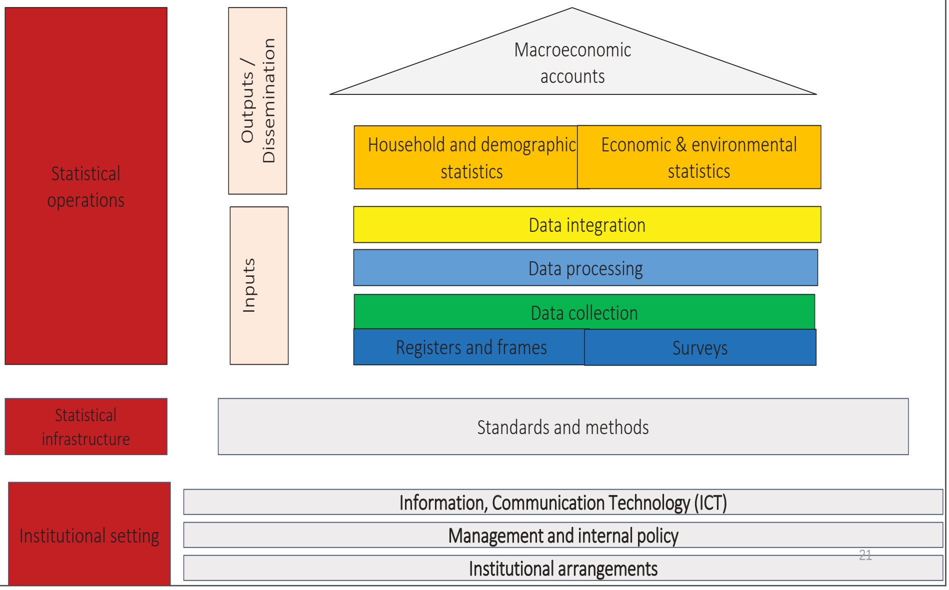
- technical, scientific and administrative methodology
  - Standard and classification
  - Statistical Business Registers, etc
- information technology
- data collection
- dissemination
- human resources

Some Challenges

1. Fast technologic developments
2. Increasing internationalization
3. Rapid growing availability of new data
4. Sharp increase in rate at which data become available
5. Growing demand for more (& quicker) information
6. Decreasing budgets (& improving cost efficiency)
7. IT Landscape

20

Integrated statistics approach



Benefits

- **Statistical business and information architecture** governs common statistical production process and centralized statistical services over time and across countries.
- Corporate, centralized services allow for **statistical professionalization, project management and coordination.**
- **Meet policy demands**: covering business and household statistics, labor statistics, short term statistics, national accounts and international statistics.
- **Cost effectiveness.**
- **Improved quality:** coordinated output; reduction of human factor; improved reproducibility.
- **Reduction of response burden** on business and household respondents.
- Offer **collaboration** in the development and application of common methods and IT tools.
- **Robust** and **flexible** and a **stable platform** for facing new developments.

22

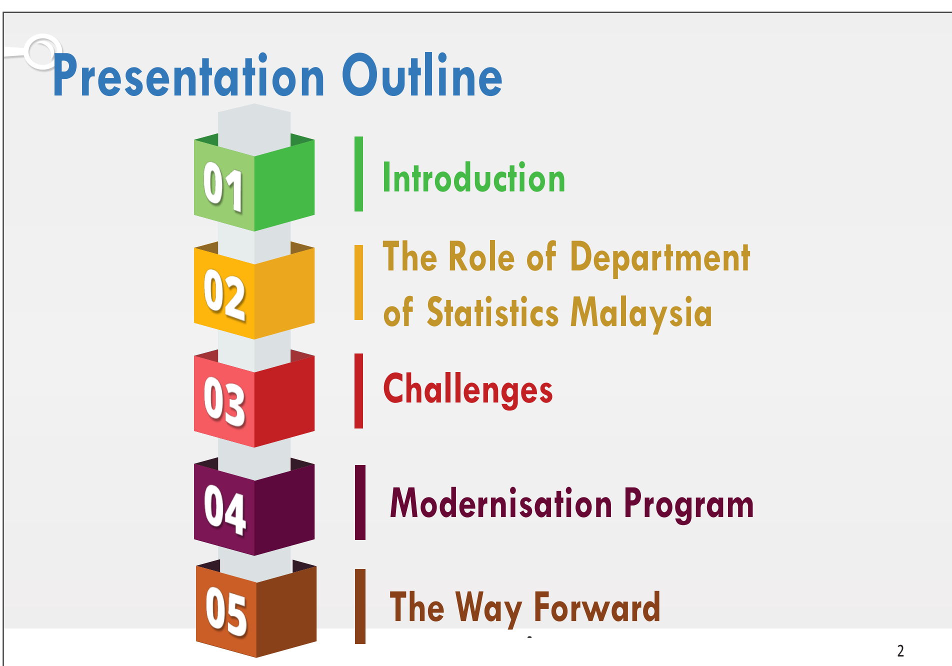

STATISTICS INDONESIA
The Agent of Trustworthy Statistical Data for All
THANK YOU
March 2017
CERDAS
RB

This page is intentionally left blank

Emerging Challenges in Data Collection, Survey Methodology, and Implication for Official Statistics

Norhayati Shuja'
Assistant Director of the Economic Indicator Division, Department of Statistics, Bank Negara Malaysia


International Statistical Institute
Regional Statistics Conference
ISI Regional Statistical Conference 2017
EMERGING CHALLENGES IN DATA COLLECTION,
SURVEY METHODOLOGY AND IMPLICATIONS
FOR OFFICIAL STATISTICS
22-24 March 2017
Bali International Convention Centre
Department of Statistics Malaysia
The Source of Malaysia's Official Statistics


Presentation Outline
01 Introduction
02 The Role of Department of Statistics Malaysia
03 Challenges
04 Modernisation Program
05 The Way Forward
2

Introduction

- The globalisation phenomenon has led the countries worldwide to face **new challenges.**
- The competitive and dynamic socio-economic environment resulted to a more complex expectations among the statistical community.
- Therefore, the needs for dynamic and complex statistics require the Department of Statistics Malaysia to further develop strategies in adapting the current environment changes.

3

Role of Department of Statistics, Malaysia (DOSM)

General
- Compilation of Statistics
- Analysis/ Interpretation
- Dissemination/ Statistics Communication

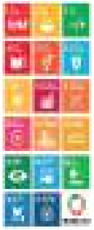
In SDGs (DOSM & Stakeholders)
- Coordinate
- Measure Development Progress
- Effectively Participation in National Monitoring Programme

The Involvement of Department of Statistics Malaysia

MDGs Millennium Development Goals

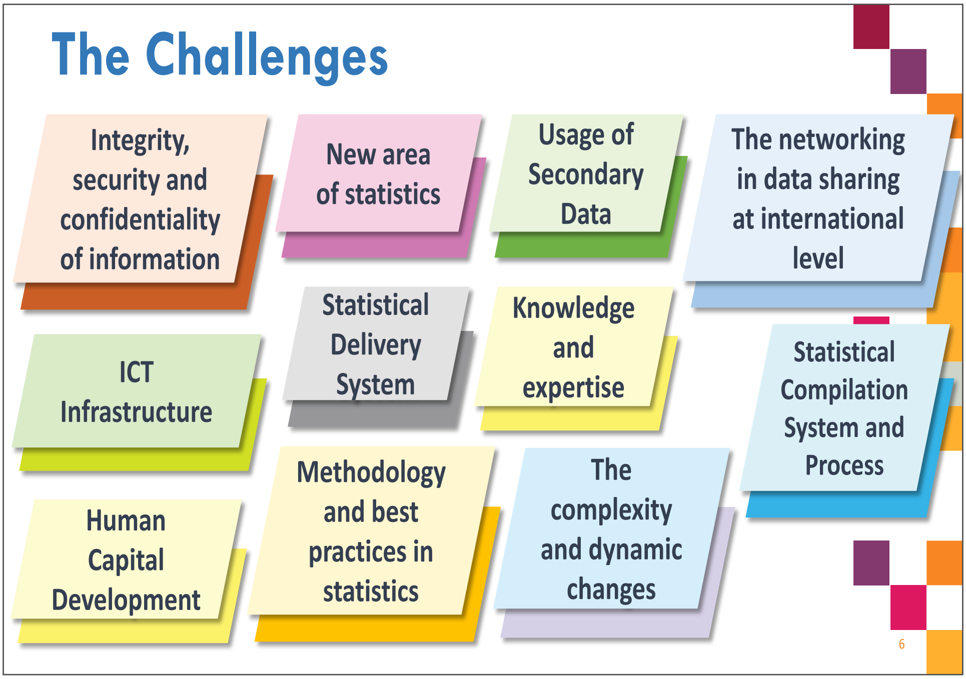
- Period : 2000 – 2015
- Involvement : Since 2007- 2015
- Focal Point : The coordination of the compiling MDGs indicators
- Member in Technical Working Groups (TWG) for providing MDGs reports at national, ASEAN and International level

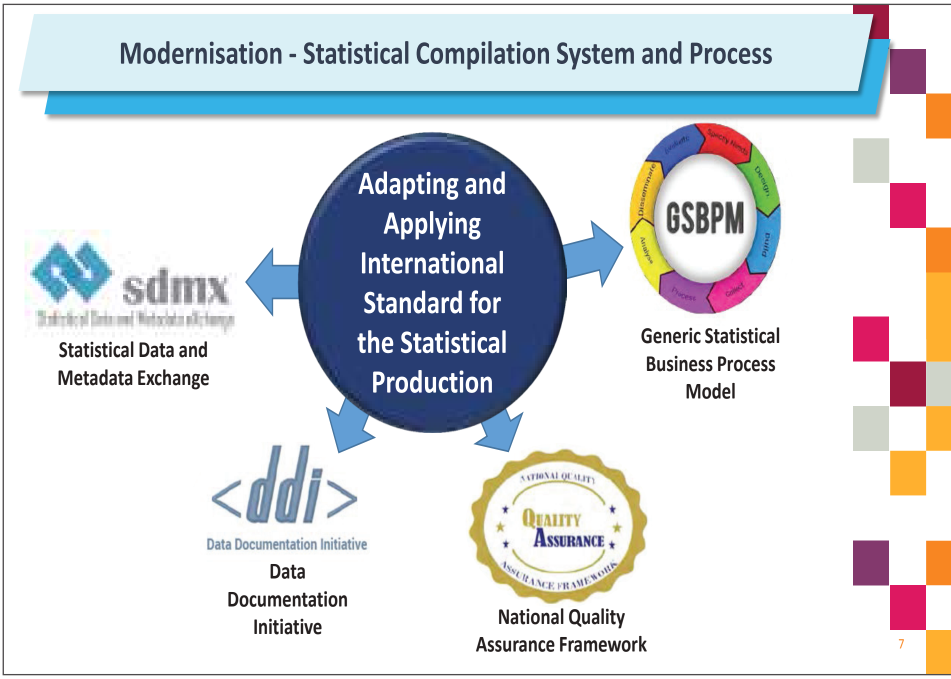
SDGs Sustainable Development Goals

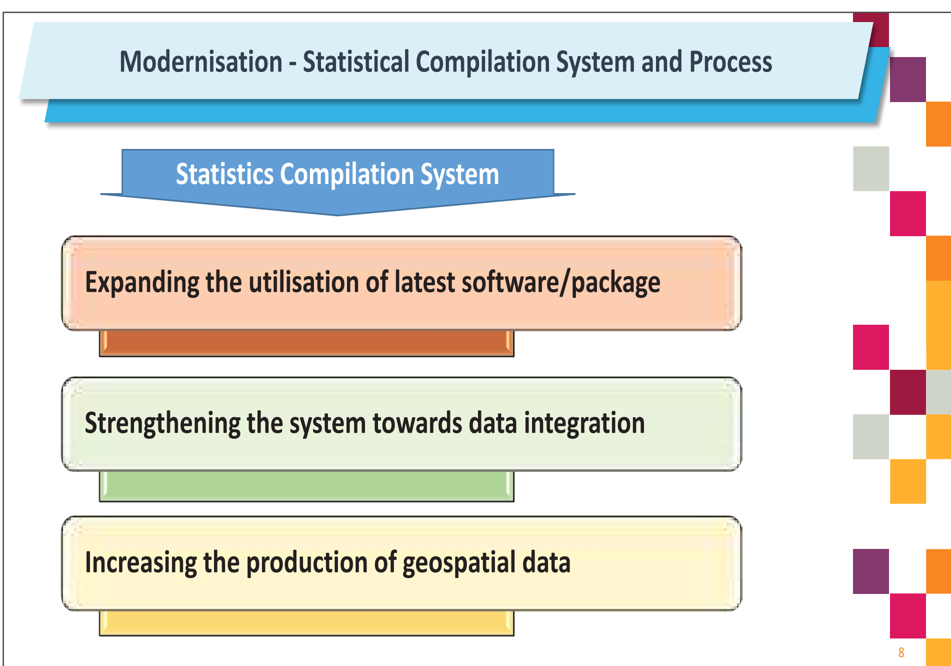


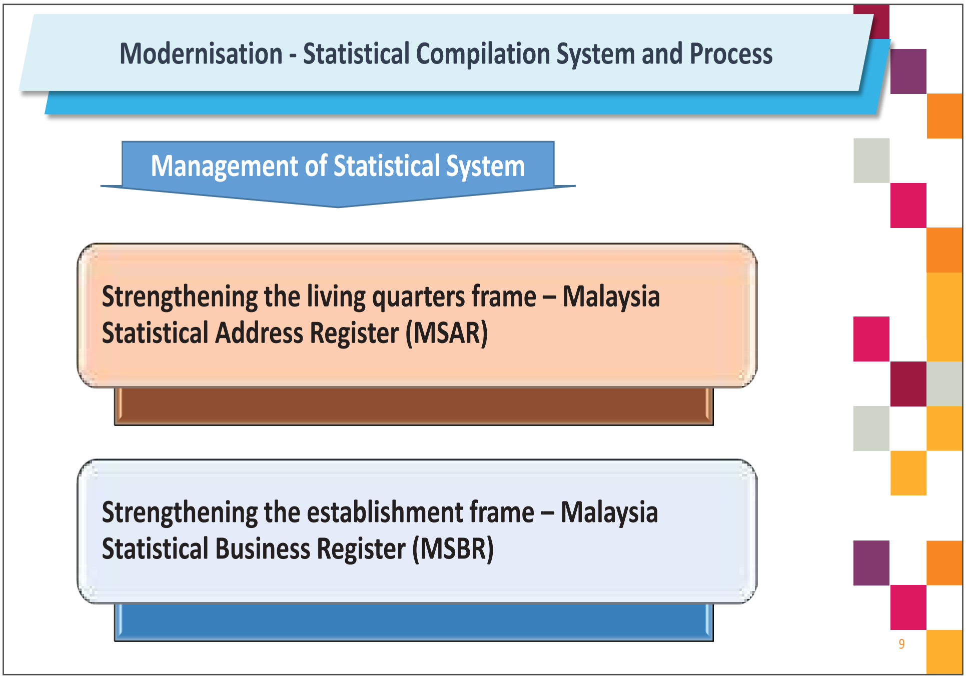
- Period : 2016 – 2030
- Involvement : April 2015
- Focal Point : The coordination of the development of SDGs indicators
- Appointed as representative of South-Eastern Asia in High Level Groups (HLG) for SDGs
- Provide inputs on assessment of SDGs indicators to Philippines (IAEG-SDGs representative of South-Eastern Asia) in the development of SDGs indicators

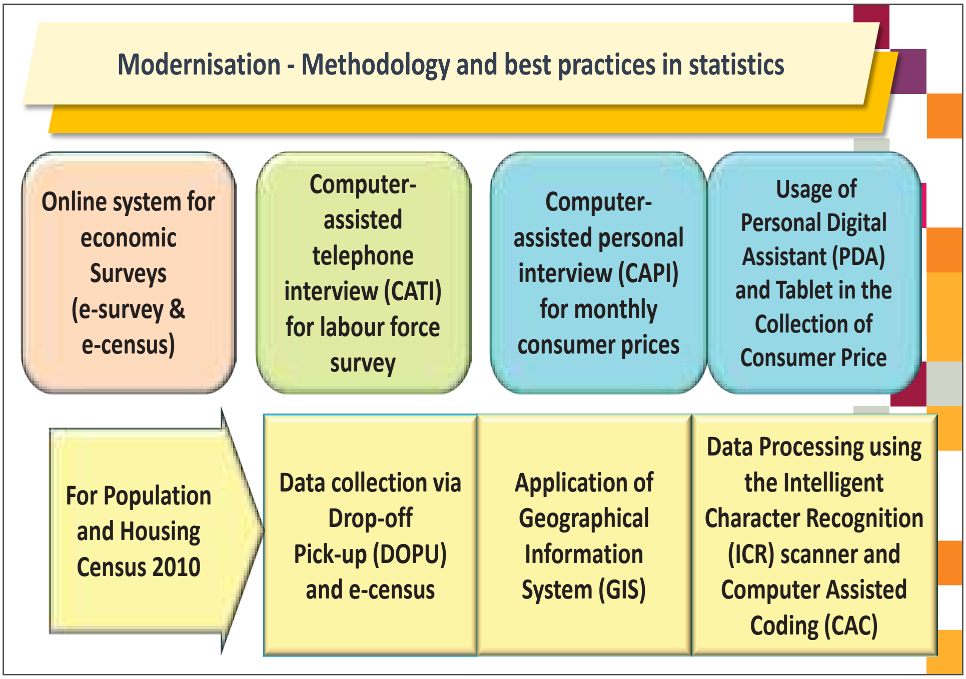

SDGs
SDGs is a set of universally applicable goals that balances the three dimensions of sustainable development:
environmental, social & economy
17 GOALS
169 TARGETS
241 INDICATORS
56% 134 Indicators Social
24% 58 Indicators Environment
20% 49 Indicators Economy
Source: United Nation

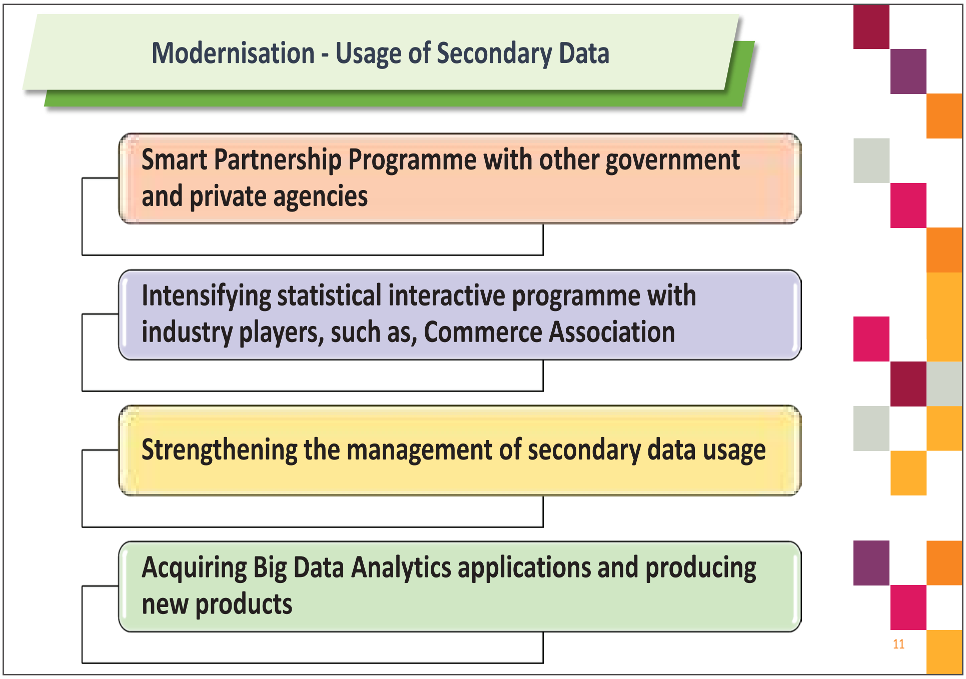

The Challenges
Integrity, security and confidentiality of information
New area of statistics
Usage of Secondary Data
The networking in data sharing at international level
Statistical Delivery System
Knowledge and expertise
ICT Infrastructure
Statistical Compilation System and Process
Methodology and best practices in statistics
The complexity and dynamic changes
Human Capital Development
6

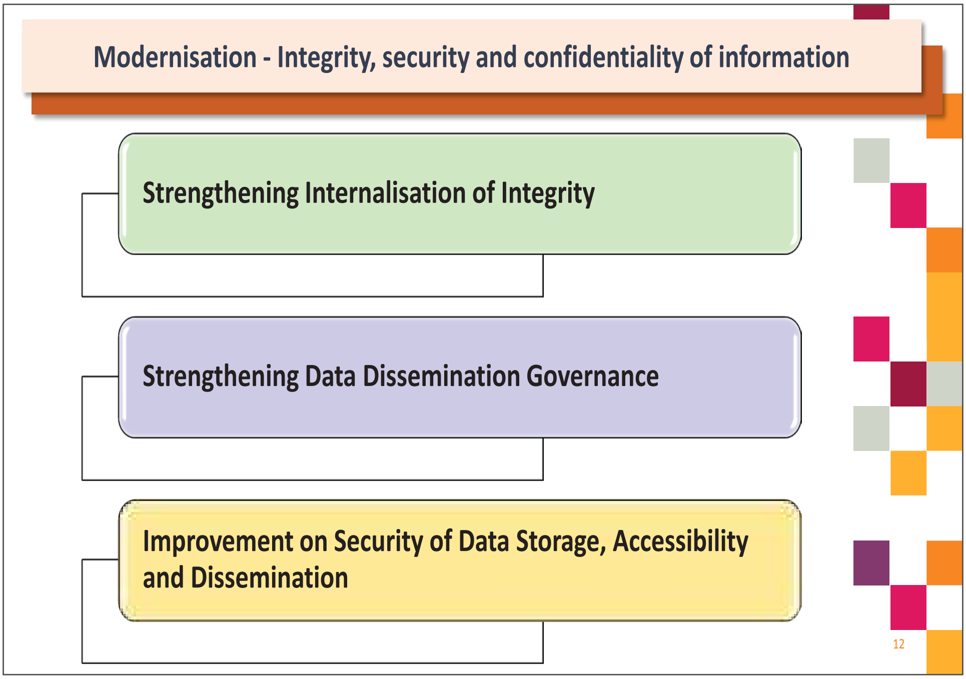

Modernisation - Statistical Compilation System and Process
Adapting and Applying International Standard for the Statistical Production
sdmx
Statistical Data and Metadata Exchange
GSBPM
Generic Statistical Business Process Model
<ddi>
Data Documentation Initiative
Data Documentation Initiative
Quality Assurance
National Quality Assurance Framework
7


Modernisation - Statistical Compilation System and Process
Statistics Compilation System
Expanding the utilisation of latest software/package
Strengthening the system towards data integration
Increasing the production of geospatial data
8


Modernisation - Statistical Compilation System and Process
Management of Statistical System
Strengthening the living quarters frame – Malaysia Statistical Address Register (MSAR)
Strengthening the establishment frame – Malaysia Statistical Business Register (MSBR)
9


Modernisation - Methodology and best practices in statistics
Online system for economic Surveys (e-survey & e-census)
Computer-assisted telephone interview (CATI) for labour force survey
Computer-assisted personal interview (CAPI) for monthly consumer prices
Usage of Personal Digital Assistant (PDA) and Tablet in the Collection of Consumer Price
For Population and Housing Census 2010
Data collection via Drop-off Pick-up (DOPU) and e-census
Application of Geographical Information System (GIS)
Data Processing using the Intelligent Character Recognition (ICR) scanner and Computer Assisted Coding (CAC)


Modernisation - Usage of Secondary Data
Smart Partnership Programme with other government and private agencies
Intensifying statistical interactive programme with industry players, such as, Commerce Association
Strengthening the management of secondary data usage
Acquiring Big Data Analytics applications and producing new products
11


Modernisation - Integrity, security and confidentiality of information
Strengthening Internalisation of Integrity
Strengthening Data Dissemination Governance
Improvement on Security of Data Storage, Accessibility and Dissemination
12

Modernisation - ICT Infrastructure

STATSBDA BIG DATA ANALYTICS

National Enterprise-Wide Statistical Systems (NEWSS) : to integrate, incorporate field activities, sampling, data collection, processing, analysis, estimation and report generation

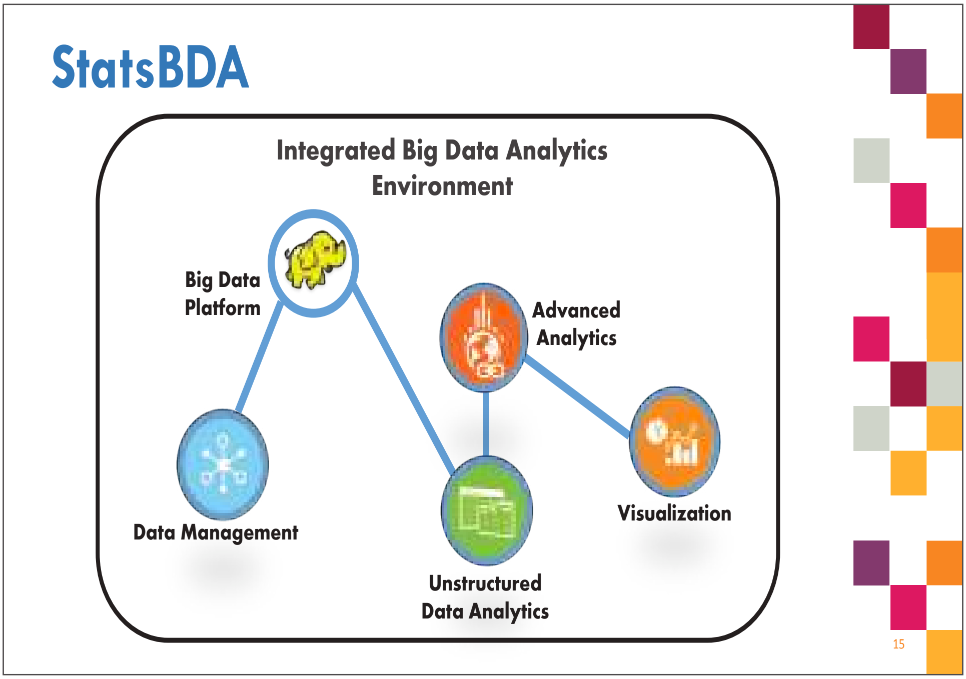
Statistics Data Warehouse (StatsDW) is an Enterprise Data Warehouse (EDW). The StatsDW enable fast and easy access to data stored

Big Data Analytics (StatsBDA) project was kicked off on December 2016. The StatsBDA will increase the use of new data sources as well as simplify the process of decision making

13

StatsDW

14


StatsBDA
Integrated Big Data Analytics Environment
Big Data Platform
Advanced Analytics
Data Management
Visualization
Unstructured Data Analytics
15


StatsBDA
Trade by Enterprise Characteristics (TEC)
e-Info
Real Time Business Status
Price Intelligence (PI)
Biz@codes@stats
Public Maturity Assessment on Official Statistics (PMAOS)
REALTIME
Real time news on official statistics (RTOS)
16

Modernisation - Statistical Delivery System

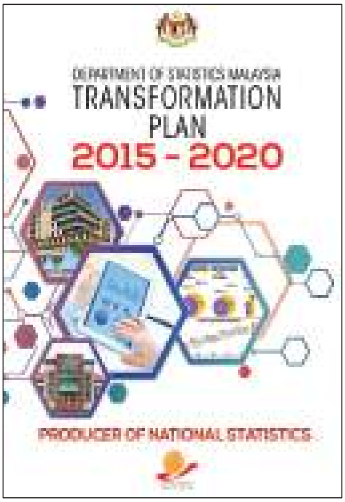
17

Modernisation - Human Capital Development

- Strengthening Professionalism, Personality, Integrity, Career Development and Integrate Smart Talents
- Empowering Statistical Services
- Strengthening the Role of Malaysia Statistical Training Institute as a Centre of Excellence
- Creating Subject Matter Expert (SME)

18

DOSM – The Way Forward



4 Strategic Thrust

- Strengthening the Role as a Producer of Statistics
- Producing Statistical Products and Services of Integrity and Reliability
- Empowering Talents
- Managing Resources and Infrastructure Efficiently

19

DOSM – The Way Forward

- Strengthen the methodology used based on current and relevant international standards
- Modernise the statistical work processes
- Expand and integrate the use of administrative data
- Capitalise on data evolution through big data
- Widen the production of new statistics by focus area
- Enhance the capability on analytics and innovation
- Strengthen the medium of statistics communication
- Increase statistical literacy awareness towards evidence-based decision making

20

DOSM – The Way Forward

Population Census 2020

Malaysia Census Transformation Programme 2020 (MyCTP 2020)

Malaysia Integrated Population Census Framework (MyIPCF)

21

Norhayati binti Shuja'
yatishuja@stats.gov.my

Emerging Challenges in Data Collection, Survey Methodology, and Implication for Official Statistics: Banking Sector Economist's Needs and Perspectives

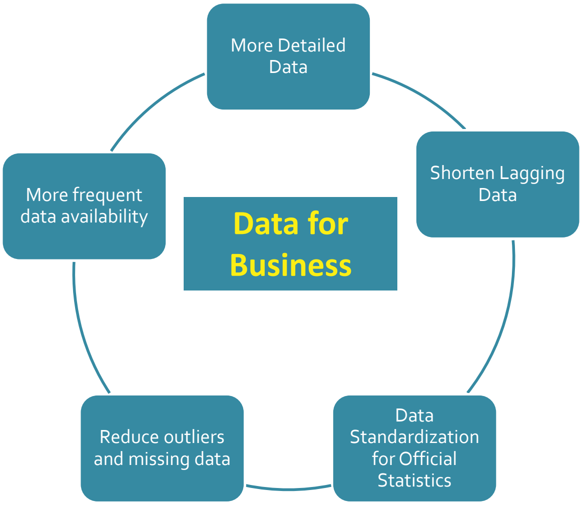
Anton Gunawan
Chief Economist, Bank Mandiri

One Heart One Mandiri

Emerging Challenges in Data Collection, Survey Methodology, and Implication for Official Statistics: banking sector economist's needs and perspectives

Anton Gunawan
Chief Economist Bank Mandiri

What business seek from official statistics?



General Challenges from Data Gathering

mandiri

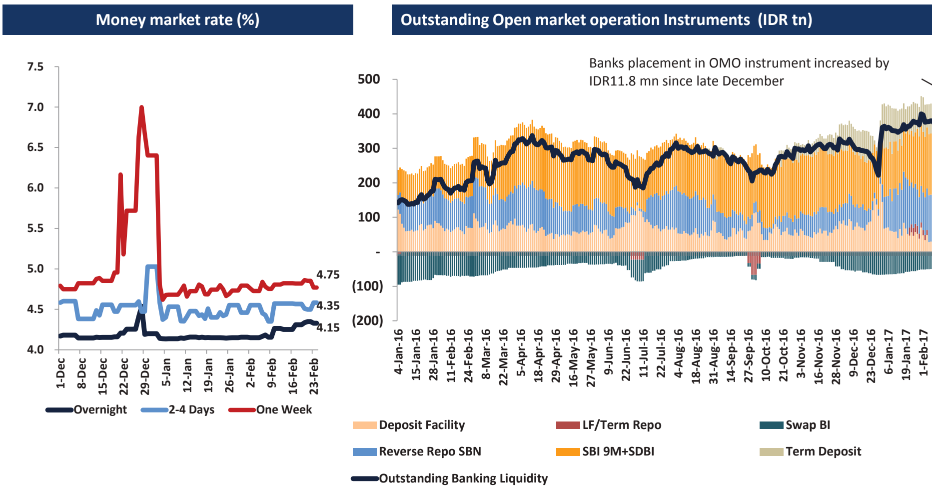
Data Classification:
Our Economists/Analysts categorize secondary data into 3:

Macro & Financial Data	Industry & Regional	Socio-Economic
• GDP, Inflation, Fiscal, Balance of Payments, External debt and Leading indicators • Banking and non-banking financial institutions statistics • Money Market and inter-bank data • Flow of funds	• Sectoral data: cement sales, auto sales, consumer confidence index, retail sales, Input-Output, industry statistics • Regional data: APBD (local budget), Province/Kabupaten in numbers	• Socio-economic secondary data: health, poverty, education and unemployment, basic foods • Social primary data: household spending, SME data
Challenges	**Challenges**	**Challenges**
• Lag in providing data more than 2 months: banking data, non-banking FI data (insurance, mutual fund, pension fund) • For KUR (lending for poor people) data, no cross sectional data	• For industry data: too many outliers, lots of missing data, and the lag is too long (2 years) • Industry statistics sometimes cannot capture provincial or Kabupaten data • Data issuance only on annual basis • Local budget data are lagging	• Some gender-based data are lagging too long (more than two years) • No Kabupaten/Municipality-based data • Decentralization creates challenges on having similar quality of local-based data (similar form)

Trying to show the banking sector liquidity using OMO instruments

Another approach is constrained by lagging data availability

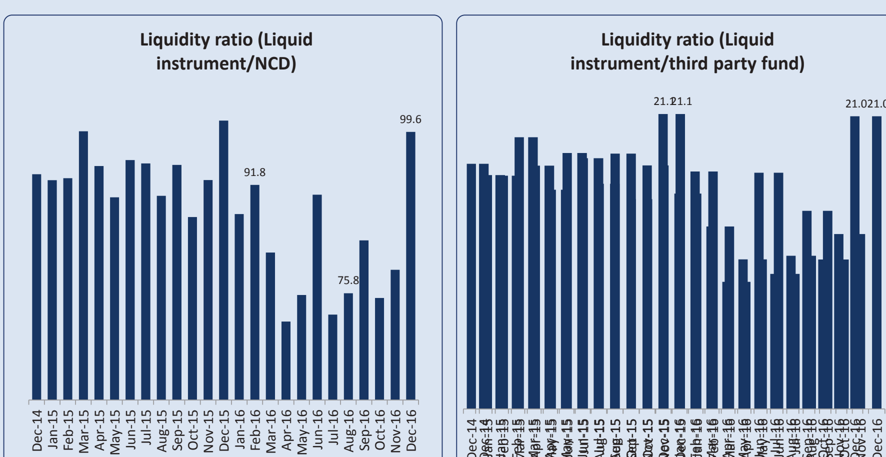
mandiri



Source : Indonesian Banking Statistic

Banking sector's Liquidity ratio increased significantly in Dec16 (two months lag)

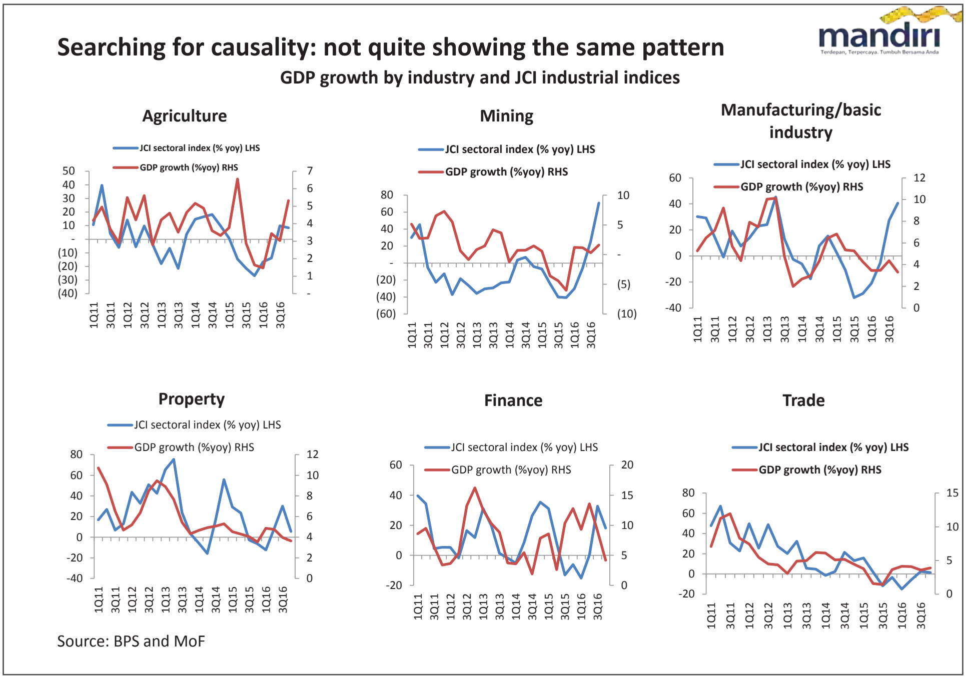
mandiri

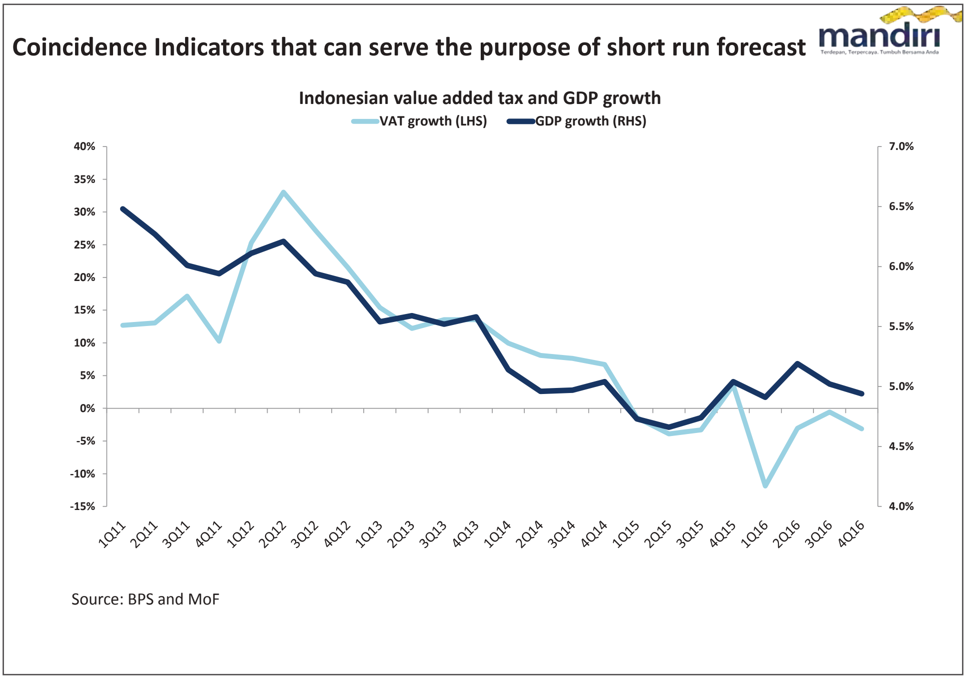


Liquid instrument = cash + placement at BI+ Excess Liquidity
NCD = Non core deposit = 30% demand deposit + 30% saving deposit + 10% time deposit

Updated : Dec 2016

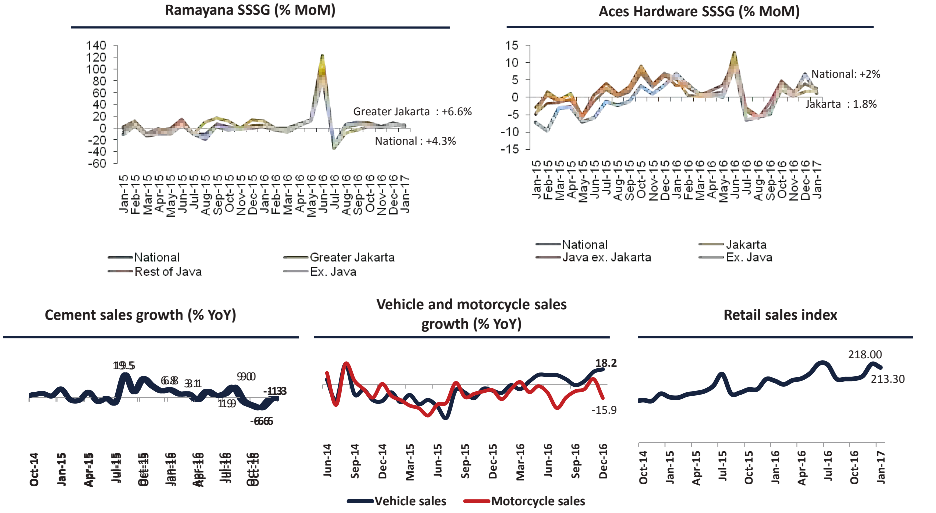
Source : Indonesian Banking Statistics, Bank Mandiri calculation


Searching for causality: not quite showing the same pattern
GDP growth by industry and JCI industrial indices
Agriculture
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Mining
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Manufacturing/basic industry
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Property
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Finance
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Trade
JCI sectoral index (% yoy) LHS
GDP growth (%yoy) RHS
Source: BPS and MoF


Coincidence Indicators that can serve the purpose of short run forecast
Indonesian value added tax and GDP growth
VAT growth (LHS)
GDP growth (RHS)
Source: BPS and MoF

Looking for supporting evidences in the more micro level data

Mixed performances in some benchmark indicators : seasonal pick up in retail sales



Source : Companies, CEIC

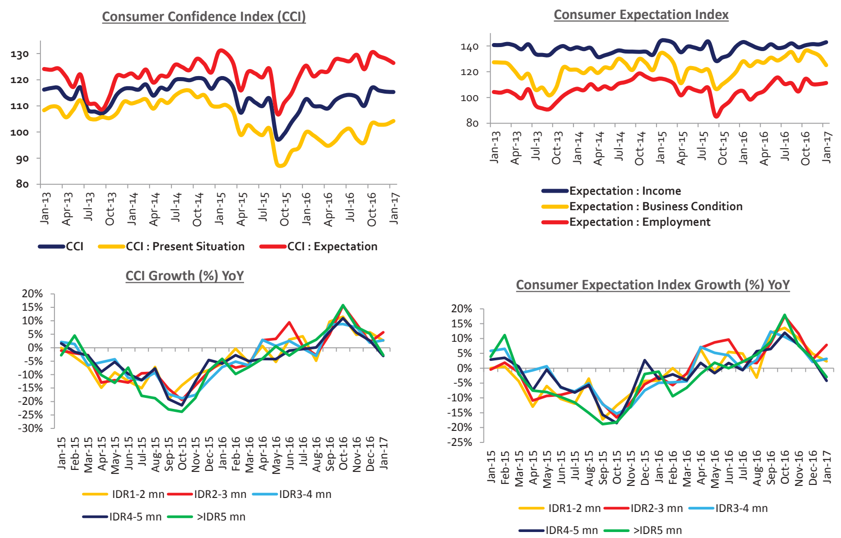
Some sectors have improved with modest level

Sector	Current Market Demand Assessment d)	Indicator	Growth (%, yoy) Nov-16	Jan-Nov 2015	2015	Jan-Nov 2016	Notes
Retail	Positive	Ramayana sales	6.2 b)		-2.0	5.8 a	Consumer confidence is flat
		Ace Hardware sales	10.9 b)		4.7	4.8 a	
Automotive	Neutral	Total car sales	14.9	-16.8	-16.1	3.7	Consumer confidence is flat
		Commercial vehicle	-18.9	-16.1	-16.6	-29.2	
		Passenger vehicle	26.5	-17	-16	15.3	
		Total motorcycle sales	6.6	-18	-17.6	-8	
Property	Negative	Total Marketing Sales of eight companies	35.5 c)		-4.3	-8.6 a	Consumer confidence is flat and interest rates are still high
Cement	Positive	Total sales	0.9 b)		1.83	1.24 a	Sales growth is supported by government infrastructure projects
Heavy Equipment	Negative	Total sales	48	-41.2	-39.5	-0.1	Demand from commodity sector has not yet fully recovered even though commodity prices increased
		Forestry segment	65	-8.0	-9.3	-32.2	
		Agro segment	222	-67.3	-63.2	-21.2	
		Mining segment	100	-52.3	-49.9	-4.0	
		Construction segment	10	-22.4	-22.3	29.1	
Sea Transport	Negative	Loading and unloading activities in the 5 main ports	3.4	-7.9	-8.2	-1.0	Volume of trade has not yet increased, especially in commodities sector
Oil	Negative	CAPEX (Investments)	n.a	n.a	-2.6		Oil price is still low
CPO	Negative	Volume of exports	17.74	14.34	15.62	-15.16	CPO price has consistently increased but still needs time to impact the export performance
		Value of exports	52.98	-12.71	-11.91	-10.06	
Coal	Negative	Volume of exports	14.59	-9.95	-10.11	0.1	Coal price remains low
		Value of exports	34.18	-22.71	-23.15	-13.17	

Notes : a) Jan-Dec b) Dec c)Fourth-Quarter 2016 d) the status is not OCE Industry Rating ; n.a. data not available

Consumer expectation had weakened because of increasing inflation

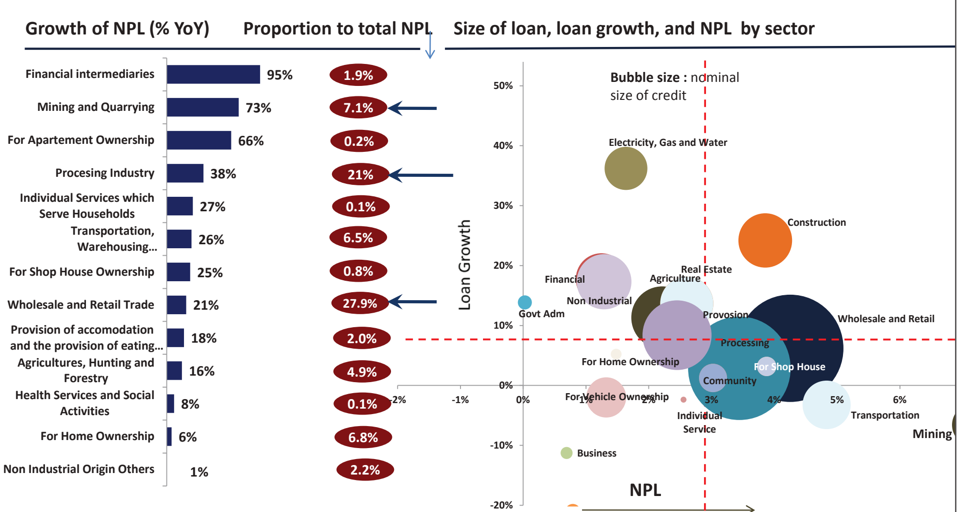
People with higher income (>IDR5 mn) seen less optimistic



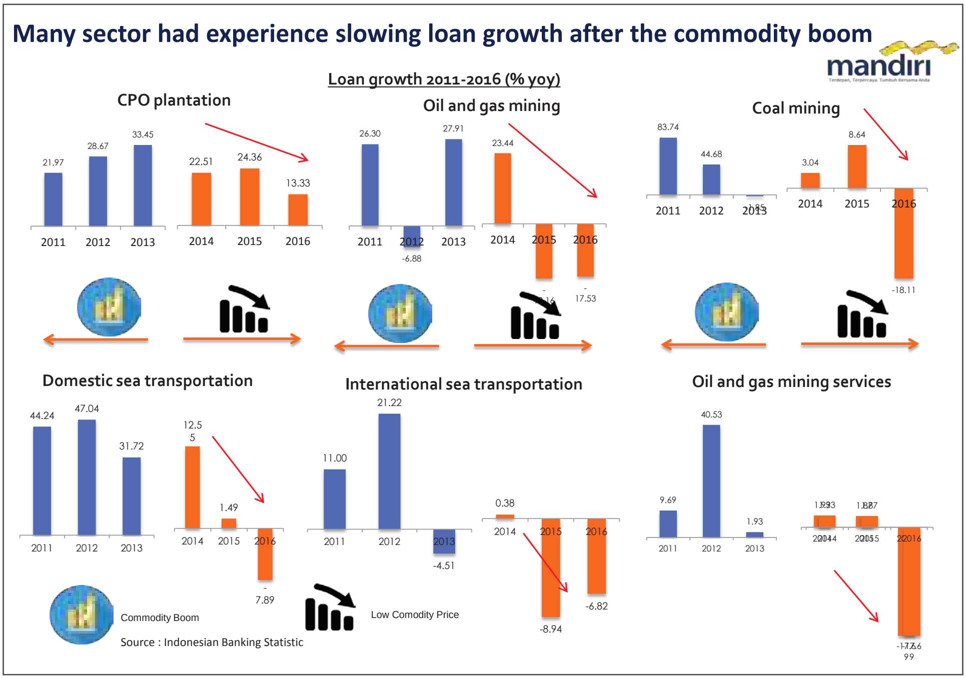
Source: Bank Indonesia, CEIC

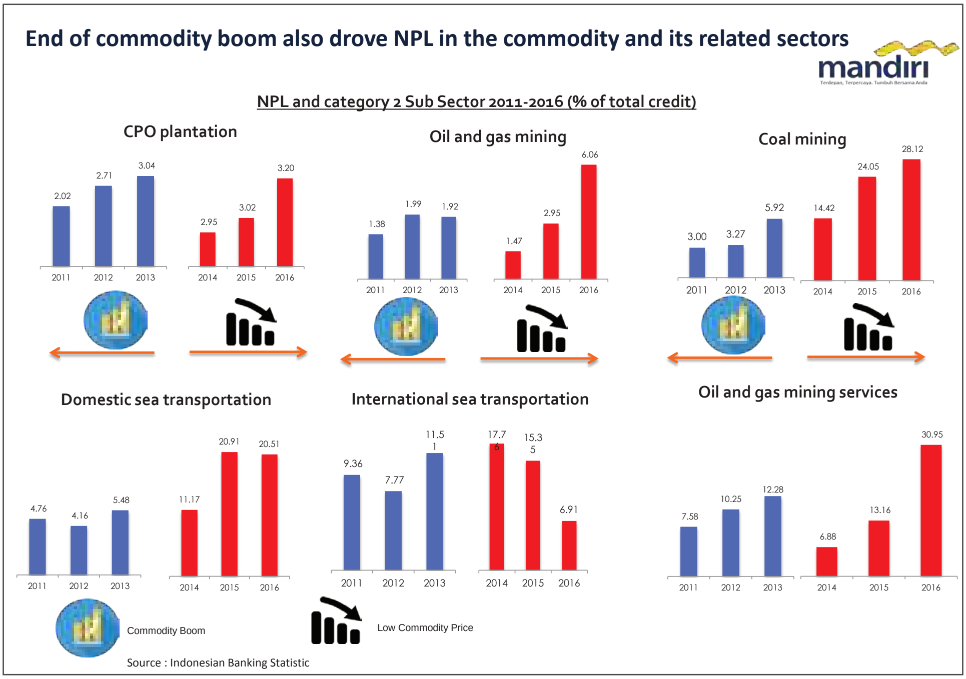
11

Understanding the whole value chain of the sector is very important: Mining sector had the highest NPL ratio, but the value of NPL in Dec16, amid contracted loan growth



Source : Indonesian Banking Statistics, Bank Mandiri calculation

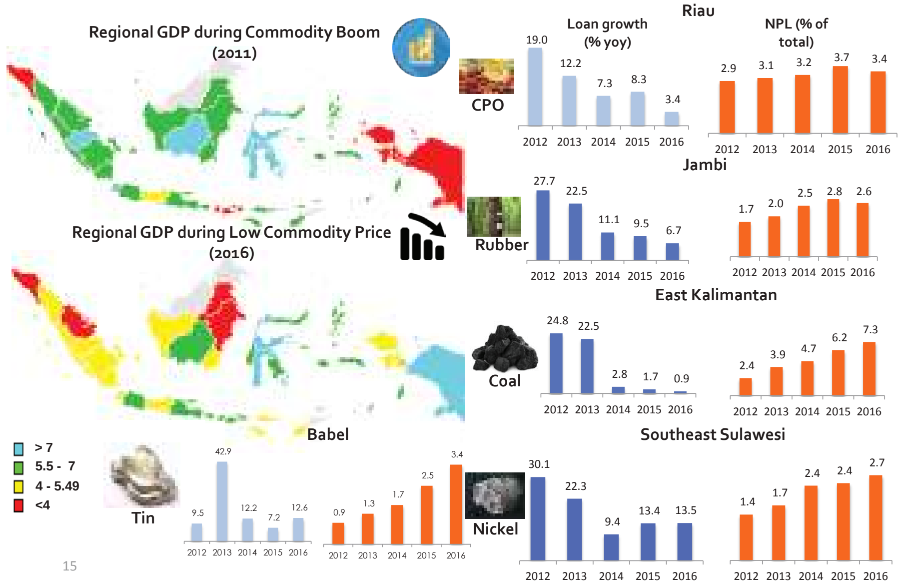

Many sector had experience slowing loan growth after the commodity boom
mandiri
Loan growth 2011-2016 (% yoy)
CPO plantation
21.97 28.67 33.45 22.51 24.36 13.33
2011 2012 2013 2014 2015 2016
Oil and gas mining
26.30 -6.88 27.91 23.44 17.53
2011 2012 2013 2014 2015 2016
Coal mining
83.74 44.68 3.04 8.64 -18.11
2011 2012 2013 2014 2015 2016
Domestic sea transportation
44.24 47.04 31.72 1.49 7.89
2011 2012 2013 2014 2015 2016
International sea transportation
11.00 21.22 -4.51 0.38 -8.94 -6.82
2011 2012 2013 2014 2015 2016
Oil and gas mining services
9.69 40.53 1.93
2011 2012 2013
Commodity Boom
Low Comodity Price
Source : Indonesian Banking Statistic


End of commodity boom also drove NPL in the commodity and its related sectors
mandiri
NPL and category 2 Sub Sector 2011-2016 (% of total credit)
CPO plantation
2.02 2.71 3.04 2.95 3.02 3.20
2011 2012 2013 2014 2015 2016
Oil and gas mining
1.38 1.99 1.92 1.47 2.95 6.06
2011 2012 2013 2014 2015 2016
Coal mining
3.00 3.27 5.92 14.42 24.05 28.12
2011 2012 2013 2014 2015 2016
Domestic sea transportation
4.76 4.16 5.48 11.17 20.91 20.51
2011 2012 2013 2014 2015 2016
International sea transportation
9.36 7.77 6.91
2011 2012 2013 2014 2015 2016
Oil and gas mining services
7.58 10.25 12.28 6.88 13.16 30.95
2011 2012 2013 2014 2015 2016
Commodity Boom
Low Commodity Price
Source : Indonesian Banking Statistic

The commodity producing regions had suffered the most as commodity boom ends

Loan growth on the regions fell and NPL increased

mandiri



15

Macroeconomic and Financial Sector Data (1)

mandiri

No	Needs	Data Sources	Availability	Weakness	Suggest
1	Transaction volume and value of internet banking	Bank Indonesia	Not available		Provide transaction volume and value for internet banking by: - BUKU - Type of bank (SOE, non SOE, etc) - Location
2	Transaction volume and value of server based and chip based electronic money	Bank Indonesia	Not available		Provide transaction volume and value for server based electronic money by: - BUKU - Type of bank (SOE, non SOE, etc) - Location
3	Transaction volume and value of electronic channel payment (e.g. electronic money, CC, Debit Card) for some special purpose (e.g in public transportation, toll payment, fuel oil transaction, etc)	Bank Indonesia	Not available		Provide transaction volume and value of electronic channel payment (e.g. electronic money, CC, Debit Card) for some special purpose (e.g in public transport, toll payment, gas station, etc)
4	More data on KUR realization	Coordinating Ministry of Economy	Current data is not presented in the form of a cross section		Provide cross data of KUR realization between period, bank, debtor, location, NPL, and sector
5	Indonesian Banking Statistics	OJK/Bank Indonesia	Published monthly	Very long time-lag (3 month)	Need shorter time lag (1 month)

Macroeconomic and Financial Sector Data (2)

No	Needs	Data Sources	Availability	Weakness	Suggest
6	Update of list of banks by BUKU	OJK/Bank Indonesia	Not available		List should be published frequently
7	Loan growth by segment (corporate, commercial, business, micro, etc)	OJK/Bank Indonesia	Not available		Data should be published frequently
8	Monthly data of individual banks (asset, loans, deposit, LDR, NPL, income statement, etc)	OJK/Bank Indonesia	Not available		Data should be published frequently
9	Open market operations	Bank Indonesia	Available by instruments		Data should be available by instruments and by tenors
10	Budget Realization	MoF	Monthly data available	But the lag is too long (more than 2 months)	Two weeks or at most one month lag
11	Flow of Funds	BPS	Quarterly and Annual	Lagging too long	The latest data is of 2Q16. possible for 1 quarter lag?

Industry Data (1)

No	Needs	Data Sources	Availability	Weakness	Suggest
1	Physical Infrastructure realization data		Not Available		The easiest way is for this report to be included in the government budget accountability report.
2	Detail of construction sectors in GDP, broken down to property, infrastructure, building, etc		Not Available		
3	Time Charter Rate (Sea Transport Statistics)		Not Available		
4	Industry statistics	BPS	Available annually but with a lag	1) Too many outliers	1) Minimalize the missing data and outliers
				2) A lot of missing data	2) reducing data lag
				3) The lag is too long (2 years)	

Industry Data (2)

No	Needs	Data Sources	Availability	Weakness	Suggest
5	Cement Price		Not Available		3 years ago Ministry of Trade still publish this data, but then it stop. We suggest that this data can be release again soon.
6	Loading Unloading data in Seaport	BPS	Until now only 5 sea port included in the statistics	There are many sea port which have not yet been included in the statistics	Provide data for at least a sea port from each provinces
7	Local Urea Fertilizer Price		Not Available		
8	Coal Production by Region	Kementerian ESDM and Local Government	Not Available		Availability of Monthly Coal Production by Region and by type of producers license (IUP or PKP2B)
9	Domestic Consumption of Coal	Kementerian ESDM	Not Available		Availability of Domestic Consumption of Coal (by Sector & by region)
10	Oil and Gas Investment Realization	SKK Migas	Oil and Gas Investment Realization	Only in Annual Basis	Availability of Oil and Gas Investment Realization in Monthly Basis and with type of investment activity
11	Fuel Consumption	BPH Migas / Kementerian ESDM	Not Available		Availability of Fuel Consumption Data in Monthly Basis, by type and by Region

Industry Data (3)

No	Needs	Data Sources	Availability	Weakness	Suggest
12	Electricity Consumption & Production	PLN/ Kementerian ESDM	Electricity Consumption & Production by Region	Only in Annual Basis	Availability of Electricity Consumption & Production Data in Monthly Basis, by type of customer and by region
13	Electricity Generating Capacity	PLN/ Kementerian ESDM	Electricity Generating Capacity by Region	Only in Annual Basis	Availability of Electricity Generating Capacity in Monthly Basis, by type of generator, by energy source and by region
14	Number of vehicles by brands per region	BPS	Number of vehicles by type on national level	Cannot identify region's potency	Data by Region
15	Vehicles sales by brands per region	GAIKINDO	National vehicles sales	Cannot identify region's potency	Data by Region
16	Interest rate for housing/apartment loan based on project/business location	Bank Indonesia	Not available		Interest rate for housing/apartment loan based on project/business location
17	Number of housing sold based on 34 province	Bank Indonesia	Growth of housing sales in total of 15 cities		Number of housing sold based in 34 province
18	Housing price value based on 34 province	Bank Indonesia	Housing price index in 15 cities		Housing price value based on 34 province

Industry Data (4)

No	Needs	Data Sources	Availability	Weakness	Suggest
19	Price value of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province	Bank Indonesia	Index price of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 7 cities		Price value of rental rate of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province
20	Supply number of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province	Bank Indonesia	Index supply of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 7 cities		Supply number of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province
21	Demand/occupancy rate of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province	Bank Indonesia	Index demand of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 7 cities		Demand/occupancy rate of office space, retail space, apartment, industrial land, hotel, convention hall, warehouse in 34 province
22	More data on development of housing based on province	Ministry of Public Housing and Public Works	Not available		Provide more detailed information on housing development in Indonesia
23	More data on housing needs based on province	BPS	Not available		Provide more detailed information on housing needsin Indonesia
24	Composition Requirements Heavy Equipment by Type and Sector (Agro, Mining, Construction, Forestry)	Kementerian PUPR	Not Available		Provide the data at least quarterly

Industry Data (5)

No	Needs	Data Sources	Availability	Weakness	Suggest
25	Supply and Demand Heavy Equipment	BPS/Kementerian PUPR	Published on articles	Inconclusive	Provide the data at least yearly
26	Raw Sugar and White Sugar prices	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least quarterly
27	Raw Sugar and White Sugar Consumption (Household and Industrial)	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least quarterly
28	White Sugar and Rafinated Sugar Consumption and Production	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least quarterly
29	Production Capacity Sugar Mill	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least quarterly
30	Raw Sugar Impor	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least quarterly

Industry Data (6)

No	Needs	Data Sources	Availability	Weakness	Suggest
31	Sugar Land Area	Ditjen Perkebunan/BPS	Published in Statistik Perkebunan	Very long time lag (almost 2 years)	Need shorter time lag (1 year at least)
32	Cost of Sugar Production in Various Areas/Region	Ditjen Perkebunan/BPS	Not Available		Provide the data at least quarterly
33	Supply and Demand Pulp and Paper	BPS	Published in Statistik Perkebunan	Very long time lag (almost 2 years)	Need shorter time lag (1 year at least)
34	Consumption Pulp and Paper, National and per Capita	BPS	Published on articles	Inconclusive	Provide the data at least yearly
35	Production Capacity Pulp and Paper (National and Big Player/5 Top Players)	Ditjen Perkebunan	Published on articles	Inconclusive	Provide the data at least yearly
36	HTI land area	Ditjen Perkebunan/BPS	Published in Statistik Perkebunan	Very long time lag (almost 2 years)	Need shorter time lag (1 year at least)
37	Raw material composition in pulp and paper production	Ditjen Perkebunan/BPS	Not Available		Provide the data at least yearly

Regional Data

No	Needs	Data Sources	Availability	Weakness	Suggest
1	Non Performing Loan Data Based on Project/Business Location	OJK	Non Performing Loan Data Based on Bank Location		Non Performing Loan Data Based on Project/Business Location in banking Statistics
2	Regency and Municipality Statistics	BPS		1) The Lag is too long	
				2) Sometimes they do not have similar format	
3	Regional Investment Realization per sector		Not Available		The data only can be accessed if we go to regional investment coordinating board. To simplify we need this data to be posted in central investment coordinating board portal.
4	Regional Government Budget Realization	DJPK Kementerian Keuangan		Not all region included in the report. Maybe because not all of the region submit it on time	

Thank You

Closing Remarks

"Enhancing Statistics Prospering Human Life"

Gantiah Wuryandani
Director of Statistics Department, Bank Indonesia
Closing Remarks

International Statistical Institute-Regional Statistics Conference (ISI-RSC) 2017,
24 March 2017, Bali, Indonesia

Distinguished Guest,

Ladies and Gentlemen,

Participants of ISI RSC Conference 2017

Assalamualaikum Wa Rahmatullaahi Wa Barakatuh

1. We are finally approaching the end of the conference. On behalf of Bank Indonesia, I would like to express my sincere gratitude to all of you who have bestowed your presence in this second ISI regional statistics conference in Bali, the island of God. Some of you having travel so far and spend your valuable time to be here, to speak, to discuss, and to participate. I believe in this conference all participants have shared views and experiences that created fruitful discussion, for the past few days.

2. It is my honor to make a few remarks in this closing opportunity. I would like to thank and highly appreciate International Statistical Institute and its Southeast Asia Regional Networks, Irving Fisher Committee of Bank for International Settlement, and European Central Bank for their valuable and significant contribution that ensure this conference took place. I also thankful to NSO, FMS and Ikatan Perstatistikan Indonesia for their support. In this occasion, I would like to express my heartfelt appreciation to the local organizing committee and especially my team for their hard work in the long journey of shaping this successful conference. And of course, to many others whom I can not mention here one by one.

3. This conference is a valuable opportunity to gather many participants from central banker, academician, researcher, and policy maker to talk over global issues in statistics, to support and generate better human's life. I guess, we all believe and it is undeniable that statistics is the key role of our decisions. I would not try to summarize the whole discussion of statistics issues in the conference, while there are plenty exotics and alluring spots in Bali island waiting for you, to enjoy it.

Distinguished,

4. You will all agree with me that we had a very valuable and useful discussion during the conference. We heard very open and insightful thought from many expert and prominent speakers. From the pre event conference on March 20-21 to the main event on March 22-24, broad issues of statistics have been discussed, from the perspective of theory, methodology, and practises in the field of macroeconomics including banking and finance, sharia economic, environmental, demography and social welfare, health and many more. Statistics is not just numbers since it could tell us everything. Researcher, analyst, policy maker highly demand broaden range of datasets that rely on statistics, particularly in the situation of making critical decision making. This has been confessed by Mr Budiono, the 11th vice president of Indonesia, in the first plenary session.

5. The conference has discussed various development in theory, computation and methodology of statistics, including the practices carried out in the international community. Hence, the production of statistics is expected to become more valid and reliable towards better human life in every field. The more coverage and the more granular statistics has become essential to be obtained, using new innovation in theory and methodology. We need to update and continuously improve our theoretical base and methodology applied, in statistical science and practice.

6. Furthermore, our focus should go beyond what have been discussed in this conference. The challenges are remained. Our capacity to develop useful statistics has to be continuously updated and innovated along with the remarkable huge development in human activities, supported by fast development in technology. The era of digital economy encourages statisticians to have revolution in theory and methodology of statistics, including data collection. For example, Big Data has become an encounter towards optimal utilization of structured and unstructured massive volume of data. For all these concepts to carry valuable result, a high level of professionalism and integrity of statisticians in national and international agencies is necessary. All participants should be able to benefit from these achievements.

Ladies and Gentlemen

7. We must not forget the big picture: We have plenty of reasons to be proud of all relevant and high quality statistics that are being produced every day, all over the world to shape policies around the globe. So let's keep up the good work and hold fast to our mission to improve human's life. This is what we are about. Better statistics indeed creates better life. Again, thank you all for participating, and we count on you to be good ambassadors for statistics.

8. Before I close this remark, since you are already here in Bali, the paradise island, I encourage you to explore fascinating sites, from incredible beaches, amazing undersea, delicious traditional culinary, magical culture, beautiful terrace paddy field or just shopping souvenirs and many more to discover and to indulge yourself. I induce you to extend your stay in Indonesia not just in Bali, but also to its surrounding neighborhood wonderful islands, such as Lombok, Komodo, and Wakatobi. You will have an unforgettable thrill moment during your visit in Indonesia and it is addictive to make you come back again.

Finally, dear distinguished participant, ladies and gentlemen,

I apologize for any inconvenient condition. I wish you all have safe journey home. May God Almighty bless you in your future and ever. With that, I close this conference officially.

Thank you, thank you, and thank you

Wassalamu alaikum wa rahmatullaahi wa barakatuh

Index

B

C

D

E

F

K

L

M

N