62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

SPECIAL TOPIC SESSION

VOLUME 1

62nd ISI WORLD STATISTICS CONGRESS 2019

18 - 23 August 2019, Kuala Lumpur

Come | Connect | Create

www.isi2019.org

PROCEEDING

ISI WORLD STATISTICS CONGRESS 2019

SPECIAL TOPIC SESSION (VOLUME 1)

Published by:
Department of Statistics Malaysia
Block C6, Complex C
Federal Government Administrative Centre
62514 Putrajaya
MALAYSIA

Central Bank of Malaysia
Jalan Dato' Onn
P.O. Box 10922
50929 Kuala Lumpur
MALAYSIA

Malaysia Institute of Statistics
Department of Mathematical Sciences
Faculty of Sciences and Technology
43600 UKM Bangi, Selangor
MALAYSIA

Portal : https://www.isi2019.org
Email : lpc@isi2019.org

Published in February 2020

Suggested citation:
Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 1, 2019. 461 pages

Preface

The 62nd International Statistical Institute World Statistics Congress (ISI WSC 2019) has a long tradition since 1887, held for the first time in Kuala Lumpur, Malaysia on 18 to 23 August 2019. ISI WSC 2019 is a global gathering of statistical practitioners, professionals and experts from industries, academia and official authorities to share insights in the development of statistical sciences.

The congress attracted an overwhelming number of participants across the regions. The scientific sessions were delivered over five days with parallel sessions and e-poster sessions running all day long. The scientific program reaches across the breadth of our discipline that comprised of Invited Paper Sessions (IPS), Special Topic Sessions (STS) and Contributed Paper Sessions (CPS). Papers presented exhibit the vitality of statistics and data science in all its manifestations.

I am very honoured to present the proceedings of ISI WSC 2019 to the authors and delegates of the congress. The proceedings contain papers presented in IPS, STS and CPS which were published in fourteen (14) volumes. Scientific papers were received from August 2018 and were carefully reviewed over few months by an external reviewer headed by Scientific Programme Committee (SPC) and Local Programme Committee (LPC). I am pleased that the papers received cover variety of topics and disciplines from across the world, representing both developed and developing nations.

My utmost gratitude and appreciation with the expertise and dedication of all the reviewers, SPC and LPC members for their contributions that helped to make the scientific programme as outstanding as it has been.

Finally, I wish to acknowledge and extend my sincere thanks to the member of National Organising Committee of ISI WSC 2019 from Department of Statistics Malaysia, Bank Negara Malaysia, Malaysia Institute of Statistics and International Statistical Institute for endless support, commitment and passion in making the ISI WSC 2019 a great success and congress to be remembered.

I hope the proceedings will furnish the statistical science community around the world with an outstanding reference guide.

Thank you.

Dr. Mohd Uzir Mahidin
Chairman
National Organising Committee
62nd ISI WSC 2019

Scientific Programme Committee of the 62nd ISI WSC 2019

Chair
Yves-Laurent Grize, Switzerland

Vice-Chairs
Kwok Tsui, Hong Kong, China
Jessica Utts, USA

Local Programme Committee Chair and Co-Chair
Rozita Talha, Malaysia
Ibrahim Mohamed, Malaysia

Representatives of Associations
Rolando Ocampo, Mexico (IAOS)
Cynthia Clark , USA (IASS)
Bruno de Sousa, Portugal (IASE)
Sugnet Lubbe, South Africa (IASC)
Mark Podolskij, Denmark (BS)
Beatriz Etchegaray Garcia, USA (ISBIS)
Giovanna Jona Lasinio, Italy (TIES)

At-Large Members
Alexandra Schmidt, Canada/Brazil
Tamanna Howlader, Bangladesh

Institutional/Ex Officio
Helen MacGillivray, ISI President, Australia
Fabrizio Ruggeri, Past SPC Chair, Italy

Liaison at ISI President's Office
Ada van Krimpen, ISI Director
Shabani Mehta, ISI Associate Director

Local Programme Committee of the 62nd ISI WSC 2019

Chairperson
Rozita Talha

Co-Chairperson
Prof. Dr. Ibrahim Mohamed

Vice Chairperson
Siti Haslinda Mohd Din
Daniel Chin Shen Li

Members
Mohd Ridauddin Masud
Dr. Azrie Tamjis
Dr. Chuah Kue Peng
Dr. Tng Boon Hwa
Prof. Dr. Ghapor Hussin
Prof. Madya Dr. Yong Zulina Zubairi
Prof. Madya Dr. Wan Zawiah Wan Zin @ Wan Ibrahim
Dr. Nurulkamal Masseran
Siti Salwani Ismail
Rosnah Muhamad Ali
Dr. Ng Kok Haur
Dr. Rossita Mohamad Yunus
Muhammad Rizal Mahmod

TABLE OF CONTENTS

Special Topic Session (STS): Volume 1

Business registers as the backbone of statistics; Challenges and opportunities regarding business register and outcome quality

Irene Salemink, Harrie van der Ven, Barry Coenen,
Rico Konen, Johan Lammers
Statistics Netherlands

Abstract

At Statistics Netherlands the Statistical Business Register (SBR) is positioned as backbone for economic statistics and plays a central role in its production. Related business economic statistics, including National Accounts, are designed and organized as a chain of statistical processes. This chain starts with a coordinated population of statistical units derived from SBR and the operation is placed under central direction; the so called chain-management. This approach governs quality management, including procedures to maintain SBR, coordination of adjustments of statistical estimates due to errors in the SBR populations and prevention of inconsistencies.

The SBR is the first place to correct errors and to repair them on a coordinated way in statistics. However, it's not only the quality of the SBR itself that counts, also the economic indicators derived from it are of importance for customers. In order to control the quality of the chain as a whole, the maintenance of the SBR is tuned with the statistical processes, e.g. by defining different maintenance groups of statistical units in the SBR or combining data of different sources (Administrative, NCB).

Major challenges concern a mixed-mode-multisource-approach for compiling economic statistics and a growing demand to measure the economy adequate. In this model large and complex enterprises are dealt with in a custom fit approach, the role of administrative sources is enlarged, the role of big data is upcoming and more flexibility in the use of the business register is needed.

The contribution by the business register to the business value of economic statistics is crucial. Priority in the development of the processes and the applications depends upon this value.

Keywords

Backbone; Chain-management; Business value

1. Economic Business Statistics; A Chain of Statistical Processes

National Accounts are compiled using various data sources of which a substantial part originates from the business economics domain. The compilation of the underlying statistical data is most commonly organized like

so called stovepipes, in the sense that the individual statistic is produced as a standalone process. The demand for integrated, consistent and coherent statistical information however increases and working in a standalone manner does not contribute to this demand, on the contrary. Due to the ever increasing complexity of the organization of global, large, complex enterprise groups, the difficulties to retrieve statistical data, the various entry points for various business statistics (International Trade, SBS, STS, FATS, FDI), increased availability and use of administrative sources, the complexity of small enterprises comprising phenomena like outsourcing, the increased economic role of the self-employed, the difficulties to describe everything correctly in the Statistical Business Register and the subsequent compilation of statistical data makes that none of the compilers of any business statistic alone can oversee this system of dependencies as a whole. It takes a helicopter view to oversee (at macro level) the increased links and dependencies between enterprises and economic sectors.

How to deal with this increased complexity, the pressure to combine all various data sources as efficiently and less burdensome as possible and still release high quality frames and statistics on time?

A possible solution is to integrate all stages of the production process of related business statistics and national accounts as part of a chain of statistical products, using chain management to orchestrate both the process and the outcome. The Statistical Business Register (SBR) is a crucial part of this chain that starts with a coordinated population derived from the SBR. Therefore in this approach the SBR is positioned as the backbone for all economic statistics and plays a central role in their production.

In order to be able to execute decisions and effects of decisions in a coordinated manner chain management and a culture of shared responsibility between all partners in the chain are a prerequisite.

2. Chain Management

2.1 What is Chain Management?

Chain management is the coordination of the various statistical processes i.e. the processing and designing of various statistical products from the perspective of the whole chain. It comprises the whole set of management and operating activities which aim for improving the cooperation of all actors in the chain so that the result of this joint effort is optimal and transparent for all users.

The chain is designed as a set of links between processes. SN defined these links as so called "steady states" in the Business Architecture (BA). Chain management concerns the links of the whole statistical production process from observation to publication.

2.2 Working under Business Architecture

An important role in the design of the chain of business statistics and a necessary condition to implement chain management has been fulfilled by the SN Business Architecture and the adoption of working under architecture. In the BA implemented at SN three layers are defined; Design, Chain management (concerning activities like planning, monitoring, data quality management) and the actual Statistics production from the start of data collection to publishing. Instead of stovepipes for each individual statistic, the architecture describes these statistical processes as a value added chain and defines several "steady states" in which the statistical data have a well-defined status. The first steady state is of course raw data, the final one the publication data. In the Business Architecture the statistical processes are split up in process steps that take the data from one steady state to the next one. The BA describes statistical processes as a value added chain in which each process step adds value by eliminating uncertainties about the data. At every step checks for possible errors are executed as well as checks for specific possible diversions from the metadata, which are corrected if necessary. This means that at every step the data have attained a specific higher level of quality.

This approach also facilitates the use of data from one process in another one; by making use of the steady states the possibilities to re-use data in other processes is largely increased. By working under architecture there exists no overlap between processes; the work is only done once and at the most efficient place in the overall process.

3. SBR as Backbone in the Chain of Economic Statistics

3.1 SBR positioning and role

The place and role of the Statistical Business Register are very important to serve and support statistical processes. The Dutch SBR is part of the economic-statistics-system and has evolved to a business register system with subsystems. By the obligatory use of the SBR for the chain of economic statistics the SBR is *the* basic infrastructural backbone in this chain. The chain of business economic statistics comprises several main processes and flows between them. Besides the process Business Register, also the processes Consistency of the largest enterprises, Direct estimates turnover (STS), production statistics (SBS), Integration Quarterly National Accounts and Integration of National Accounts are distinguished. Figure 1 shows the main constituents as well as the main flows from each system towards the others.

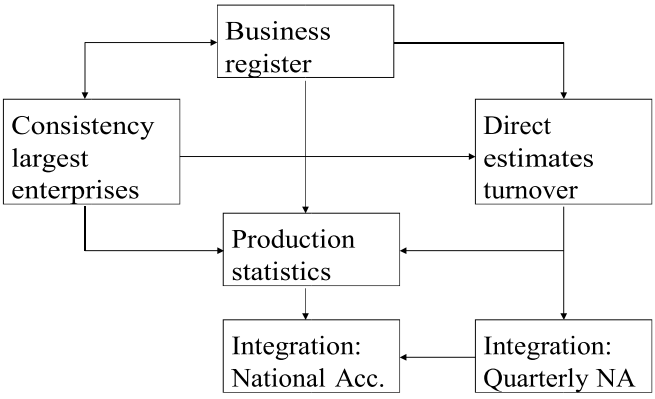


Figure 1: Main flows in the chain, central role for the SBR

In order to satisfy the requirement as structural backbone the SBR has to be the central place where statistical units and their main characteristics are stored and maintained. The Dutch SBR contains therefore the identifying and structural data on businesses at the level of three types of Statistical Units; the Enterprise Groups (EG), Enterprises (Ent) and Local Enterprise Units (LEU). The mandatory utilization of these Statistical Units and the frames derived from the backbone are in turn a precondition for coordination of the statistical output of the chain. And last but not least, the SBR has been expanded and should also facilitate the linkage of administrative units and their information to statistical units. The SBR thus fulfills the role of both:

- Population frame and;
- Statistical frame and as well as of;
- Coupling frame

4. Business value of SBR

The challenges require agile responses, driven by maximal increases of the business value of SBR. Statistics Netherlands uses a strategic agenda. Improvements have to fit within this agenda. A top-down approach. Besides that, the business value of the business register can be determined and any proposal for a change can be evaluated by estimating the contributing of products and services from SBR towards official statistics. A bottom-up approach.

4.1 Business strategy and SBR

The main topics and their corresponding objectives of the strategic agenda, regarding SBR are:

Innovation to create new and faster services, cost reduction and low administrative burden;

Phenomenon-oriented to adopt a phenomenon and society-oriented approach;

Paid services to increase usage of products and services;

Account management to position Statistics Netherlands as a reliable partner and show the added value of collaboration;
Quality to safeguard the quality of processes and output, and make them transparent;
Processes and organization to collaborate effectively and efficiently. The core values are reliable, society-oriented and innovative.

4.2 Business value and SBR

The main categories of business value contributed by the business register are presented in table 2.

Category	Description
Coherence of input-sources	Maximises the use of input sources and the opportunities to combine them.
Comparability over time and between domains	Full support for dynamics in the populations and for sub selections
Consistency between statistics	Coordination by the use of statistical units and variables for stratification and sub selection
Timeliness	Timely processing of changes in populations and of available register information
Completeness	Full coverage and minimised over coverage
Efficiency	Minimises costs to produce business register products and services
Flexibility, Transparency and Openness	Ability to adapt to changes in input, processing and needs in co-operation with partners

5. Challenges

5.1 Increased supply of data sources

The availability of administrative registers, data on financial and economical transactions, sensordata, data in the Internet (Webscraping) and in social media is increasing. For small and medium size enterprises these as the main sources to maintain the SBR and to compile economic statistics. For large and complex Enterprises a more dedicated strategy, containing direct observation and interactive editing will be useful. This approach implies that various sources (multisource) and various modes (mixed mode) altogether are being the input for making economic statistics. The SBR has to supply means (units, coupling) to process the information of these sources and to shorten the 'time-to-market' for new sources. Maintenance of SBR itself benefits from this too.

5.2 A growing demand to measure the economy adequate

"Measuring the economy has become even more challenging in recent times, in part as a consequence of the digital revolution. Quality improvements and product innovation have been especially rapid in the field of information technology. Not only are such quality improvements themselves difficult to measure, but they have also made possible completely new ways of exchanging and providing services. Disruptive business models, such as those of Spotify, Amazon Marketplace and Airbnb, are often not well-captured by established statistical methods, while the increased opportunities enabled by online connectivity and access to information provided through the internet have muddied the boundary between work and home production. Moreover, while measuring physical capital – machinery and structures – is hard enough, in the modern economy, intangible and unobservable knowledge-based assets have become increasingly important. Finally, businesses such as Google operate across national boundaries in ways that can render it difficult to allocate value added to particular countries in a meaningful fashion. Measuring the economy has never been harder." [4]

Towards the SBR this challenges to be more flexible in defining and selecting populations, to be able to use information from administrative sources and from various kinds of (big data) sources. Meanwhile the consistency and the coherence within the system of economic statistics needs to be preserved.

References

1. van Delden, A., Lammertsma, A., van de Ven, P. (2009): Chain management in statistics: best practices, SN discussion paper 090403.
2. Konen, R. (2014): Maintenance strategy of the Dutch SBR, 24th Meeting of the Wiesbaden Group on Business Registers, Vienna, September 15th.
3. Vennix, K. (2012): The treatment of large enterprise groups within Statistics Netherlands, ICES IV, Montreal, June 13th 2012.
4. Professor Sir Charles Bean (2016): Independent Review of UK Economic Statistics

Quality assurance framework of the Japanese business register

Takashi IOKA[1]
Statistics Bureau of Japan, Ministry of Internal Affairs and Communications

Abstract

In this paper, two kinds of new frameworks which will start from 2019 to ensure the quality of the Statistical Business Register (BR) in the Statistics Bureau of Japan (SBJ) will be introduced. The first framework is the Economic Census 2019 which will be conducted from June 2019. In Japan, the Economic Census has been conducted almost twice every five years (2009, 2012, 2014 and 2016). The results of the Economic Census have been recorded in the BR as its basic data source. In the results of the last Economic Census conducted in 2016, all the enterprises located in Japan have been recorded in the BR with basic information such as their name, address, number of employees, industrial classification, turnover, and so on. In the Economic Census 2019, a new administrative data called Corporate Number will be used for the purpose of ensuring the quality of the BR frame. The National Tax Agency assigns a Corporate Number (13-digit) to each enterprise, and publishes the number with the name and address of the company's head office to improve the efficiency of management of information on enterprises, reduce the cost associated with reference and exchange of enterprise information, etc. By using this new data for the Economic Census 2019, it is expected that the frame of the BR will be updated in an even more efficient and precise manner. The second one is the profiling which will also start from 2019. The latest basic information of large enterprises will be regularly grasped by the profiling team in the National Statistics Center through an interactive profiling system. The profiling team and the large enterprises can communicate with each other by accessing the system. The information gathered by the profilers will be promptly recorded in the BR. The role of the profilers is not only gathering information from large enterprises, but also reducing their response burden by helping and supporting large enterprises to respond to the surveys conducted by the Japanese government. In this effort, it is expected that the accuracy of the data about large enterprises in the BR will improve more than ever. In this paper, details of those two frameworks will be described.

Keywords

Economic Census; Corporate Number; Profiling

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of the Statistics Bureau of Japan.

1. Introduction

The current BR in Japan, formally called the Establishment Frame Database, has been operated since January 2013 and the importance of and need for the BR goes on increasing year after year. In this paper, two kinds of frameworks which will start from 2019 to ensure the quality of the BR in the SBJ will be introduced. The first framework is the Economic Census 2019 which will be conducted from June 2019, and the second one is the Profiling. Before mentioning the frameworks, outlines of the BR and the Economic Census will be introduced first.

1.1 Outline of the Business Register

In the Statistics Act, it is stipulated that the Minister of Internal Affairs and Communications shall develop the BR in Japan[2]. Based on the Act, the SBJ is in charge of operating the BR. The current BR aims mainly to provide the latest business frame every year, which is called the "Annual Frame," for the sampling frame of business surveys conducted by the national and local governments as well as incorporated administrative agencies, and to reduce the burden on respondents of statistical surveys conducted by these organizations. After the renovation in 2013, the BR has been storing various survey data (survey results) such as the Economic Census, Financial Statements Statistics of Corporations by Industry, and administrative data consisting of the Labor Insurance Data[3], the Commercial and Corporation Registration Data[4], and the Electronic Disclosure for Investors Network Data[5]. The information on all establishments and enterprises in Japan (about 6 million establishments) is stored in the BR. The chart on the next page summarizes the maintenance cycle of the BR for reference.

---

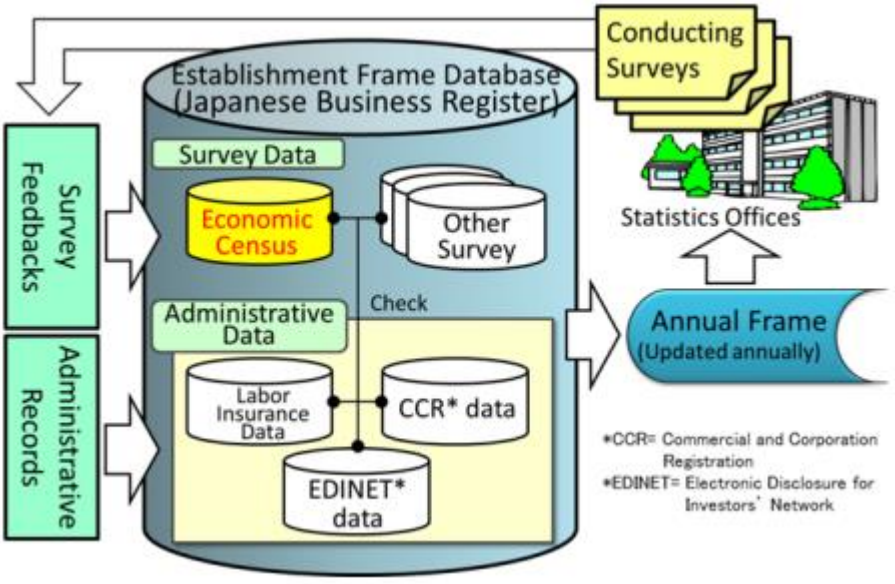
[2] The Statistics Act (Article 27): The Minister of Internal Affairs and Communications shall develop an establishment frame database by utilizing questionnaire information pertaining to fundamental statistical surveys or general statistical surveys, questioning juridical persons and other organizations or through other methods, for the purpose of contributing to the accurate and efficient production of statistics by administrative organs, local public entities, and incorporated administrative agencies, etc. and reduction of the burden on respondents of statistical surveys.

[3] The Labor Insurance Data: The data about businesses which were newly established and discontinued is obtained from the Ministry of Health, Labour and Welfare every month. After obtaining the data, SBJ send a letter of inquiry to those new businesses to obtain the data about industrial classification, number of employees and so on.

[4] The Commercial and Corporate Registration Data: The data about businesses which were newly established, transferred, changed trade names, merged and discontinued is obtained from the Ministry of Justice every month. After obtaining the data, SBJ send a letter of inquiry to those new businesses to obtain the data about industrial classification, number of employees and so on.

[5] System for investors to obtain disclosure documents including information on enterprise name, address, and financial information based on the Financial Instruments and Exchange Act.

Chart: Maintenance Cycle of the Business Register



1.2 Outline of the Economic Census as the Main Data Source for the Business Register

The main data source for the BR is based on the results of the Economic Census. Objectives of the Economic Census are to identify the actual situation of business activities of establishments and enterprises as well as identifying the comprehensive industrial structure in Japan, and to organize information on the population for conducting various statistical surveys for establishments and enterprises.

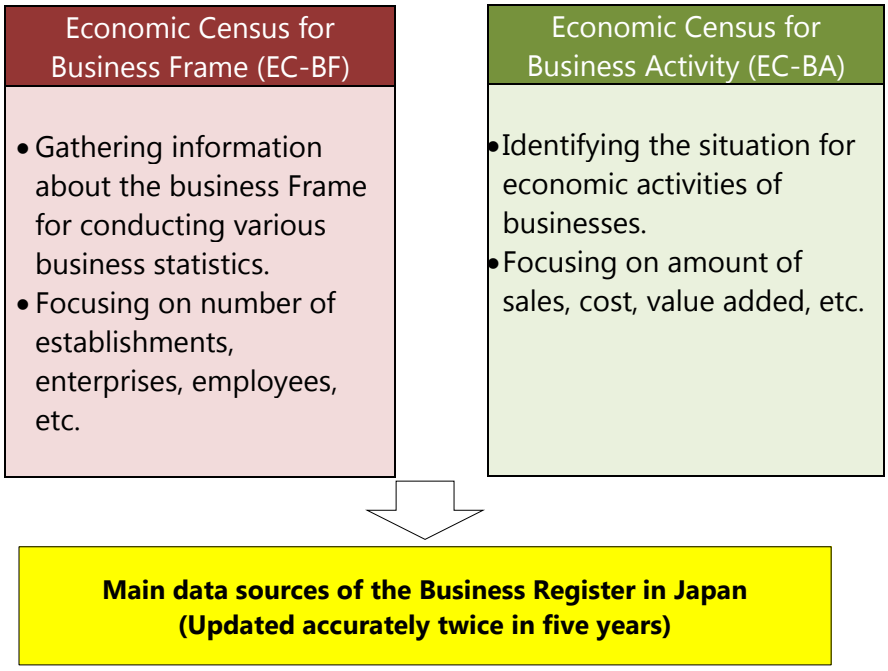
In Japan, the Economic Census consists of two surveys as follows:

a) The Economic Census for Business Frame, to identify the basic structure of establishments and enterprises.

b) The Economic Census for Business Activity, to identify the situation of economic activities of establishments and enterprises.

These two surveys have been conducted approximately once every five years, beginning in 2009. Thus, the Economic Census for Business Frame was conducted in 2009 and 2014, and the Economic Census for Business Activity was conducted in 2012 and 2016. Then, in 2019, the Economic Census for Business Frame will be conducted. As a result, BR has been accurately updated twice in five years by the main data source (Economic Census data).

Figure 1: Outline of the Economic Census



In the interim between two surveys, BR has been updated regularly by using the administrative data and other survey results. This method has been very effective for the maintenance of the BR in Japan.

1.3 New Quality Assurance Frameworks for the Business Register

The economic impact of aging society accompanied by falling child birth rates, and the rapid changes of the economic and social structure are advancing in Japan. To deal with those issues, it became more important than ever before to grasp accurate business trends by improving the official economic statistics.

In this meantime, in 2016, Prime Minister Shinzo Abe held the 22nd meeting of the Council on Economic and Fiscal Policy at the Prime Minister's Office. At the meeting, there was discussion on improvement of economic statistics, focused on GDP statistics, and the Basic Policy for the Fundamental Reform of Economic Statistics was decided. In the policy, effective use of the "Corporate Number" to improve the accuracy of the official statistics was mentioned.

In 2017, according to the basic policy, the Chief Cabinet Secretary suggested to use the Corporate Number, and to conduct the Profiling to improve the quality and coverage of the BR as the basic data source for the official statistics. The SBJ started to put the Chief Cabinet Secretary's suggestion into practice.

2. Methodology

2.1 Economic Census for Business Frame 2019

In the Economic Census 2019, the Corporate Number will be used mainly for the purpose of ensuring the quality of the BR frame. The National Tax Agency assigns a Corporate Number (13-digit unique ID) to each enterprise, and publishes the number with the name and address of the enterprise's head office to improve the efficiency of management of information on enterprises, reduce the cost associated with reference and exchange of enterprise information, etc.

One of the characteristics of this new administrative data is that anyone can use it with the name and address of the enterprises through the internet[6].

And the Corporate Number is a unique ID for each enterprise. One enterprise can have only one unique Corporate Number. Therefore, duplications or omissions of the enterprises stored in the BR will be reduced by using the Corporate Number.

In the Economic Census 2019, the SBJ will newly add the names and addresses of the enterprises grasped by the Corporate Number to the lists of establishments for the enumerators who will practically find out the statuses of the enterprises.

It is prospected that not a few enterprises added by the Corporate Number do not actually exist (fictitious enterprises, dummy companies, enterprises with their actual business places located in different places or dead companies and so on). Therefore, it is effective that the enumerators practically walk the streets and check the actual status of the enterprises, especially the status of the enterprises added by the Corporate Number, before storing them in the BR.

If the enterprises added by the Corporate Number exist, in addition to the newly established establishments, the enumerators will deliver the Questionnaire to those enterprises to gather the basic and essential information for the BR frame such as the number of employees, Corporate Number, industrial classification, turnover and so on.

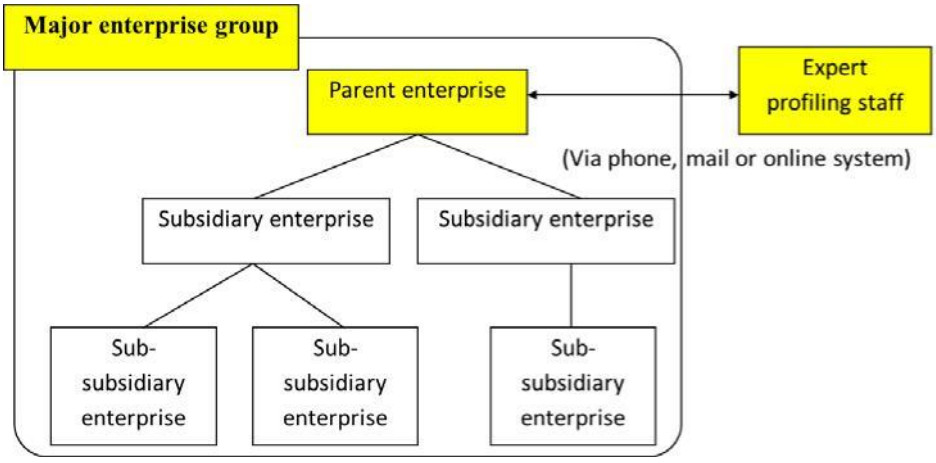
For reference, the Economic Census 2019 will be the first Census for which tablet terminals (portable devices) will be developed and provided for enumerators. The maps and the lists of establishments will be downloaded to the tablet terminals beforehand, and all the enumerators will use them so that the information they ascertain will be shared with the SBJ immediately.

[6] Corporate Number Publication Site (https://www.houjin-bangou.nta.go.jp/en/) is available.

2.2 Profiling

The SBJ is planning to introduce the Profiling from 2019. The members of the expert profiling staff[7] along with the National Statistics Center will regularly contact and communicate with the headquarters of major enterprise groups[8] through an interactive profiling system and grasp the latest information on the structure and activities of major enterprise groups. The information gathered by the profilers will be promptly recorded in the BR. The role of the profiling staff is not only to gather information from large enterprises but also to reduce their response burden by helping and supporting them to respond to the surveys conducted by the Japanese government.

Figure 2: Image of the Profiling



3. Results

New quality assurance frameworks for the BR in Japan will start from 2019. After conducting the Economic Census 2019 by using the Corporate Number, it is expected that the frame of the BR will be updated in an even more efficient and precise manner. This means that a high-quality BR frame will be available for the next Economic Census for Business Activity 2021 and other official statistics by using the results of the Economic Census 2019. As for the Profiling, it is expected that the accuracy of the data on large enterprises in the BR will improve more than ever. In addition, it is expected that the response burden of the large enterprises will be reduced by helping and supporting them to respond to the surveys conducted by the Japanese government.

[7] It is expected that around fifty staff members will be engaged in the profiling.

[8] It is expected that around three to five thousand enterprises will be the subjects for the Profiling.

4. Discussion and Conclusion

In this paper, two kinds of frameworks which will start from 2019 to ensure the quality of the BR were introduced. By these efforts, it is expected that we can grasp information on enterprises more precisely and contribute to improving the quality and coverage of the BR as the basic data source for the official statistics. On the other hand, we need to continue finding better ways to grasp the status of establishments more precisely, especially in the interim of the Economic Census. The coverage of the administrative data is currently not sufficient to thoroughly update the BR. For example, there is less useful administrative data at the branch offices, and also the data of individual proprietors is difficult to update because individual proprietors are not legally required to complete registration (they have no Corporate Number). In addition, it seems also difficult to acquire information on deaths of businesses because bankrupt companies tend to fail to submit the notification of discontinuance of business. It seems that the impacts for the economy in Japan by branch offices and individual proprietors are very small, but it is important for us to gather all the information on all the establishments located in Japan including branch offices and individual proprietors for providing an accurate BR frame.

References

1. Isao Takabe, Takashi Ioka, (2016). Restructuring the maintenance methods of the Business Register in Japan, 25th Meeting of the Wiesbaden Group on Business Registers 2016.
2. Masatsugu Kitahara, Masao Takahashi (2018). Efforts to Enhance the Quality of the Economic Census in Japan, 26th Meeting of the Wiesbaden Group on Business Registers 2018.

Identification of multiple unusual observations in spatial regression

A.H.M. Rahmatullah Imon[1], Ali S Hadi[2]
[1]Department of Mathematical Sciences, Ball State University, USA
[2]Department of Mathematics and Actuarial Science, The American University in Cairo,Egypt

Abstract

Traditional outlier detection methods cannot be directly applied to spatial data because of its global nature. Spatial outlier detection methods concentrate on discovering neighborhood instabilities [see Shekhar et al. (2002)]. However, most of the traditional detection methods may not accurately locate outliers when multiple outliers exist. Robust spatial z test proposed by Hadi and Imon (2018) has largely resolved this issue. But lots of unresolved issues exist in spatial regression where likewise linear or generalized linear models, the entire inferential procedure is generally affected in the presence of unusual observations called outliers (y-outliers) and high leverage points (x-outliers) or both. A large body of literature are available now for the identification of unusual observations in linear and/or generalized linear regression but this is still an unexplored area in spatial regression. In this paper we propose a new method for the identification of multiple spatial outliers and spatial high leverage points based on robust and clustering algorithms. We also propose a very simple but attractive graphical display to locate these two types of outliers in the same graph.

Keywords

Spatial outlier; Differencing; Masking; High leverage points; Clustering; GP-GSR plot.

1. Introduction

Conceptually spatial outliers are very different from classical outliers. A commonly used definition is that outliers are a minority of observations in a dataset that have different patterns from that of the majority of observations in the dataset. The assumption here is that there is a core of at least 50% of observations in a dataset that are homogeneous (that is, represented by a common pattern) and the remaining observations (hopefully few) have patterns that are inconsistent with this common pattern.

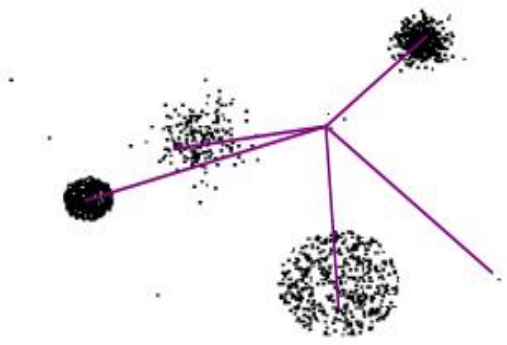


Figure 1. Outliers in data clusters

Spatial outliers are those observations whose characteristics are markedly different from their spatial neighbors. The identification of spatial outliers is important because it can reveal hidden but valuable knowledge in many applications such as identifying aberrant genes or tumor cells, discovering highway traffic congestion points, locating extreme meteorological events such as tornadoes, and hurricanes etc. Although outliers could be easily identified in univariate, bivariate, or even trivariate data through graphical examination of the data, visual inspection does not usually work for more than three dimensions. Not only that automated identification of outliers is tricky even for a two dimensional data if they data form clusters as shown in Figure 1. Here the idea of majority minority simply does not work, bad clusters are identified as outliers [see Hadi et al. (2009)] based on classification techniques. Things could even be cumbersome in regression models where outliers can occur along the ydimension, or along the x-dimension, or both and/or among the relationship between x and y. An excellent review of different aspects of spatial outliers is available in Shekhar et al. [16] and Hadi and Imon (2018). Conceptually, spatial outliers match with outliers in big data and for this reason outlier detection techniques designed for big data are often routinely employed in spatial data. In big data the concept of outlier is local, not global so as in spatial data. The distance and/or density based methods such as k–nearest neighbourhood, local outlier factor (LOF), spatial outlier factor (SOF) methods have become more popular. But all these methods are designed to identify outliers along the y-axis and hence is not readily applicable for spatial regression. For example, temperatures and amount of rainfalls of different regions may vary due to their distances from sea or mountain. Once we fit this relationship by regression we may observe not only strange temperature or rainfall pattern, the distance factor may also be unusual. Attempts have been made to identify outliers based on residuals but it only focuses on the outliers in y, but not in x or both and the whole concept is rather global than local. To overcome this problem in this paper we propose a method which not only

focuses on both x and y dimensions at the same time, but also considers classification techniques to identify outliers.

2. Methodology

Consider a standard regression model

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is an n-vector of observed responses, X is an $n \times p$ matrix representing p explanatory variables including the constant, β is a p-vector of unknown finite parameters and ε is an n-vector of random disturbances with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 I$. The traditionally used Ordinary Least Squares (OLS) estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the vector of fitted values is $\hat{Y} = X\hat{\beta} = HY$. The matrix

$$H = X(X^T X)^{-1} X^T \qquad (2)$$

is often referred to as weight or leverage matrix whose diagonal elements h_{ii} are termed leverages. The OLS residual vector $\hat{\varepsilon}$ is defined as $\hat{\varepsilon} = Y - \hat{Y}$. Observations corresponding to exceptionally large $\hat{\varepsilon}$ values are termed outliers. However, the question still remains how large is large? For this reason we often consider the standardized version of residuals. One very popular choice is deleted Studentized residuals (DSR) defined as

$$t_i = \frac{y_i - x_i^T \hat{\beta}^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}}, i = 1,2,\dots,n \qquad (3)$$

Where $\hat{\beta}^{(-i)}$ and $\hat{\sigma}^{(-i)}$ are the OLS estimates of β and σ respectively with the $i-th$ observation deleted. We call an observation outlier when its corresponding deleted Studentized residual value exceeds 3 in absolute value. Observations corresponding to exceptionally large h_{ii} values are termed high leverage points which are essentially outliers and high leverage points simultaneously rather than separately. Gray (1986) proposed the Leverage-Residual $(L-R)$ plot where the leverage value h_{ii} for each observation i is plotted against the square of a normalised form of its corresponding residual. The bulk of the cases will be associated with low leverage and small residuals so that they cluster near the origin (0,0). The unusual cases will have either high leverages or large residual components and so will tend to be separated from the bulk of the data. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located in the area to the right.

The $L-R$ plot may be effective in the identification of single outlier but it may be ineffective in the presence of multiple outliers unless we remove a group of suspect outliers prior fitting the model. Denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Also suppose that R contains $(n-d)$ cases after $d<(n-k)$ cases in D are deleted. Without loss of generality, assume that these observations are the last d rows of X and Y so that we can partition the matrices as

$$X=\begin{bmatrix}X_R\\X_D\end{bmatrix}, Y=\begin{bmatrix}X_R\\X_D\end{bmatrix}, H=\begin{bmatrix}H_R & H_{RD}\\H_{DR} & H_D\end{bmatrix}$$

where $H_R=X_R(X^TX)^{-1}X_R^T$ and $H_D=X_D(X^TX)^{-1}X_D^T$ and symmetric matrices of order $(n-d)$ and d respectively, and $H_{RD}=X_R(X^TX)^{-1}X_D^T$ is an $(n-d)\times d$ matrix. However, $(X_R^TX_R)^{-1}$ can be expressed as

$$(X_R^TX_R)^{-1}=(X^TX-X_D^TX_D)^{-1}=(X^TX)^{-1}+(X^TX)^{-1}X_D^T(I_D-U_D)^{-1}X_D(X^TX)^{-1} \qquad (4)$$

where I_D is an identity matrix of order d. Using (4), Imon (2002) defined a group deleted version of high leverage points called generalized potentials defined as

$$P_{ii}^*=\begin{cases}\dfrac{h_{ii}^{(-D)}}{1-h_{ii}^{(-D)}} & i\in R\\ h_{ii}^{(-D)} & i\in D\end{cases} \qquad (5)$$

where $h_{ii}^{(-D)}=x_i^T(X_R^TX_R)^{-1}x_i, i=1,2,\dots\dots,n.$ In other words $h_{ii}^{(-D)}$ is the $i-th$ diagonal element of $X(X_R^TX_R)^{-1}X^T$ matrix. The vector of estimated parameters after the deletion of d observations, denoted by $\hat{\beta}^{(-D)}$, is obtained using (4) as

$$\hat{\beta}^{(-D)}=(X_R^TX_R)^{-1}X_R^TY_R=\hat{\beta}-(X^TX)^{-1}X_D^T(I_D-U_D)^{-1}\hat{\varepsilon}_D \qquad (6)$$

where $\hat{\varepsilon}_D=Y_D-X_D\widehat{\beta}$. Using (4), (5) and (6), Imon (2005) introduced a group deleted version of resdiuals called generalized Studentized residuals (GSR) defined as

$$t_i^* = \begin{cases} \dfrac{y_i - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)}\sqrt{\left(1 - h_{ii}^{(-D)}\right)}} & i \in R \\ \dfrac{y_i - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)}\sqrt{\left(1 - h_{ii}^{(-D)}\right)}} & i \in D \end{cases} \quad (7)$$

where $\hat{y}_i^{(-D)} = x_i^T \hat{\beta}^{(-D)}$ and $\hat{\sigma}^{(-D)}$ are the fitted values of $y$and the scale parameter σ respectively after the omission of the suspected outlier group indexed by D. Although the expression of generalized potentials is available for any arbitrary set of deleted cases, D, the choice of such a set is clearly important since the omission of this group determines the weights for the whole set. We call an observation outlier when its corresponding generalized Studentized residual value exceeds 3 in absolute value. No such value exists for generalized potentials. We follow Hadi (1992) to declare an observation as a high leverage point if its corresponding p_{ii}^* exceeds a threshold given as

$$p_{ii}^* > Median\ (p_{ii}^*) + 3\ MAD\ (p_{ii}^*). \quad (8)$$

where MAD stands for the median absolute deviation.

These above results enable us to define a simple graphical display of classifying group deleted leverages and residuals for possible identification of them. Generalized potentials are used as leverages and the generalized Studentized residuals as deletion residuals in a 'generalized potentials – generalized Studentized residuals (GPGSR)' plot. Since the high leverage points need not to be outliers and outliers may not be points of high leverage we may expect different deletion sets D from the computation of these two quantities. Since D is the group of suspected outliers we prefer to include all observations considered to be suspect either along the y dimension or along the x dimension. We employ the blocked adaptive computationally-efficient outlier nominators (BACON) proposed by Billor et al. (2000) as a classifier. Another possibility could be the application of support vector regression for the same, especially when the data is big. The main advantage of the GPGSR plot is that it is suitable for the data where masking (false negative) and/or swamping (false positive) make single case diagnostic plots misleading. This plot, unlike the L-R plot retains the signs of residuals, which can be very important when their interpretation is concerned. Since the bulk of the cases will be associated with low leverage and small residuals, most of the pairs $(t_i^*,\ p_{ii}^*)$ will cluster near the origin $(0, 0)$. The unusual cases will have either high leverages or large residual components and will tend to be separated from the bulk of the cases. High leverage cases will be located at the right corner of the plot and observations with large residuals will be located either

at the upper or lower corner of the plot depending on their signs; large positive outliers will be plotted at the upper corner and large negative outliers will be located at the bottom corner of the plot.

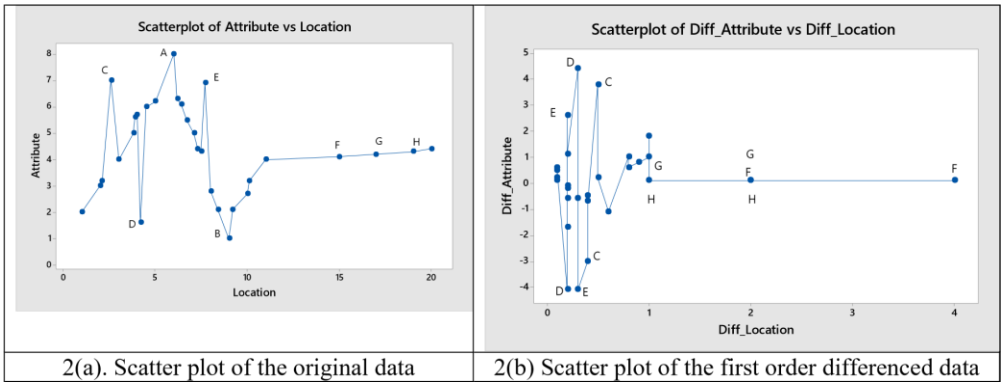
3. Results

In this section we would like to present an example to demonstrate how our proposed method works in the classification of spatial regression outliers in both x and y dimensions. Here we consider a spatial outlier data given by Hadi and Imon (2018) extending the idea of Shekhar et al. (2002). We present the data in Table 1 and also in Figure 2.

Table 1: Hadi and Imon (2018) spatial outlier data

Index	Location	Attribute	Diff_Location	Diff_Attribute
1	1.0	2.0	*	*
2	2.0	3.0	1.0	1.0
3	2.1	3.2	0.1	0.2
4	2.6	**7.0 C**	0.5	**3.8 C**
5	3.0	4.0	0.4	**-3.0 C**
6	3.8	5.0	0.8	1.0
7	3.9	5.6	0.1	0.6
8	4.0	5.7	0.1	0.1
9	4.2	**1.6 D**	0.2	**-4.1 D**
10	4.5	6.0	0.3	**4.4 D**
11	5.0	6.2	0.5	0.2
12	6.0	8.0 A	1.0	1.8
13	6.2	6.3	0.2	-1.7
14	6.4	6.1	0.2	-0.2
15	6.7	5.5	0.3	-0.6
16	7.1	5.0	0.4	-0.5
17	7.3	4.4	0.2	-0.6
18	7.5	4.3	0.2	-0.1
19	7.7	**6.9 E**	0.2	**2.6 E**
20	8.0	2.8	0.3	**-4.1 E**
21	8.4	2.1	0.4	-0.7
22	9.0	1.0 B	0.6	-1.1
23	9.2	2.1	0.2	1.1
24	10.0	2.7	0.8	0.6
25	10.1	3.2	0.1	0.5
26	11.0	4.0	0.9	0.8
27	**15.0 F**	4.1	**4.0 F**	0.1
28	17.0	4.2	2.0	0.1
29	19.0	4.3	2.0	0.1
30	20.0	4.4	1.0	0.1

This example gives a clear distinction between classical outlier and spatial outlier. In Figure 2(a) attribute values are plotted against their locations. For global outliers, traditional statistics will essentially look at the attribute values in the *y* axis and if we do that we observe that the points which are very high such as A or very low such as B. In contrast to that, the spatial outliers are like the spikes C, D, and E. They look like spatial outliers because they violate the law of geography that the nearby things should be very similar. When we take the first order difference of the attributes as shown in Figure 2(b) clearly C, D, and E look very different than their neighbors. It is also interesting to note that the possible global outliers A and B do not look like outliers anymore. In general, we do not search for outliers along the *x*-axis. But when we carefully look at Figure 2(a), we observe that the point F has a marked difference from its neighbors. Points G and H look unusual too. This difference is visible more clearly when we look at the first order difference of the locations as shown in Figure 2(b). Point F now clearly looks like a high leverage point or an outlier along the *x*-space. Points G and H look more extreme as well.



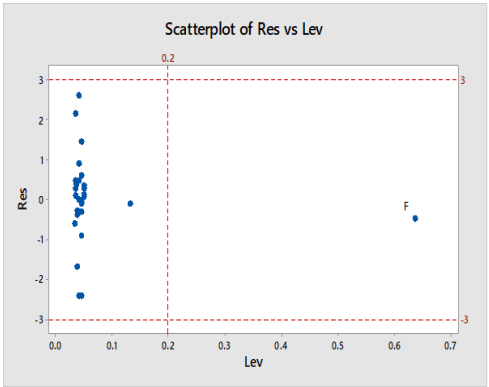
2(a). Scatter plot of the original data | 2(b) Scatter plot of the first order differenced data

Figure 2: Scatter plot of the original and the first order differenced data.

Table 2: Residuals and leverages for Hadi and Imon (2018) spatial outlier data

Index	Del St. Residual	Leverage	GSR	GP
1	*	*	*	*
2	0.45678	0.040885	1.09925	0.06658
3	0.11424	0.051079	0.20420	0.06290
4	2.15139	0.035779	**5.31765 C**	0.03590
5	-1.69711	0.037989	**-4.20445 C**	0.03835
6	0.47421	0.035612	1.13209	0.04571
7	0.32834	0.051079	0.72348	0.06290
8	0.06085	0.051079	0.07645	0.06290
9	-2.41622	0.045639	**-6.03394 D**	0.05185
10	2.61223	0.041275	**6.49375 D**	0.04367
11	0.07639	0.035779	0.07645	0.03590
12	0.89298	0.040885	0.13783	0.06658

Index	Del St. Residual	Leverage	GSR	GP
13	-0.92108	0.045639	2.34566	0.05185
14	-0.10826	0.045639	-2.56908	0.05185
15	-0.33030	0.041275	-0.32208	0.04367
16	-0.28573	0.037989	-0.85865	0.03835
17	-0.32172	0.045639	-0.74085	0.05185
18	-0.05503	0.045639	-0.84428	0.05185
19	1.43538	0.045639	**3.58453 E**	0.05185
20	-2.42202	0.041275	**-6.02091 E**	0.04367
21	-0.39244	0.037989	-1.00843	0.03835
22	-0.62533	0.034647	-1.61273	0.03631
23	0.58737	0.045639	1.39415	0.05185
24	0.26077	0.035612	0.59882	0.04571
25	0.27470	0.051079	0.59176	0.06290
26	0.35834	0.037710	0.84471	0.05471
27	-0.49040	**0.636897 F**	-0.91806	**1.75404 F**
28	-0.12121	0.131865	-0.24012	**0.34271 G**
29	-0.12121	0.131865	-0.24012	**0.34271 H**
30	-0.02276	0.040885	-0.06854	0.06658



3(a). L-R plot | 3(b). GP-GSR plot

Figure 3: Diagnostic plots for the spatial regression data

Now we run a spatial regression of attributes on locations. Since our interest is to understand the neighbourhood instability at consider the first order difference of attributes and locations as given in columns 4 and 5 of Table 1. We then run a regression of diff in attributes on diff in location and the resulting deleted Studentized residuals and leverages are given in columns 2 and 3 of Table 2. Although DSR are very popular outlier measure they fail to identify even a single observation as outlier. Here the cut-off for the leverage is 0.2 and it can identify F as a high leverage point. We see exactly the same picture in the L-R plot as shown in Fig 3(a). Now we compute GSR and GP and the results are presented in columns 4 and 5 of Table 2. We use BACON

classifier to obtain the D set first and then compute GSR and GP as outlined in equations (5) and (7). It is worth mentioning that the cut-off value for GP is 0.1 based on equation (8). We also present the GP-GSR plot for this data in Figure 3(b). These results clearly show the merit of our proposed method. It can successfully identify 3 spatial outliers (C, D, and E) and 3 spatial high leverage points (F, G, H).

4. Discussion and Conclusion

The main objective of our research was to develop a method for the joint identification of outliers and high leverage points for spatial regression. In section 2 we develop a new method to identify both of them and propose a new graphical display called GP-GSR plot to locate both of them in the same graph. In spatial statistics literature observations with neighborhood instability are diagnosed as outliers. For this reason we employ our method on the first order difference of x and y. A numerical example clearly shows the advantage of using our proposed method. It clearly shows that the proposed method can successfully identify outliers and high leverage points simultaneously while the existing methods fail to do so.

References

1. Billor, N., Hadi, A. S., and Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators, *Comput. Statist. Data. Anal.*, 34, 279-298.
2. Gray, J. B. (1984). A simple graphic for assessing influence in regression. *J. Statist. Comput. Simul.* 24, 121134.
3. Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Comput. Statist. Data. Anal.* 14, 1-27.
4. Hadi, A.S. and Imon, A.H.M.R. (2018). Identification of multiple outliers in spatial data, *Int. Jour. Statist. Sci.*, 16, 87-96.
5. Hadi, A. S., Imon, A. H. M. R., and Werner, M. (2009). Detection of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1, 57–70.
6. Imon, A. H. M. R**.** (2002). Identifying multiple high leverage points in linear regression, *Jour. Statist. Stud.*, 22, 207–218.
7. Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression, *J. App. Stat.*, 32, 929–946.
8. Shekhar, S., Lu, C., and Zhang, P. (2002). Detecting graph-based spatial outlier, *Intelligent Data Analysis*, 6, 451-468.

Identification of high leverage points in linear functional relationship model for Big Data

Abu Sayed Md. Al Mamun[1], A.H.M. Rahmatullah Imon[2]
Abdul Ghapor Hussin[3], Yong Zulina Zubairi[4]
[1]University of Rajshahi, Bangladesh
[2]Ball State University, Muncie, USA
[3]National Defence University of Malaysia
[4]University of Malaya, Malaysia

Abstract

Linear functional relationship is having wider applications in statistics because explanatory variables with measurement error are more prevalent in real life problems. So there is a greater scope that unusual errors (outliers) could generate unusual observations in the X-space called high leverage points. High leverage points often exert too much influence and consequently become responsible for misleading conclusion about the fitting of a regression model, causing multicollinearity problems, masking and/or swamping of outliers etc. Although a good number of literature are available on the identification of high leverage points in linear regression model, but this is still an unsolved issue in linear functional relationship model. In this paper, we suggest a procedure for the identification of high leverage points based on group deletion. The usefulness of the proposed method for the detection of multiple high leverage points is studied by some well-known data sets and Monte Carlo simulations. Since our statistic is based on median and median absolute deviation instead of mean and standard deviation respectively it is computationally less extensive and more suitable for big data.

Keywords

Leverages; Masking; Swamping

1. Introduction

The linear functional relationship model (LFRM) is an extension of a linear regression model (LRM) which allows for sampling variability in the measurements of both the response and explanatory variables. In regression the model is poorly fitted because of the presence of outliers. It is a common practice over the years to use residuals for the identification of outliers. Residuals are in fact estimates of the true errors that occur in the Y-space. We anticipate at this point that fitting of the LFRM could be even more complicated because here outliers could occur in the X-space more frequently than the linear regression model. Outliers in the X-space are called high leverage points in the regression literature since they exert too much weight

on the fitting of the model. When we use the ordinary least squares (OLS) or the maximum likelihood (ML) method for fitting a regression line, the resulting residuals are functions of leverages and true errors. Thus high leverage points together with large errors (outliers) may pull the fitted line in a way that the fitted residuals corresponding to those outliers might be too small and this may cause masking (false negative) of outliers. For the same reason the residuals corresponding to inliers may be too large and this may cause swamping (false positive). Peña and Yohai (1995) pointed out that high leverage cases are mainly responsible for masking and swamping of outliers in linear regression. The unfortunate consequences of the presence of high leverage points in linear regression have been studied by many authors. The presence of a high leverage point could increase (often unduly) the value of. Chatterjee and Hadi (1988) mentioned the existence of collinearity-influential observations whose presence could induce or break the collinearity structure among the explanatory variables. Kamruzzaman and Imon (2002) and Imon and Khan (2003a) pointed out that high leverage points may be the prime source of collinearity-influential observations. Imon (2009) pointed out that in the presence of high leverage points the errors not only become heteroscedastic, they might produce big outliers as well. This could make the procedures for the detection of heteroscedasticity very complicated. That is why the identification of high leverage points is essential before making any kind of inference.

In this paper our main objective is to identify high leverage points in a linear functional relationship model. Although some efforts have been done on the identification of outliers and influential observations in LFRM e.g. (Abdullah,1995; Vidal, 2007; and Wellman, 1991), but so far as we know, there is no reported work in the identification of high leverage points in LFRM. Let us consider a simple linear regression model

$$y_i = \alpha + \beta X_i + \varepsilon_i \qquad (1)$$

where y_i is the response, X_i is (supposed) explanatory variable assumed to be constant and specific assumption made on ε_i. We feel that the assumption of X_i being constant in model (1) may not appropriate in reality, instead we introduce a linear functional relationship model.

Consider the following model $y_i = Y_i + \varepsilon_i, \quad x_i = X_i + \delta_i,$

and $$Y_i = \alpha + \beta X_i, \text{ for } i = 1,2,...,n \qquad (2)$$

where the two linearly related unobservable variables X and Y are considered as the true part and the corresponding random variables x and y are observed with random errors δ_i and ε_i. The unobservable X and Y are fixed (nonstochastic) and (2) is called a functional relation. So the main difference

between a LRM and a LFRM is that in LRM it is assumed that the explanatory variable is free from error but in LFRM it is subjected to error.

2. Methodology

In regression analysis it is sometimes very important to know whether any set of *X*-values are exerting too much influence on the fitting of the model. According to Hocking and Pendleton (1983) "*high leverage points are those for which the input vector* x_i *, in some sense, far from the rest of the data."* Let us consider a *k* variable regression model

$$Y = X\beta + \in \tag{3}$$

A set of influential *X*-values is known as a high leverage point. The OLS residual vector can be expressed in terms of the true disturbance vector as

$$\hat{\in} = Y - \hat{Y} = (I - W) \in \tag{4}$$

where the matrix $W = X(X^T X)^{-1} X^T$ given in (4) is generally known as weight matrix or leverage matrix. The weight matrix *W* reflects joint effect of *k* regressors on the fitted responses. Writing the data matrix of *k* explanatory variables as $X = [x_1, x_2, \cdots, x_n]^T$, the *i*-th diagonal element of the weight matrix *W* is defined as $w_{ii} = x_i^T (X^T X)^{-1} x_i$ (5)

For a perfect balanced design, w_{ii} can be written as

$$w_i = \frac{1}{n} + \frac{(x_{i1} - \bar{x}_{.1})^2}{\sum (x_{i1} - \bar{x}_{.1})^2} + \frac{(x_{i2} - \bar{x}_{.2})^2}{\sum (x_{i2} - \bar{x}_{.2})^2} + \ldots + \frac{(x_{ip} - \bar{x}_{.p})^2}{\sum (x_{ip} - \bar{x}_{.p})^2}$$

and thus the diagonal elements w_{ii} of the weight matrix *W* are considered as leverage values, which measure influence of each observation in the *X*-space. A good number of works have been done in the detection of a single high leverage point. It is easy to show that the average value of w_{ii} is k/n, where *k* is the number of the regressors (including the intercept term) and *n* is the total number of observations. Data points having large w_{ii} values are generally considered as high leverage points. Since finding the theoretical distribution of w_{ii} is difficult, all of the high leverage detection techniques are based on rules of thumb. Hoaglin and Welsch (1978) considered observations to be unusual when w_{ii} exceeds $2k/n$ which is also known as the twice-the-mean (2M) rule. Vellman and Welsch (1981) preferred the thrice-the–mean (3M) rule where w_{ii} is considered to be large when it exceeds $3k/n$. Huber (1981) suggested breaking the range of possible values, $(0 \le w_{ii} \le 1)$ into three intervals. Values $w_{ii} \le 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and values above 0.5 should be avoided. Well known Mahalanobis

distances are also suggested to use as measures of leverages in the literature, however, Rousseeuw and Leroy (1987) showed that Mahalanobis distance for each of the points has a one-one relationship with w_{ii} and do not yield any extra information in the leverage structure of a data point. Hadi (1992) pointed out that traditionally used measures of leverages are not sensitive enough to the high leverage points. He introduced a single case deleted leverage measure, named as potential, which is believed to be more sensitive to the high leverage point. Imon and Khan (2003b) showed that in the presence of multiple high leverage points, observations are masked in such a way that even potential values may not focus on all of them. As a remedy to this problem, Imon (2002) proposed generalized potentials for the identification of multiple high leverage points in linear regression. Further developments of the generalized potentials are done by Habshah et al. (2009) and Bagheri et al. (2002). As we already know that in linear functional relation model the explanatory variable is measured with error, or in other words is not fixed, we cannot readily apply the leverage measures discussed in section 2 since they are designed for fixed explanatory variable. In this section we obtain the estimated values of X so that these values can be used as fixed-X in the subsequent studies. Let us assume,

$$E(\delta_i) = E(\varepsilon_i) = 0, \text{var}(\delta_i) = \sigma_\delta^2, \text{var}(\varepsilon_i) = \sigma_\varepsilon^2, \forall i$$
$$\text{cov}(\delta_i, \delta_j) = \text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$
$$\text{cov}(\delta_i, \varepsilon_j) = 0, \forall i, j \quad (6)$$

Model (2) is also known as the unreplicated linear functional relationship when there is only one relationship between the two variables X and Y .There are $(n+4)$ parameters to be estimated, which are $\beta, \alpha, \sigma^2, \tau^2$ and $X_1, X_2, ..., X_n$. Several methods of parameter estimation have been developed (Fuller, 1987) but our primary interest is the maximum likelihood (ML) method. Let (2) and (6) hold, and that δ_i and ε_i are independent normal variables, viz.

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \text{ and } \delta_i \sim N(0, \sigma_\delta^2) \quad (7)$$

Since X_i are non-random variables, $\sigma_x^2 = 0$ and there are $(n+4)$ parameters, namely $\beta, \alpha, \sigma^2, \tau^2$ and the n values of X_i to be estimated. the estimator of σ_ε^2 is derived as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2n}\left[\lambda \sum (x_i - \hat{X}_i)^2 + \sum (y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i)^2\right]$$

which is not consistent. Kendall and Stuart (1979) showed a consistent estimator of σ_ε^2 can be derived by multiplying $\frac{2n}{n-2}$ to (15), that is

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-2}\left[\lambda\sum(x_i - \hat{X}_i)^2 + \sum(y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i)^2\right]$$

and we obtain the estimated values of X as

$$\hat{X}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{(\lambda + \hat{\beta}^2)}$$

where $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and

$$\hat{\beta} = \frac{\left(\sum \hat{X}_i y_i - \hat{\alpha}\sum \hat{X}_i\right)}{\sum \hat{X}_i{}^2}$$

$$= \frac{(\lambda + \hat{\beta}^2)(\lambda S_{xy} + \hat{\beta}S_{yy} + n\hat{\beta}^3\bar{x}^2 + n\lambda\hat{\beta}\bar{x}^2)}{\lambda^2\sum y_i{}^2 + 2\lambda\hat{\beta}S_{xy} + \hat{\beta}^2 S_{yy} + n\hat{\beta}^4\bar{x}^2 + 2n\lambda\hat{\beta}^2\bar{x}^2}$$

gives $S_{xy}\hat{\beta}^2 + (\lambda S_{xx} - S_{yy})\hat{\beta} - \lambda S_{xy} = 0$ and that implies

$$\hat{\beta} = \frac{(S_{yy} - \lambda S_{xx}) + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}}$$

where, $\bar{y} = \frac{\sum y_i}{n}, \bar{x} = \frac{\sum x_i}{n}, S_{yy} = \sum y_i{}^2 - n\bar{y}^2, S_{xx} = \sum x_i{}^2 - n\bar{x}^2$ and $S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$

Identification of High Leverage Points in LFRM

In this section we suggest a procedure for the identification of high leverage points in linear functional relation model. From a set of observed x and y (both assumed to be measured with error), we have estimated the fixed-X

$$\hat{X}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{(\lambda + \hat{\beta}^2)}$$

Since we have a single explanatory variable here, the formula (5) for the computation of leverage values can be simplified as

$$\hat{w}_{ii} = \hat{x}_i^T(\hat{X}^T\hat{X})^{-1}\hat{x}_i = \frac{1}{n} + \frac{\left(\hat{X}_i - \bar{\hat{X}}\right)^2}{\sum_{i=1}^{n}\left(\hat{X}_i - \bar{\hat{X}}\right)^2}$$

Since the above formula contains mean and sum of squares of means which could be very sensitive to high leverage points. For this reason we propose a new formula for the leverages analogous to formula (20), but here the non-robust components are replaced by their corresponding robust alternatives. Hence the formula is

$$\tilde{w}_{ii} = \frac{1}{n} + \frac{|\hat{X}_i - Med(\hat{X}_i)|}{nMAD(\hat{X}_i)}$$

It is easy to show that mean ($\hat{w}_{ii}$) = median ($\tilde{w}_{ii}$) = 2/n.

We consider several measures of the identification of high leverage points, the twice the mean (2M) rule, the thrice-the-mean (3M) rule, and then introduce a new cut-off point. Since it may not be easy to find the theoretical distribution of $\tilde{w}_{ii}$ and often excessive high leverage values can affect measures like mean and standard deviation, we define a confidence bound type cut-off point

$$\tilde{w}_{ii} > \text{Median}(\tilde{w}_{ii}) + 3\ \text{MAD}(\tilde{w}_{ii})$$

which is analogous to forms used by Hadi (1992), Imon (2002,2005) and others.

In this paper, we consider five identification rules which are listed below:

Rule 1 (Classical 2M): $\hat{w}_{ii} > 4/n$

Rule 2 (Classical 3M): $\hat{w}_{ii} > 6/n$

Rule 3 (New 2M based on Median): $\tilde{w}_{ii} > 4/n$

Rule 4 (New 3M based on Median): $\tilde{w}_{ii} > 6/n$

Rule 5 (New Median based Cut-off point): $\tilde{w}_{ii} > \text{Median}(\tilde{w}_{ii}) + 3\ \text{MAD}(\tilde{w}_{ii})$

We compare the performances of the above rules in terms of correct identification of high leverage points and swamping rate of good leverages.

3. Results

We consider a real world data to investigate the performance of our proposed method. In order to make the relationship as model (2), we assume that measurement error can occur in either variable of these two examples. The data is taken from Hand et al. (1994) where the data for 50 results of iron content of crushed blast furnace slag measured by two different techniques, which are chemical test (Y) and magnetic test (X). The X values are estimated by the maximum likelihood formula (17). Now we compute the leverage values for this data set. Here the cut-off point for rule 1 and 3 is 0.08, for rule 2 and 4 is 0.12 and rule 5 is 0.0975 respectively. We observed that the traditional leverage values $\hat{w}_{ii}$ do not identify any high leverage points, but the 2M rule swamps in 6 good cases. The newly proposed leverage measures $\tilde{w}_{ii}$ do not identify any high leverage points but the 2M rule swamps in 1 good case. The 3M rule does not identify any high leverage point for both of these two leverage measures. We observe exactly the same performance from the rule based on the new cut-off point as well. Now we modified the original iron in slag data by inserting few high leverage points. We consider three different

situations. In case 1, 5 low leverage cases (10%) are replaced by high leverage points. In case 2 and case 3 we replace 20% and 30% low leverage points by points of high leverages respectively. Now we compute the leverage values for this data set. Here the cut-off point for rule 1 and 3 is 0.08 and for rule 2 and 4 is 0.12 as they were before. The cut-off points for rule 5 are 0.1052, 0.1024, and 0.0845 for 10%, 20% and 30% high leverage points respectively. We observed from that for the 10% contamination, the traditional leverage values $\hat{w}_{ii}$ can identify high leverage points successfully, but their performances tend to deteriorate with the increase in the level of contamination. For 20% contamination it fails to identify 4 high leverage cases out of 10 and for the 30% contamination it fails to identify 10 out of 15 high leverage points. The newly proposed leverage measures $\tilde{w}_{ii}$ perform very well in this regard. All high leverage points are successfully identified irrespective of the level of contamination.

In this section we report a Monte Carlo simulation which is designed to investigate the performances of different measures of leverages in linear functional relation model. For four different sample sizes, n = 20, 30, 50 and 100, we generated the X values from Uniform (20, 40). Here we consider three different percentages, i.e., 10%, 20%, and 30% high leverage points. The X value corresponding to the lowest high leverage value is then set at 100 and the next values have an increment of 5 each. To generate a model like (2), we then define $x_i = X_i + \delta_i$, where δ_i is $N(0, 1)$. The values of y_i are generated as $y_i = 20 + 2X_i + \varepsilon_i$, where ε_i is also $N(0, 1)$. For each different sample we apply all five leverage identification rules mentioned in section 4 and compute the correct identification rate (IR) and the swamping rate (SR) in terms of percentages. We run 10,000 simulations for each combination. When no high leverage point exists, we observe from the above table that for n = 20, all methods considered in the simulation perform well. However, rule 1, i.e., the traditional leverage measure based on the 2M rule has about 5% swamping rate. The newly proposed rule 4 performs the best as its swamping rate is the lowest followed by rule 2, rule 5 and rule 3. The performance of all these rules tend to improve with the increase in sample sizes but still rule 1 has relatively very high swamping rate which clearly shows that the 2M rule is too prone to declare low leverage points as points of high leverages. In case of 10% high leverages, almost all methods perform very well. Each method maintains 100% identification rate with low swamping rate. Only when n = 100, the identification rate for rule 2 is 90%. But when 20% or 30% high leverage points are present in the data both the 2M and the 3M rule break down. The rule 2, i.e., the 3M performs worst as often its correct identification rate is 0%. The performance of the rule 1 is also poor as it can identify around 13% cases

correctly when there is 30% contamination. The performances of the newly proposed all three rules, i.e., rules 3, 4 and 5 are very satisfactory. They have almost 100% correct identification rate with very small swamping rates, if at all.

4. Discussion and Conclusion

In this paper, our main objective was to propose a method of leverage measures and then to develop an identification rule for the detection of high leverage points in linear functional relationship model. After obtaining a method of finding the fixed-X values, we propose three different identification rules based on robust measures of leverages. Both numerical and simulation results show that the traditionally used measures may often fail to identify even a single high leverage point when 20% to 30% high leverage points are present in the data. The 2M rule based on traditional leverage measure possesses relatively very high swamping rate as well. However, the proposed methods perform very well in every occasion. Our study clearly shows that they can correctly identify all high leverage points without swamping low leverage cases.

References

1. Abdullah, M. B. (1995). Detection of influential observations in functional errors-in- variables model. Communications in Statistics: Theory and Methods. 24:1585–1595.
2. Bagheri, A., Habshah, M. and Imon, A.H.M.R. (2009). Two-step robust diagnostic method for identification of multiple high leverage points. Journal of Mathematics and Statistics. 5: 97–206.
3. Chatterjee, S. and Hadi, A. S. (1988). Sensitivity Analysis in Linear Regression, Wiley, New York.
4. Fuller, W.A. (1987). *Measurement error models*, Wiley, New York.
5. Habshah, M., Norazan, R. and Imon, A.H.M.R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics. 36: 507–520.
6. Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis. 14: 1-27.
7. Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994) A Handbook of Small Data Sets, Chapman and Hall, London.
8. Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. The American Statistician. 32:17-22.
9. Hocking, R.R. and Pendleton, O.J. (1983). The regression dilemma. Communications in Statistics-Theory and Methods. 12: 497-527.

10. Huber, P.J. (1981). Robust Statistics. Wiley, New York.
11. Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression. Journal of Statistical Studies. 3(Special Volume): 207–218.
12. Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression. Journal of Applied Statistics. 32: 929 – 946.
13. Imon, A. H. M. R. (2009). Deletion residuals in the detection of heterogeneity of variances in linear regression. Journal of Applied Statistics. 36:347–358.
14. Imon, A. H. M. R. and Khan, M.A.I. (2003a). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. International Journal of Statistical Sciences. 2:37–50.
15. Imon, A.H.M.R. and Khan, M.A.I. (2003b). A comparative study on the identification of high leverage points in linear regression. Journal of Statistical Studies. 23: 27–32.
16. Kamruzzaman, M. and Imon, A. H. M. R. (2002). High leverage point: Another source of multicollinearity. Pakistan Journal of Statistics.18: 435–448.
17. Kendall, M.G. and Stuart, A. (1979). The Advance Theory of Statistics, Vol.2, Griffin, London.
18. Peña, D.and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. Journal of the Royal Statistical Society Ser- B. 11(57): 18-44.
19. Ryan, T.P. (1997). Modern Regression Methods, Wiley, New York.
20. Rousseeuw, P.J. and Leroy, A. (1987). Robust Regression and Outlier Detection, Wiley, New York.
21. Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics. The American Statistician. 35:234-42.
22. Vidal, I., Iglesias, P. and Galea, M. (2007). Influential observations in the functional measurement error model. Journal of Applied Statistics. 34:1165-83.
23. Wellman, M. J. and Gunst, R. F. (1991). Influence diagnostics for linear measurement error models. Biometrika. 78: 373–380.

Regression analysis of clustered interval-censored failure time data with linear transformation models in the presence of informative cluster size

Hui Zhao[1], Chenchen Ma[2], Junlong Li[2], Jianguo Sun[2]
[1]Central China Normal University, Wuhan 430079, P.R.China
[2]Department of Statistics, University of Missouri, Columbia, Missouri 65211, U.S.A.

Abstract
This paper discusses regression analysis of clustered interval-censored failure time data, which often occur in medical follow-up studies among other areas. For such data, sometimes the failure time may be related to the cluster size, the number of subjects within each cluster or we have informative cluster sizes. For the problem, we present a within-cluster resampling method for the situation where the failure time of interest can be described by a class of linear transformation models. In addition to the establishment of the asymptotic properties of the proposed estimators of regression parameters, an extensive simulation study is conducted for the assessment of the finite sample properties of the proposed method and suggests that it works well in practical situations. An application to the example that motivated this study is also provided.

Keywords
Clustered data; Interval-censoring; Informative cluster size; Linear transformation models; Within-cluster resampling

1. Introduction

This paper discusses regression analysis of clustered interval-censored failure time data, which often occur in medical follow-up studies among other areas (Williamson et al., 2003; Zhang and Sun, 2010). For such data, the failure times of interest are clustered into small groups instead of being independent and also are known only to lie within certain intervals instead of being observed exactly or right-censored. In these situations, sometimes the failure time may be related to the cluster size, the number of subjects within each cluster, too. In other words, in addition to clustering and interval censoring, we may also face or have to deal with informative cluster sizes. In the following, a semiparametric inference procedure is presented for the problem.

Clustered failure time data arise in a failure time study when some failure times of interest are dependent with each other. An example of such data is given by randomized multi-center clinical trials where patients are recruited and grouped by study centers. In these situations, the patients from the same

center may share similar medical environment and thus their failure times may tend to be correlated with each center serving as a cluster. Furthermore, the cluster size, the number of subjects from a center, could be different from one center to another and may contain some relevant information about the failure time of interest. Similar data can occur in a dental study concerning all teeth of an individual (Zhang and Sun, 2010) and such an example will be discussed below in details.

For the analysis of clustered failure time data, a commonly used approach is the marginal model approach in which estimation is usually carried out based on estimating equations-based (GEE) procedures. One major advantage of these methods is their robustness against the misspecification of the correlation structure and also it is relatively easy to use as one can leave the association structure to be arbitrary (Williamson et al., 2003). On the other hand, it is apparent that such methods can be less efficient and more importantly, it is difficult to take into account the informative cluster size. Corresponding to these, we present a within-cluster resampling (WCR) method when the failure time of interest follows a class of linear transformation models (Fine et al., 1998; Zhang et al., 2005). One advantage of these models is their flexibility as they include many commonly used models such as he proportional hazards model and the proportional odds model as special cases. The WCR method uses a single observation to represent each cluster and is a cluster-based approach (Hoffman et al., 2001; Cong et al., 2007; Chen et al., 2016; Chen et al., 2017). Like the GEE-based methods, the new method can be easily implemented and leave the correlation structure arbitrary, and in the meantime, it still works or is valid when the cluster size is informative.

2. Methodology

Consider a failure time study consisting of m clusters and n_i subjects within cluster i. For subject l in cluster i, let T_{il} denote the failure time of interest and suppose that there exists a p −dimensional vector of categorial covariates denoted by $Z_{il}, l = 1, \dots, n_i, \ i = 1, \dots, m$. Some comments on the covariates will be given below. Let $n = n_1 + \cdots + n_m$ and assume that T_{il} follows the linear transformation model given by

$$\mu_0(T_{il}) = Z_{il}^T \beta_0 + \epsilon_{il} \qquad (1)$$

In the above, $\mu_0(\cdot)$ denotes an unknown strictly increasing function, β_0 is a vector of unknown regression parameters, and ϵ_{il} denotes a random error assuming to have a completely known distribution function F. An advantage of the model above is its flexibility as it includes some commonly used models as special cases. For example, it gives the Cox model if $F(t) = 1 -$

$\exp\{-\exp(t)\}$, an extreme value distribution, while one can obtain the proportional odds model by letting $F(t) = \{1 + \exp(-t)\}^{-1}$, the standard logistic distribution. Let Hz denote the distribution function of T given Z. Then it is easy to see that model (1) can be equivalently expressed as $g\big(1 - H_Z(t)\big) = \mu_0(t) - Z^T\beta_0$, where $g^{-1}(s) = 1 - F(s)$.

In the following, for inference about model (1), it will be assumed that one only observes clustered interval-censored data given by $\{O_{il} = (L_{il}, R_{il}, Z_{il}; l = 1, \dots, n_i); i = 1, \dots, n\}$ where $(L_{il}, R_{il}]$ denotes the observed interval for T_{il} as $L_{il} < T_{il} \leq R_{il}$. Also we will assume that the cluster sizes n_i may contain some relevant information about T_{il} or is informative, but given Z_{il}, L_{il} and R_{il} are independent of T_{il} or we have independent interval censoring (Sun, 2006; Zhang et al., 2005). In other words, we have

$$P(T_{il} \leq t | L_{il} = l_{il}, R_{il} = r_{il}, L_{il} < T_{il} \leq R_{il}, Z_{il}) = P(T_{il} \leq t | l_{il} < T_{il} \leq r_{il}, Z_{il})$$

with respect to censoring intervals $(L_{il}, R_{il}]'s$. Furthermore we will assume that the failure times of interest T_{il}'s may be related to the cluster sizes $n_i's$.

Before presenting the proposed WCR estimation procedure, we will first briefly consider univariate interval-censored data or the situation where $n_i = 1$ for all i In this case, for estimation of model (1), note that for any pair (T_i, T_j) we have

$$E\{I(T_i \geq T_j) | Z_i, Z_j\} = \phi\left(\beta^T Z_{ij}\right) \qquad (2)$$

under mode (1), where $Z_{ij} = Z_j - Z_i$ and $\phi(t) = \int_{-\infty}^{\infty}\{1 - F(u + t)\}\, dF(u)$. On the other hand, one can show that

$$E\{I(T_i \geq T_j) | Z_i, Z_j\} = \mathrm{E}\left\{(a_i a_j)^{-1} \int_{L_i}^{R_i} \int_{L_j}^{R_j} I(t_i \geq t_j) dH_{Z_i}(t_j) | Z_i, Z_j\right\}, \quad (3)$$

where H_{Z_i} denotes the distribution function of T_i given Z_i and $a_i = H_{Z_i}(R_i) - H_{Z_i}(L_i)$. Combining (2) and (3), Zhang et al.(2005) suggested that one can estimate β based on the estimating equation

$$U^{(1)}(\beta) = \sum_{i=1}^{m}\sum_{j=1}^{m} \phi'(\beta^T Z_{ij}) \left\{\frac{1}{\hat{a}_i \hat{a}_j} \int_{L_i}^{R_i} \int_{L_j}^{R_j} I(t_i \geq t_j) d\hat{H}_{Z_i}(s) d\hat{H}_{Z_j}(t) - \phi(\beta^T Z_{ij})\right\} Z_{ij} = 0, \quad (4)$$

where $\hat{H}_{Z_i}$ denotes a consistent estimator of H_{Z_i}, the $\hat{a}_j's$ the $a_j's$ with the $H_{Z_i}'s$ replaced by $\hat{H}_{Z_i}'s$ and $\emptyset'(t)$ the first derivative of $\emptyset(t)$ having the form with f denoting the density function of the $\epsilon_{il}'s$. In the above, Since we assume

Z is a categorical variable here, it is natural to obtain the nonparametric maximum likelihood estimator (NPMLE) of H_Z by using the self-consistency algorithm (Turnbull, 1976).

Next, we will generalize the estimation procedure above to clustered interval-censored data and present the WCR estimation procedure. The idea is to randomly select one subject from each of m clusters with replacement and estimate unknown parameters based on the set of m sampled independent subjects. By repeating this process, one can then perform the estimation by using the average of the resample–based estimates. More specifically let B be a pre-specified positive integer, the number of the resampling process described above, and $O^b = \{L_i^b, R_i^b, Z_i^b; i = 1, \dots, m\}$ the independent sample generated from the bth resampling process. Also let $\hat{\beta}_b$ denote the estimator of β given by the solution to estimating equation (4) based on the sample $O^b, b = 1, \dots, B$.

Note that in practice, in addition to estimation of β, one may also be interested in or need estimating the function $\mu_0(t)$. For this, first note that for the T_i^b corresponding to $(L_i^b, R_i^b]$, by the assumptions, we have

$$P\left(T_i^b \le t \middle| L_i^b = l_i^b, R_i^b = r_i^b, L_i^b < T_i^b \le R_i^b, Z_i^b\right) = \frac{I(l_i^b < t \le r_i^b)}{a_i^b} \int_{l_i^b}^{t} dH_{Z_i^b}(s) + I(t > r_i^b)$$

where a_i^b is defined similarly as $a_i, i = 1, \dots, m$. Furthermore, under model (1), one can show that $P(T_i^b \le t | Z_i^b) = F(\mu_0(t) - \beta_0^T Z_i^b)$ and $E_{T|Z}\{I(T \le t)|Z\} = E_{L,R|Z}\{E_{T|L,R,Z}[I(T \le t)|L, R, Z] | Z\}$. These naturally suggest the following estimating equation

$$U_b^{(2)}(\mu(t);\ \beta) = \sum_{i=1}^{m} \left\{ \frac{I(L_i^b < t \le R_i^b)}{\hat{a}_i^b} \int_{L_i^b}^{t} d\widehat{H}_{Z_i^b}(s) + I(t > R_i^b) - F(\mu(t) - \beta^T Z_i^b) \right\} = 0$$

for estimating $\mu(t)$. Here and below, we take $\widehat{H}_{Z_i^b}(\cdot)$ to be the NPMLE of $H_{Z_i^b}(\cdot)$, which can be easily obtained by the self-consistency algorithm (Turnbull, 1976) among other methods. Let $\hat{\mu}_b(t;\beta)$ denote the resulting estimator of $\mu_0(t)$. We propose to estimate β_0 and $\mu_0(t)$ by the following WCR estimators

$$\hat{\beta}_w = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_b \quad \text{and} \quad \hat{\mu}_w(t) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b\left(t; \hat{\beta}_b\right).$$

3. Results

About the theoretical properties of the estimators, we can show that as $m \to \infty$ and under some regularity conditions, $\hat{\beta}_w$ and $\hat{\mu}_w(t)$ are consistent estimators of β_0 and $\mu_0(t)$, respectively. Furthermore we have $\sqrt{m}\left(\hat{\beta}_w - \beta_0\right) \to N(0, \sum_w)$ in distribution.

Now we apply the proposed estimation procedure to a set of clustered interval-censored failure time data from a lymphatic filariasis (LF) study discussed by Williamson et al. (2008) and Zhang and Sun (2013) among others. The LF is a debilitating parasitic disease in which several worms live together in several nests and an effective treatment is expected to kill the worms in all of the nests. The study consists of two treatments, the co-administration of diethylcarbamazine (DEC) and albendazole (ALB) (new treatment) versus DEC alone (standard treatment), and the main goal of the study is to compare their effects on the treatment of LF. The drug ALB is an anti-parasitic one, manufactured by GlaxoSmithKline, and it is commonly used to treat interstima worm infections. When coadministered with DEC, it helps break to the cycle of LF transmission between mosquitoes and humans, and by using ultrasound, the doctor can detect the movement of the living adult worms.

The study followed 47 patients, 22 given the new treatment and the other 25 given the standard treatment, for a year since their treatments and they were periodically examined by ultrasound to see if the worms were still alive. Thus with respect to the times to the clearance of the worms in each nest, the variables of interest, only clustered intervalcensored data were observed with each patient serving as a cluster and the cluster size being the number of nests of adult filial worms in the patient. In addition, these times to the clearance may be correlated to the number of nests as pointed out by Williamson et al. (2008) and can be seen from Table 1, reproduced from Williamson et al. (2008), which gives the average percentages of the nests cleared during the one year follow-up for all study individuals. It is apparent that the time to clear the worms seems to be positively correlated with the number of nests or the cluster size. In total, 78 adult worm nests were detected by ultrasound with the cluster size n_i ranging from 1 to 5.

Table 1. Percentages of nests cleared in the LF study during the follow-up

Number of Nests	Percentage Cleared
1	81.8
2	62.5
3	50.0
4 or 5	33.3

Table 2. Estimated treatment effects for the LF Data

Model	# of within-cluster resamples	Estimate	SE	p-value
Cox model	B=40	-0.5052	0.3148	0.1085
	B=80	-0.4940	0.3145	0.1163
	B=160	-0.4997	0.3175	0.1154
Probit model	B=40	-0.6279	0.4770	0.1805
	B=80	-0.6319	0.4836	0.1913
	B=160	-0.6202	0.4573	0.1750

For the analysis, define Z_i to be 0 if subject i was given the new treatment and 1 otherwise. Note that here we only have cluster-specific covariates. Table 2 contains the results obtained by the application of the proposed estimation procedure and includes the estimated treatment effect on the time to the clearance of the worms, the estimated standard error (SE), and the $p-$values for testing the covariate effects equal to zero. They suggest that there seems no significant difference between the two treatment groups. Williamson et al. (2008) and Zhang and Sun (2013) gave similar conclusions. Note that here we used different B values but the results seem to be robust. On the other hand, one may be careful about the conclusions due to the small number of subjects.

4. Discussion and Conclusion

A main feature of the models considered is their generality and flexibility as they allow one to describe covariate effects in various forms. For inference about regression parameters, a WCR-based estimating equation approach was presented, and although the method may be computationally intensive, it is highly intuitive and can be easily implemented. Also similar to the partial likelihood approach, the proposed method has the advantage that it does not require the estimation of the nonparametric function involved.

In the above, the focus has been on regression parameters, but sometimes one may be interested in making inference about the unknown function $\mu_0(t)$ too. One such situation is when the survival prediction is of interest. On the other hand, the derivation of the limiting distribution of $\hat{\mu}_w(t;\hat{\beta}_b)$ is quite challenging even if under right censoring mechanism. A main reason for this is that the estimator $\widehat{H}_Z(t)$, the NPMLE of $H_Z(t)$, used above has a non-normal limiting distribution only with a convergence rate of $m^{1/3}$. Thus it is reasonable to postulate that the estimator $F(\hat{\mu}_b(t;\hat{\beta}_b) - \hat{\beta}_b^T Z_i^b)$ also has a very complicated asymptotic distribution with a convergence rate of $m^{1/3}$.

References

1. Chen, L., Sun, J., Xiong, C. (2016). A Multiple imputation approach to the analysisof clustered interval-censored failure time data with the additive hazards model. *Computational Statistics and Data Analysis*, 103, 242-249.
2. Chen, L., Feng, Y., Sun, J. (2017). Regression analysis of clustered failure time datawith informative cluster size under the additive transformation models. *Lifetime Data Analysis*, 23, 651-670.
3. Cong, X., Yin, G., Shen, Y. (2007). Marginal analysis of correlated failure time datawith informative cluster sizes. *Biometrics* 63, 663–672.
4. Fine, J.P., Ying, Z., Wei, L.J. (1998). On the linear transformation model for censoreddata. *Biometrika* 85, 980–986.
5. Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failuretime data. *Biometrics* 42, 845–854.
6. Hoffman, E.B., Sen, P.K. Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika* 88, 1121–1134.
7. Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.
8. Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped,censored and truncated data. *Journal of the Royal Statistical Society, Series B* 38, 290– 295.
9. Williamson, J.M., Datta S., Satten, G.A.(2003). Marginal analyses of clustered datawhen cluster size is informative. *Biometrics* 59, 36–42.
10. Williamson, J.M., Kim, H.Y., Manatunga, A., Addiss, D.G. (2008). Modeling survival data with informative cluster size. *Statistics in Medicine* 27, 543–555.
11. Zhang, X., Sun, J. (2010). Regression analysis of clustered interval-censored failuretime data with informative cluster size. *Computational Statistics and Data Analysis* 54, 1817–1823.
12. Zhang, X., Sun, J. (2013). Semiparametric regression analysis of clustered intervalcensored failure time data with informative cluster size. *The International Journal of Biostatistics* 9, 205–214.
13. Zhang, Z., Sun, L., Zhao, X., Sun, J. (2005). Regression analysis of interval-censoredfailure time data with linear transformation models. *The Canadian Journal of Statistics* 33, 61–70.

Comparative analysis of methodologies used for collecting data on formal and informal employment and tertiary education efficiency

Maria Frolova
FrolovaEDU company

Abstract

The rates of economic growth in many aspects depend on the quality of human capital and the match of professional training accomplished by educational institutions to the labour market demand. The last factor is the reason for change of share of informal employment in total employment. The phenomenon of highly educated informal employees is broadly spread in developed countries. For operation management and efficient control of investments made into education sector on macro and micro level by government, companies and households we need to have the up-to-date data on the matches and mismatches of the employers' demand and professional qualification of the manpower. The goal of the research is to evaluate the efficiency of tertiary education system in meeting the demands of labour market with use of indicators based on the procedures of statistical matching and integration the results of different surveys and administrative data. In the article the author makes comparative analysis of the methodologies developed and used to collect data on tertiary education and to measure skills and qualification mismatches of persons in employment. The guidelines, regulations, definitions and data of International Labour Office (ILO) and International Standard Classification of Education (ISCED) 2011 are used.

Keywords

Labour market demand; Qualification mismatch; Skills mismatch; Labour statistics; Education statistics

1. Introduction

In modern global and digital world the speed of economy development in countries, companies and households is closely related with the talents, skills and personal development speed of human capital. On one hand, company performance fully depends on qualifications and skills of manpower entering labour market, they need employees able and ready to accomplish and support their current activities, so we need the full match of company demands and skills and qualifications of graduates. Underqualified and underskilled personnel always imply stagnation/slow down/unemployment or lead to extra company/personal costs for obtaining the necessary level for implementing company plans/employment.

On the other hand, with overeducated and overskilled human capital we meet both benefits and risks:

1. only overeducated and overskilled professionals can make the breakthrough;
2. lost productivity and lower growth for companies is possible in case overeducated and overskilled workers lose job satisfaction and motivation, because of lower wages or unrealized expectations.

The phenomenon of highly educated informal employees is broadly spread in developed countries. Currently we observe the change of the share of informal employment in total employment, including both "informal employment in the formal sector of economy" and "employment in the informal sector".

The rates of economic growth in many aspects depend on the quality of human capital and the match of professional training accomplished by educational institutions to the labour market demand. Valentina Stoevska, Department of Statistics, International Labour Office (ILO) distinguishes the following negative consequences and potential cost that the situation with persistent qualification and skill mismatches can lead to: "For workers (for overeducated and overskilled) - lower wages lower job satisfaction, loss of motivation, higher on-the-job search, higher the risk of being out of employment, unrealized expectations, lower returns on investment in education. For employers – lost productivity, increased absenteeism, higher turnover, lower growth, less innovation. For society - wasted education costs, higher unemployment benefits, lost income tax revenues. Total cost depends on the number of mismatched individuals, the type and severity of mismatches."

For operation management and efficient control of investments made into education sector on macro and micro level by government, companies and households we need to evaluate the efficiency of tertiary education programs in relation to meeting labour market challenges. To accomplish this goal the following steps were suggested and made by the author:

1. the study of methodologies used for collecting data on the matches and mismatches of the employers' demand and professional qualification of the manpower;
2. the study of methodologies used for collecting data on tertiary education.

2. Methodology

2.1. In 2018 International Labour Office (ILO) introduced draft Guidelines concerning measurement of qualifications and skills mismatches of persons in employment, represented during the 20th International Conference of Labour Statisticians in Geneva. The suggested approach is based on methodological

work undertaken in a number of member countries, Organization for Economic Cooperation and Development (OECD) and the European Centre for the Development of Vocational Training, existing standards related to labour statistics and education statistics.

The Guidelines distinguish between Qualifications and Skills, there are used the definitions made by UNESCO in International Standard Classification of Education (ISCED) 2011 and International Standard Classification of Education: Fields of Education and Training 2013 (ISCED-F 2013).

Qualification is defined as the official confirmation usually in the form of a document, obtained through (i) successful completion of a full education program; (ii) successful completion of a stage of an education program (intermediate qualifications); or (iii) validation of acquired knowledge, skills and competences, independent of participation in an education program (acquired through non-formal education or informal learning). For formal qualifications the data is collected about the level of education and field of study. Non-formal qualifications are not officially recognized as equivalent to formal qualifications.

Skills are defined as the innate or learned ability to apply knowledge acquired through experience, study, practice or instruction, and to perform tasks and duties required by a given job. The skills are classified in the following way:

- job specific/technical skills that relate specifically to certain types of jobs or job fields, it is difficult to transfer them from job to job;
- basic skills are the prerequisite for further education and training, and for acquiring transferable and technical and vocational skills;
- transferable/portable skills are relevant to a broad range of jobs and occupations and can be easily transferred from one environment to another.

In ISCO 08, skill specialization is considered in terms of four concepts: the field of knowledge required, the tools and machinery used, the materials worked on or with: and the kinds of goods and services produced. Due to job requirements different levels of skills proficiency are required: low level, moderate level, high/advanced level, none.

ILO methodology suggests classifying the mismatch of persons in employment separately in qualifications and skills.

Due to ILO definition qualification mismatch refers to a situation in which a person in employment, during the reference period, occupied a job whose qualification requirements did not correspond to the level and/or type of qualification they possessed. Over-qualification and under-qualification due to the level of study and field of study mismatch could be observed. Skill mismatch refers to a situation in which a person in employment, during the

reference period, occupied a job whose skills requirements did not correspond to the skills they possesses.

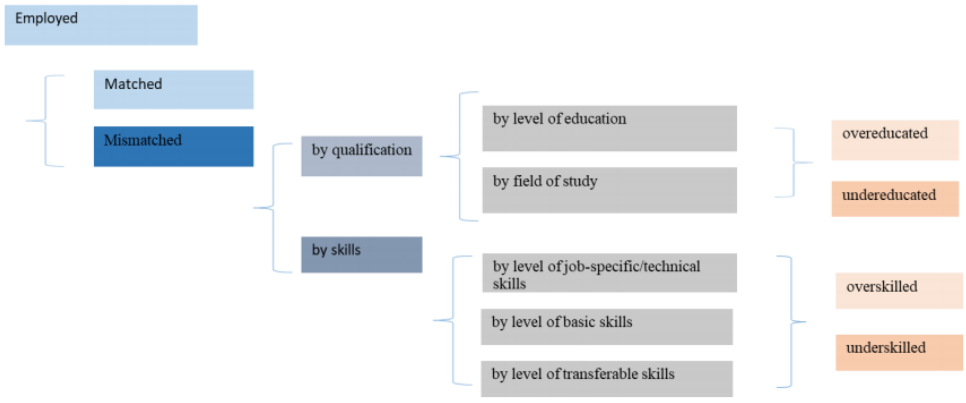


Fig.1 Match and mismatch of persons in employment, International Labour Office Guidelines 2018

Mismatch could be observed of overall skills or types of skills due to the classification represented in fig. 1. Two mismatch levels are distinguished: overskilling, when the level of skill is higher and underskilling, when level of skill is lower.

The measurement of both qualification and skills mismatches is recommended to be based on the data from household/establishment surveys, administrative records, secondary sources.

For measurement the qualification mismatch both by level of education and by field of study it is suggested to use normative, statistical or subjective approaches. The highest level of educational attainment of a person in employment is considered as the educational level.

Due to normative approach the thresholds used as a boundary between matched and mismatched are determined based on educational requirements as specified in relevant legislation or national practice. For statistical approach the thresholds are empirically determined on the basis of the modal level of education/field of study of all persons in employment in an occupation or occupational group. For subjective approach the thresholds are determined on the basis of the modal value of the self-assessed level of education /field of study required to perform the job by all persons employed in a given occupation or occupational group.

Subjective approach could be direct and indirect. Direct approach: a person in employment is considered to be overeducated/undereducated if they report having a level of education that is higher/lower than that required to perform their current job. Indirect approach: a person in employment is considered to be overeducated/undereducated if their level of education is above/below the modal value of the self-reported level of education

appropriate to get the job or to perform the job reported by all workers in the same occupation or occupational group.

For measuring the skills mismatch we compare the skills required for competent job performance and skills possessed by a person in employment. The mismatch can be measured for:

1. specific types of skills (skill as multi-dimensional concept), when a person in employment is overskilled/undeskilled if the level of specific type of skills required to do their job are lower/higher than the level of skills they possess.
2. overall skills (skill as uni-dimensional concept), when a person in employment is overskilled if they assess having the skills to perform more complex tasks or underskilled if they report that, for competent performance at the job, some of their skills need to be further developed.

ILO Guidelines 2018 suggest the following basic indicators for reporting labour underutilization related to the inadequate use and mismatch of qualifications and skills of persons in employment:

- persons in employment mismatched by level of education, over and undereducated,
- persons in employment mismatched by field of study,
- persons in employment mismatched by both level of education and field of study,
- persons in employment mismatched by technical skills, over and underskilled,
- persons in employment mismatched by basic skills, over and underskilled,
- persons in employment mismatched by transferable skills, over and underskilled.

To understand the relationship between qualification and skills mismatches, it is suggested to identify separately and report headcounts and rates for the groups:

- persons in employment undereducated but matched/mismatched by type/level of skills,
- persons in employment overeducated but matched/mismatched by type/level of skills,
- persons in employment matched by level of education but matched/mismatched by type/level of skills,
- persons in employment mismatched by field of study but matched/mismatched by type/level of skills.

2.2 Education statistics is collected due to the methodology provided by International Standard Classification of Education (ISCED) 2011, the standard framework developed by UNESCO and used to categorize and report cross-

nationally comparable education statistics. The UNESCO Institute for Statistics and UNESCO-OECD-Eurostat (UOE) data collection programs, as well as education statistics of UNESCO member states are collected according to the suggested standards. Education programs and qualifications are organized by education levels and fields. ISCED level reflects the degree of complexity and specialization of education program content. Statistics on educational programs provides information on the links between inputs (entrants into the system), the process (participation) and the output (the qualification). Due to this approach the qualification becomes the sign that the program graduate possesses the level of professional skills enough to start professional career.

As we are interested in labour market analysis, the focus of our research is methodology used to collect the data on tertiary education that includes both academic and vocational (professional) education. The main actors in tertiary education sector responsible for the professional training of human capital are institutions and companies providing formal, non-formal education and informal learning service to population. Formal education is provided by public organizations and recognized private bodies. Vocational education is often recognized as a part of formal education system. Programs that take place partly in the workplace are also considered as formal education if they lead to a qualification recognized by national education authorities. Apprenticeships – programs provided in cooperation between educational institutions and employers.

Table 1 Tertiary education structure, formal education programs

Level	Program	Orientation
Level 5	Short cycle tertiary education	General Vocational
Level 6	Bachelor's degree	Academic
Level 7	Master's degree	Professional
Level 8	Doctor's degree	

Table 1 represents ISCED levels 5, 6, 7 and 8: short-cycle tertiary education, Bachelor's or equivalent level, Master's or equivalent level, and doctoral or equivalent level. Labour market entry is possible from every level, or the program is used as a pathway to other tertiary education programs. At level 2 the following types of orientation are possible: general and vocational. At levels 6-8 the terms academic and professional are used correspondingly.

Vocational education is defined as education programs designed for learners to acquire the knowledge, skills and competences specific to a particular occupation, trade or class of occupations or trades. Successful completion leads to labour market-relevant, vocational qualifications, acknowledged by national authorities or labour market. General education is defined as education programs designed to develop learners' general

knowledge, skills and competences. General education includes education programs that are designed to prepare participants for entry into vocational education but do not prepare for employment in a particular occupation or trade.

Non-formal education is provided by an educational provider. It is an addition, alternative and/or complement to formal education. It may not have a continuous pathway structure, typically in the form of short-courses, workshops or seminars. Though the qualifications of non-formal education programs are mostly not recognized as formal qualifications, they are more focused on life and work skill training, social or cultural development. Non-formal education includes training in a workplace to improve and adapt existing qualifications and skills and training for unemployed or economically inactive persons. ISCED 2011 does not give methodology for organizing mapping for non-formal programs and qualifications, recommends to use the criteria of content equivalency.

Informal learning and incidental or random learning are not measured by ISCED. Informal learning includes intentional/deliberate learning activity that is not institutionalized, it occurs in the family, workplace, local community and daily life. Incidental or random learning occurs as a by-product of day-to-day activities, event or communication. In professional context both informal and incidental/ random learning have high efficiency, because of high personal motivation.

The relationship between tertiary education sector and labour market is represented in fig. 2.

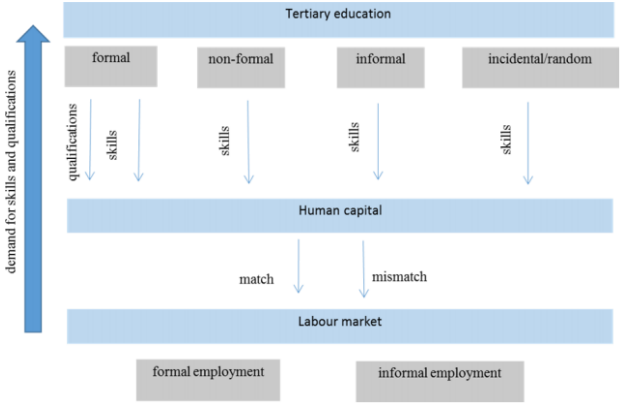


Fig 2. Interrelation between tertiary education sector and labour market

3. Results

3.1 Currently ILO Draft Guidelines concerning measurement of qualifications and skills mismatches of persons in employment 2018 do not make difference between formal and informal employment. Relevant information on quality and skill mismatch of people with informal employment

could be from interviewing the households. Direct measurement of employment in the informal sector as well as informal employment is low among the developed countries. Though informal employment is difficult for measuring, the ILO notes the importance of the topic and the increasing demand for information by many users, it is important to understand the contribution to GDP of the informal sector.

3.2 International data collection on education is mainly focused on formal education. The heterogeneity of non-formal education makes it difficult to develop general guidelines for data collection and analysis. Non-formal education programs have great influence on labour market in terms of skills, as the programs are focused less on theoretical learning and more on skill training.

3.3 ISCED is designed to classify the activities of educational institutions due internationally agreed categories, it has no focus to assess the competences of individuals. It collects data on participants, entrants, graduates and educational attainment due to educational level, orientation and field of study. The methodology cannot be used for evaluating the efficiency of educational programs in relation to measuring the skills obtained, only for the qualifications.

3.4 To evaluate the efficiency of tertiary education system in meeting the demands of labour market we suggest to add the following information to the survey measuring the skill level of the person in employment: with every skill that the respondent assesses he puts information about the place/program/institution where he developed it. Not only this data will help to understand the efficiency of educational programs, but also to compare the efficiency of formal, non-formal and informal educational sectors.

3.5 If the survey contains information about formal/informal employment of the respondent, it will be possible to make the analysis of professional skills to be developed when switching from one employment category to another.

4. Discussion and Conclusion

We are living in the age when in developed countries employment can be characterized by two major factors: 1. in comparison to previous generations the working period in the life of the person has increased (we observe both the shift of retirement age and a big share of people continue working after retirement both in formal and informal sectors); 2. in situation of digital revolution there is a transformational change in the skills demanded by labour market. As the speed of changes is increasing, people are having more than one job during their life. In this situation both for the economy of countries and households it is important that people are ready and have access to efficient life-long educational programs providing the good return on investments. For this it is crucially important to have the up-to-date picture on

labour market demand and tertiary education efficiency. When developing the methodology for data collection and analysis we should take into account that we deal with a complex situation, where, on one hand, there are formal and informal employment opportunities, on the other, we come across various educational opportunities in formal, non-formal and informal learning programs.

References

1. Johan G. Wissema. (2009). Towards the Third Generation University. Edward Elgar.
2. Klaus Schwab. (2016). The Fourth Industrial Revolution. World economic forum.
3. International Standard Classification of Education 2011. UNESCO Institute for Statistics.
4. Draft guidelines concerning measurement of qualifications and skills mismatches of persons in employment. 20th International Conference of Labour Statisticians, 10-19 October 2018, Geneva.

Suggestions for improving the methodology for estimating informal employment based on a sample survey of the labour force

Musikhin Sergey
Analytical Centre by Government of Moscow, Moscow, Russian

Abstract
Assessment of the impact of the variability of the annual sex-age structure of the population on the final indicators of informal employment. The prospect of combining methodological approaches to the population.

Keywords
Indicators of informal employment; distribution methodology for sample surveys (LFS); official statistics.

1. Introduction

Informal (shadow) employment is a type of employment in the informal economy, when the fact of establishing labour relations between an employee and an employer is hidden from the official authorities. Usually, these relationships are hidden by the initiative of the employer or employee in order not to pay taxes or circumvent this or that law. In this case, the calculation is usually made in cash, often the employer is not interested in the past of the employee and his documents.

One of the elements of informal employment is the number of workers who perform additional work in the main place of work with official registration, but without official supplements for overtime work. Also to informal employment include semi-formal labour relations, which are formed with the accrual of the minimum wage to the employee and compensation in the "envelope" full equivalent of the value of his work.

In Russia, as in most countries, the government pays great attention to the legalization of informal employment, which requires reliable statistical evaluations of this employment category.

2. Methodology

In the Russian Federation, methods for calculating one component of informal employment - "Number of employees not reflected in the reports of organizations" are presented in the methodology for forming the balance of time spent. The calculation of this indicator is based on a comparison of the indicator "Number of employees of enterprises", obtained from the results of the LFS (Labour Force Survey) [Hussmanns R. (2004)], and the same indicator obtained from the data of the mandatory state statistical reporting enterprises.

The method stipulates that there are methodological differences between these indicators, and additional work is required to bring them into a comparable form. It also indicates the need to adjust the number of employees who worked part-time in accordance with the terms of the contract, as well as adjusting the structure of the population coverage and taking into account interregional labour migration.

Also, a similar method of calculation is described in the average monthly income of employment [Zarova E. & Musikhin S. & Smelov P. (2016); Laikam K.E et al. (2017)].

My research is devoted to solving the methodological question of determining informal employment according to a sample survey of labour related to the distribution of the survey results. The ILO recommendations on conducting LFS recommend a direct question, allowing to reveal the formality of the employment relationship between the employee and the employer. Based on the answer choices, it is possible to form the indicator "The number of employees working in organizations by verbal agreement (component of informal employment No2), people".

Knowing the features of the distribution of sample survey data on the demographic structure of the population of last year, I set a goal to study the impact of this methodological assumption and myself distributed the survey results to the current demographic structure of the population. I suggested that due to the high difference in the adjacent years of the sex-age pyramid in the cohort groups of the population of the Russian Federation (Fig. 1, [Website Rosstat]), the LFS results should be weighed on the demographic structure set at the beginning of the current year. This procedure is described in the methodological guidelines of international statistical services [ILO (2018), p.28; ILO (2017), p.8]. However, at present, Russian official statistics use the demographic structure at the beginning of the year preceding the year of the survey in order to promptly weigh the results of the LFS without waiting for the results of the demographic statistics.

The Number and Composition of the Population of Russia by Sex and Age in 2017, people

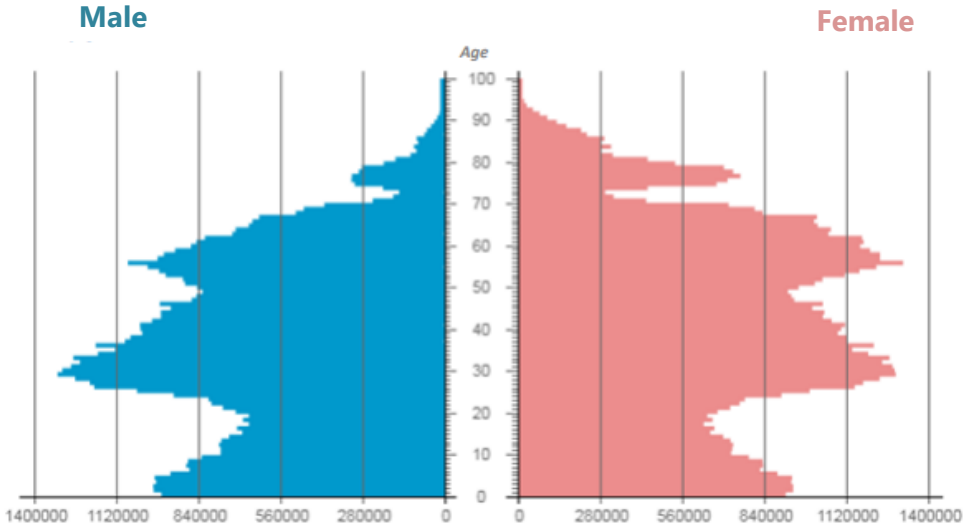


Figure 1. - The demographic structure of the population of the Russian Federation for 2017.

The tables below show the number of employees working in enterprises by verbal agreement (component of informal employment), estimated according to the current methodology and proposed by the author. (see Table 1)

Table 1 –Number of employees of enterprises by verbal agreement (informal employment components No2), person

Five year population group	2014 year		2015 year		2016 year	
Invoice method	**t-1**	**t**	**t-1**	**t**	**t-1**	**t**
Total	**286 272**	**284 110**	**279 947**	**276 966**	**233 623**	**230 666**
From 15 to 19 years	5 899	5 727	8 920	8 605	5 501	5 433
From 20 to 24 years	48 838	45 119	43 536	40 222	33 247	30 334
From 25 to 29 years	49 921	49 948	47 773	47 523	37 239	36 463
From 30 to 34 years	37 564	38 718	37 185	37 870	33 885	34 281
From 35 to 39 years	31 957	32 464	36 035	36 380	27 170	27 493
From 40 to 44 years	29 928	30 353	23 121	23 646	22 293	22 533
From 45 to 49 years	24 778	23 865	20 092	20 248	18 990	19 007
From 50 to 54 years	23 551	23 085	23 792	22 868	19 091	17 979
From 55 to 59 years	19 597	20 092	20 092	19 510	18 869	19 132
From 60 to 64 years	9 619	10 060	10 060	13 574	12 968	13 210
From 65 to 72 years	4 622	4 680	4 680	6 520	4 368	4 802

Invoice method: t-1 - Estimate applied by Rosstat, obtained by extending the respondents' answers to the demographic structure of the population as of January 1 of the year preceding the observation.

t - Estimate obtained by distributing the respondents' answers to the demographic structure of the population as of January 1 of the year of the survey (used in ILO methodological recommendations)

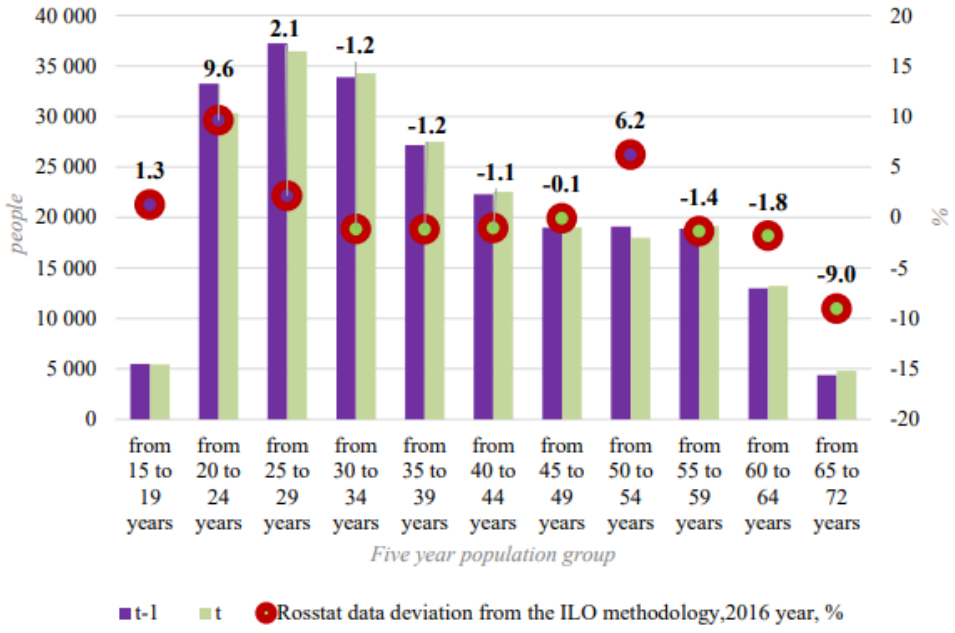


Figure 2. - Number of employees of enterprises by verbal agreement - Rosstat data deviation from the ILO methodology (%), 2016 year, people and percent

The official estimate of the number of employees of enterprises by verbal agreement exceeds the average estimates by 1%, which I obtained using the methodological approach of weighting by the demographic structure of the current year. In absolute terms, the differences in assessment are in 2014 - 2,162 people, in 2015 - 2,980 people, in 2016 - 2,957 people.

Table 2 - Structure of the number of employees of enterprises by verbal agreement (informal employment components),%

Five year population group	2014 year		2015 year		2016 year	
Invoice method	**t-1**	**t**	**t-1**	**t**	**t-1**	**t**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
From 15 to 19 years	2.1%	2.0%	3.2%	3.2%	2.4%	2.4%
From 20 to 24 years	17.1%	15.9%	15.6%	15.6%	14.2%	13.2%
From 25 to 29 years	17.4%	17.6%	17.1%	17.1%	15.9%	15.8%
From 30 to 34 years	13.1%	13.6%	13.3%	13.7%	14.5%	14.9%
From 35 to 39 years	11.2%	11.4%	12.9%	13.1%	11.6%	11.9%
From 40 to 44 years	10.5%	10.7%	8.3%	8.5%	9.5%	9.8%
From 45 to 49 years	8.7%	8.4%	7.4%	7.3%	8.1%	8.2%
From 50 to 54 years	8.2%	8.1%	8.5%	8.3%	8.2%	7.8%
From 55 to 59 years	6.8%	7.1%	6.9%	7.0%	8.1%	8.3%
From 60 to 64 years	3.4%	3.5%	4.8%	4.9%	5.6%	5.7%
From 65 to 72 years	1.6%	1.6%	2.2%	2.4%	1.9%	2.1%

Invoice method:

t-1 - Estimate applied by Rosstat, obtained by extending the respondents' answers to the demographic structure of the population as of January 1 of the year preceding the observation.

t - Estimate obtained by distributing the respondents' answers to the demographic structure of the population as of January 1 of the year of the survey (used in ILO methodological recommendations)

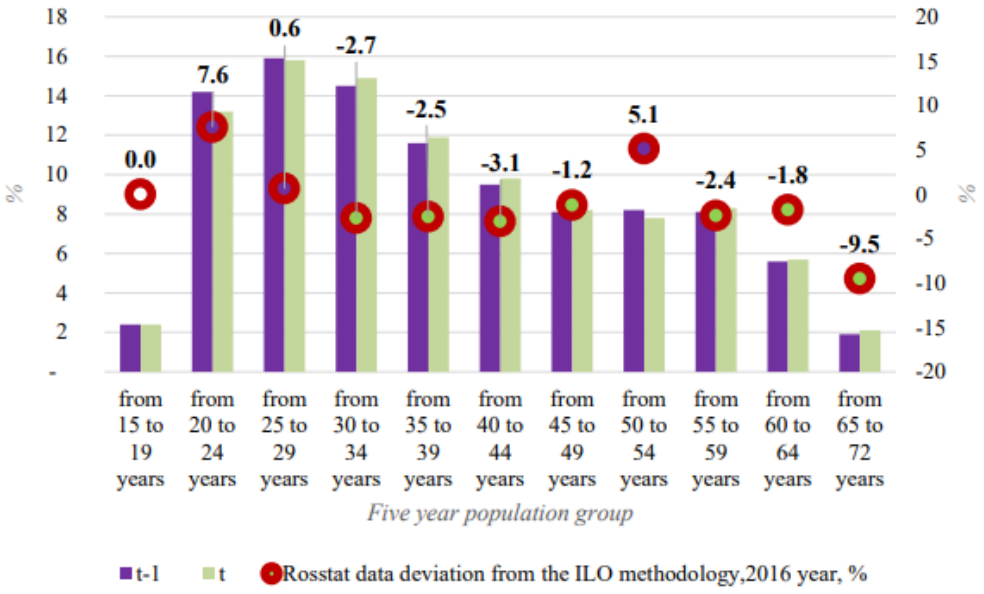


Figure 3. – Structure the number of employees of enterprises by verbal agreement - Rosstat data deviation from the ILO methodology (%), 2016 year, percent

For the majority of Russian regions, Rosstat's estimates are also more optimistic than those I received - in 2016, there were 69 of 85 of these regions of the Russian Federation. It should also be noted that Rosstat's estimates in 12 regions are overestimated by more than 3%. The rest of the regions of the Russian Federation, in which the Rosstat estimates are more pessimistic than I received, are few in numbers, where this phenomenon cannot be characterized in terms of enough representativeness. (with the current level of population coverage in this survey of Rosstat and low prevalence (or detectability) of this phenomenon in the whole country)

3. Results

As a result of the analysis, we can conclude that the unacceptability use of the counting method according to the demographic structure of last year for the purpose of forming the indicator "Number of employees of enterprises by verbal agreement (informal employment components), people" at regional level. This assumption is strongly distorts the results of the LFS. It is advisable to use new weights of distribution of LFS data immediately after summing up the demographic structure of the population at the beginning of the current survey year. It will also provide consistent assessment of the labour force survey at the country level with the methodology adopted by the International Labour Organization.

4. Discussion and Conclusion

The issue of disseminating LFS data was raised on the methodological councils of Rosstat, when discussing indicators of interregional labour migration in September 2017. But this question did not find support from the methodological community. The question of the speed of publication of statistical information and the consistency of the initial estimates throughout the year in terms of the "employment level" and "unemployment rate" were more important than the objectivity and accuracy of the methodological coordination of the data with the final ILO recommendations

References

1. ILO (2018). Global Estimates for International Migrant Workers - Results and Methodology. 2nd ed. International Labour Office - Geneva: ILO.
2. ILO (2017). Labour Force Estimates and Projections: 1990-2030. 2017 ed. Methodological description. International Labour Office - Geneva: ILO, 11.2017.
3. Laikam K.E et al. (2017). Reasoning Behind Changes in Calculation of the Average Monthly Accrued Wages of Employees. Laikam K.E., Zarova E.V., Zainullina Z. Zh., Ryzhikova Z.A., Musikhin S.N. Voprosi Statistiki №6 – Moscow, Russian Federation, 2017, p.3-8.

4. Zarova E. & Musikhin S. & Smelov P. (2016). Statistical Resources and Methods for Forecasting Wages of Employees Indicator. Voprosi Statistiki №12 – Moscow, Russian Federation, 2016, p.19-28.
5. Hussmanns R. (2004). Measuring the Informal Economy: From Employment in the Informal Sector to Informal Employment. Bureau of Statistics Working Paper, no. 53. Geneva: International Labour Office.
6. Website of the Federal State Statistics Service - http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/apps/6ca5fc804a47df3aa95cabf75a2eeced

Informal employment and Sustainable Development Goals: Mutual influence and consistency of indicators

Elena Zarova, Elvira Dubravskaya
Analytical Centre by Moscow City Government

Abstract

Informal employment is an integral part of the labour market in the economy of all countries without exception, having in many countries a significant impact on the level of socio-economic development and economic growth. The 2030 Sustainable Development Agenda resolution, adopted by the United Nations General Assembly on September 25, 2015, consists of 17 sustainable development goals and 169 targets, many of which are directly or indirectly related to informal employment.

The report presents the results of the study, which aims to develop methods for assessing the indicators of mutual influence and consistency of informal employment and indicators of sustainable development goals at the country level. At the same time, it is proved that this mutual influence is substantively and quantitatively different in countries with different levels of economic development.

The results of this work can serve as the basis for the development of a comprehensive analysis methodology for sustainable development, taking into account the fact that "they are complex and indivisible and will balance three aspects of sustainable development: economic, social and environmental".

The authors present the results of experimental calculations based on official statistical information about countries in terms of the scale and structure of informal employment, as well as indicators of the Sustainable Development Goals (SDGs). The results of these calculations show that levels of informal employment and the SDG indicators are interrelated, and this relationship has deep economic and social causes. The conclusion contains recommendations on methods for identifying and assessing the hidden factors of this interdependence, as well as the necessary information base.

Keywords

Informal employment; Sustainability; Indicators; Canonical Correlation; Synergistic effect

1. Introduction

The relationship of the SDG indicators is considered in scientific publications and practical documents as a basis for analyzing the achievement

of the sustainable development goals of the 2030 Agenda. For example, the ILO's publication "Decent Work and the Sustainable Development Goals: A Guidebook on SDG Labour Market Indicators" states that "Many of these indicators are intrinsically related to others, which is why it is important to interpret them as a coherent set so as to paint a comprehensive picture. In many cases, interpreting a given labour market indicator along with others sheds light on patterns and helps to avoid misinterpretations" 2.

The trans-boundary (between separate goals) and cross-boundary (between countries) effect of the mutual influence of SDG indicators is noted in the OECD report "Measuring Distance to the SDG targets"3 [p.18]. The relevance of our research is confirmed by the fact that the said OECD report noted that "… identification of synergies and trade-offs for many of SDG targets is an empirical question that has been little researched so far." 3[p.20].

As one of the few examples, we can give an estimate of paired correlation coefficients between the SDG indicator values in the study "Achieving a Sustainable Urban America" 4[p.18]. From the above publications, it becomes clear that the question of analysing the SDG indicators, taking into account their mutual influence and consistency with the real economic situation, is quite significant, but there are no methods for solving this yet. We offer an information base and methods for solving this problem on the example of the relationship between informal employment indicators and SDG indicators, bearing in mind that informal employment is a "litmus" of the level of a country's development and socio-economic situation in it.

2. Methodology

a. Database

The study was conducted by countries, the choice of which was determined, on one hand, by the representativeness of these countries in typological groups by level of economic development (according to the World Bank methodology), and on the other hand, by the availability of statistical data on all indicators used in the study.

In the research economies are divided into four income groupings: low (LI), lower-middle (MI), uppermiddle (UMI), and high (HI). Income is measured using gross national income (GNI) per capita, in U.S. dollars, converted from local currency using the World Bank Atlas method[1]. This paper uses World Bank definition for 2018 fiscal year[2].

[1] https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries

[2] Following the World Bank definition, for the current 2018 fiscal year, low-income economies are defined as those with a gross national income (GNI) per capita, calculated using the World Bank Atlas method, of US$1,005 or less in 2016; lower middle-income economies are those with a GNI per capita between US$1,006 and US$3,955; upper middle-income economies are

Figure 1 shows color-coding of income groupings. The set of countries is limited to 108 due to the availability of information regarding informal employment (the gray color indicates the lack of data).

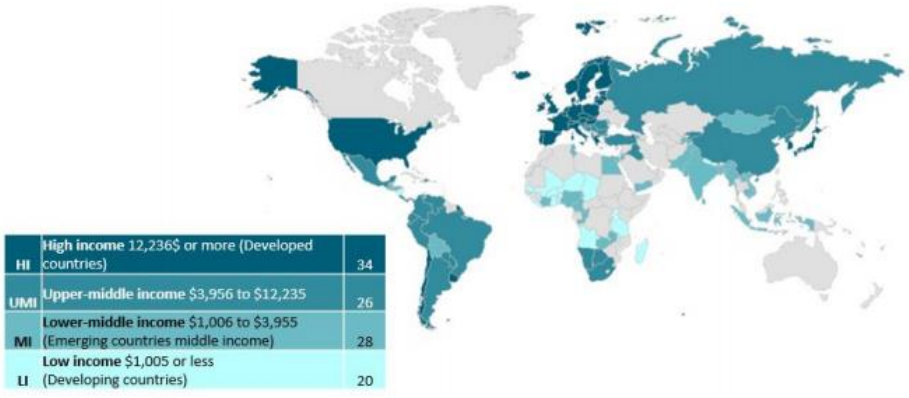


Fig. 1 Country Classification (four income groupings)

Sources: World Bank Analytical Classifications, Women and men in the informal economy: a statistical picture (third edition) / International Labour Office – Geneva: ILO, 2018

The study is based on two groups of indicators: 1 - indicators of informal employment in countries (including employment in the informal sector and informal employment in the formal sector); 2- SDG indicators.

The harmonized data of informal employment is taken from the book "Women and men in the informal economy: A statistical picture" published after the 20-th International Conference for Labour Statisticians (International Labour Office, 2018) 5. The estimates are result of joint collaboration of several ILO Departments and Women in Informal

Employment: Globalizing and Organizing (WIEGO) who applied harmonized definition informal employment on micro datasets from more than 100 countries[3]. 90 variables of cross-section data are coded according to the type of indicator. The detailed information is given in the Table 1/Figure 2.

those with a GNI per capita between US$3,956 and US$12,235; high-income economies are those with a GNI per capita of US$12,236 or more.

[3] The range of years of micro data used as a basis for the estimates is from mid-2000 to 2016. Data for more than half of the countries are from 2013 onwards and from 2010 onwards for 90 per cent of the countries considered. For each indicator, global and regional estimates of proportions result from the weighted average of national proportions for the latest year available. Those regional and global estimates are weighted by the denominator of the considered indicator using 2016 data from the ILO's Trends Econometric Models as relevant. Absolute numbers presented in the report refer to 2016 by multiplying the estimated regional or global estimate by absolute numbers for 2016 from the ILO's Trends Econometric Models as appropriate according to the denominator.

Table 1. Informal employment indicators/ International Labour Office indicators

Type of indicator	Description	Number of variables
B (base)	Share of informal employment in total employment and in non-agricultural employment by sex	8
U (urban)	Share of informal employment in total employment and in non-agricultural employment by urban or rural location	24
SA (including agriculture)	Share of informal employment in total employment by status in employment: including agriculture	17
NA (excluding agriculture)	Share of informal employment in total employment by status in employment: excluding agriculture	17
D (distribution)	Distribution of workers in informal employment and in formal employment by employment status and sex (including agriculture)	8
T (type of activity)	Share of informal employment in agriculture, industry and services by sex	4
BS (broad sector)	Distribution of workers in informal employment and in formal employment by broad sector of activity	6
SE (share of employment)	Share of employment in the informal sector, in the formal sector and in households by sex	6

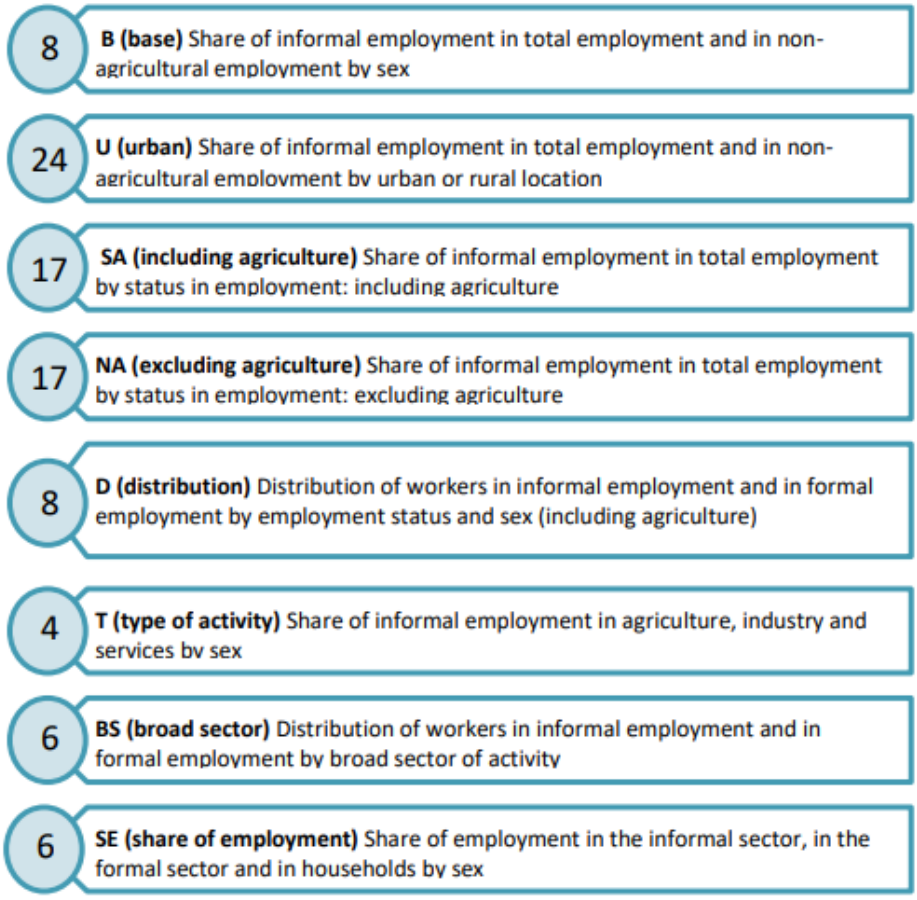


Fig. 2 Informal employment indicators/ International Labour Office indicators Sustainable Development Goals indicators*

* The number of variables representing the corresponding type given in the blue circle

The authors also created the set of indicators from various sources (UN Data, OECD, FAO, World Bank, Eurostat, UNESCO Institute for Statistics, UN, etc.), illustrating each of the SDGs. From SDG indicators only 14 were selected, for which methodologically consistent data is available for all countries in the groups under consideration (Figure 2).

Table 2 Sustainable Development Goals indicators

Sustainable Development Goal	**Variable**	**Description**
Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture	v2.a.2	Total official flows (disbursements) for agriculture, by recipient countries (millions of constant 2016 United States dollars)
	v2.c.1	Consumer Food Price Index

Sustainable Development Goal	Variable	Description
Goal 7. Ensure access to affordable, reliable, sustainable and modern energy for all	v7.1.1	Proportion of population with access to electricity, by urban/rural (%)
	v7.3.1	Energy intensity level of primary energy (mega joules per constant 2011 purchasing power parity GDP)
Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	v8.1.1.	Annual growth rate of real GDP per capita (%)
	v8.2.1	Annual growth rate of real GDP per employed person (%)
	v8.4.2	Domestic material consumption per unit of GDP, by type of raw material (kilograms per constant 2010 United States dollars) *Indicator in the global indicator framework repeat the v12.2.2.*
	v8.10.2	Proportion of adults (15 years and older) with an account at a financial institution or mobile-moneyservice provider (% of adults aged 15 years and older)
Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	v9.2.1a	Manufacturing value added as a proportion of GDP (%)
	v9.2.1b	Manufacturing value added per capita (constant 2010 United States dollars)
	v9.b.1	Proportion of medium and high-tech industry value added in total value added (%)
	v9.c.1	Proportion of population covered by a mobile network, by technology (%)
Goal 12. Ensure sustainable consumption and production patterns	v12.2.2	Domestic material consumption per unit of GDP, by type of raw material (kilograms per constant 2010 United States dollars)
Goal 17. Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development	v17.8.1	Internet users per 100 inhabitants

b. Methods

Initial information on groups of countries is represented by two blocks of indicators: 1- indicators of the relative size and structure of informal employment ("IE indicators") and 2- indicators of the SDGs ("SDGs indicators"), characterizing economic, social and environmental progress. Based on the

multidimensionality of both blocks, it was decided to conduct research on the following algorithm:

1. Assess the statistical relationships between the indicators of the above two blocks, determine their feature by groups of countries (use the methods of correlation analysis and the principal components analysis).
2. Identify the principal components separately in each of the multidimensional blocks: “IE indicators” and “SDGs indicators”.
3. Using the methods of canonical correlation, assess the degree of connection between these blocks of indicators and the main components defining this connection in the composition of the “IE indicators” and “SDGs indicators” feature spaces (“determinants of consistency”).
4. Build multifactor regression equations for the impact of indicators that determine the “determinants of consistency” of informal employment and SDG indicators on integral target indicators in the SDGs (for example, Annual growth rate per GDP per capita (v8.1.1) or Annual growth rate of real GDP per employed person (v8.2.1).
5. Assess the impact of the synergistic effect of interconnection of informal employment indicators and SDG indicators that form the "determinants of the consistency" on the integral target indicators.

3. Results

The main result of the implementation of the above algorithm is the identification and statistical evaluation of the determinants of the consistency of informal employment indicators and SDG indicators (Table 3). In assessing the canonical correlation, the first roots were statistically significant. On the basis of the maximum values of the weight coefficients, the main components from both blocks forming the canonical interconnection were identified. This made it possible to build regression models of the mutual influence of indicators of informal employment and the SDGs, taking into account their systemic effect on the aggregate indicators characterizing the socio-economic and environmental sustainability of countries.

Table 3 Determinants of consistency of informal employment indicators and SDG indicators

Group of countries	Coefficient of canonical correlation	Principal components contributing most to the canonical correlation ("determinants of consistency")	
		Block 1- "IE indicators"	Block 2 - "SDGs indicators"
HI- high income	0,795	F1 - informal urban employment in industry (in informal sector)	S1 - level of material wellbeing
UMI - upper-middle income	0,706	F1 - informal urban employment in industry (in informal and formal sectors)	S3 - economic growth rate
MI - lower-middle income	0, 649	F1 - informal urban employment in service (in informal sector)	S3 - level of informatization in the country
LI – low income	0,736	F3 - informal employment in agriculture	S1 - the level of welfare of the population

The Russian Federation was not included in the calculations, because even for that minimum of indicators, which turned out to be common for the countries in question, Russia does not yet have an information base. However, based on the fact that Russia's position in income groups changes all the time, according to the World Bank, it is necessary when assessing the prospects for achieving the SDGs to take into account the effect of informal employment on all determinants (S) shown in Table 3.

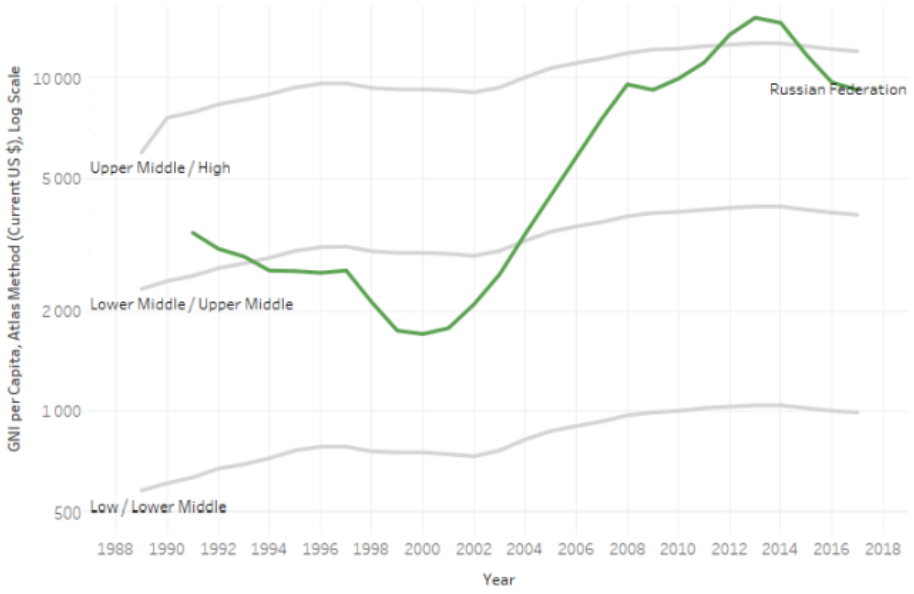


Fig. 3 Position of the Russian Federation in groups of countries by income level (World Bank data)

4. Discussion and Conclusion

The presented results provide an insight into a possible methodological approach for identifying and evaluating the complex multidimensional effect of the mutual influence of informal employment indicators and indicators of sustainable development at the country level. The results obtained by the authors contain "two news", as always: good and bad. The good news is that the phenomenon, which has already been indicated in the publications and which was hypothetically put forward by the authors of this article exists. This phenomenon is the mutual influence of informal employment indicators and SDG indicators, which has deep unobservable causes. The authors proved that it can be measured by statistical methods that allow us to identify the determinants of consistency, i.e. those characteristics of informal employment and sustainability of development, which in mutual influence determine the effectiveness of the system as a whole. Informal employment is only an example of a systemic socioeconomic phenomenon that, through implicit but statistically significant links, predetermines the achievement of the SDG goals. The proposed algorithm can be used to analyse the mutual influence of other complex subsystems.

The bad news is that, until now, the number of SDG indicators, for which there is a methodologically consistent single base for all countries, is small, much less than 169. The same has been found in terms of informal employment. Consequently, the primary task for the implementation of Agenda 2030 should be international cooperation forming such databases. This will provide the opportunity for in-depth system analysis at the cross-

country level, and this is necessary for the conscious provision of the achievement of the objectives of Agenda 2030.

References

1. Transforming our world: the 2030 Agenda for Sustainable Development., Resolution 70/1 adopted by the General Assembly on 25 September 2015, UN., https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalc ompact/A_RES_70_1_E.pdf
2. .Decent Work and the Sustainable Development Goals: A Guidebook on SDG Labour Market Indicators. https://www.ilo.org/stat/Publications/WCMS_647109/lang--en/index.htm
3. Measuring Distance to the SDG Indicators (June 2017). OECD., https://read.oecdilibrary.org/development/measuring-distance-to-the-sdg-targets-2017_9789264308183- en#page1
4. Prakash, M. Teksoz, K., Sachs, J.,Shank, M., Schmidt-Traub, G.(2017). The U.S. Cities Sustainable Development Goals Index 2017. Achieving a Sustainable Urban America., https://www.jstor.org/stable/resrep15885?refreqid=excelsior%3A870e53ced2900fe264f61372 6a796e29
5. International Labour Office. (2018). Women and men in the informal economy: A statistical picture. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_234413.pdf

Statistical literacy in the digital age in Argentina

Terán, Teresita
National University of Rosario. Rosario. Santa Fe Province. Argentina

Abstract

The teaching of Statistics must take place in a new social context, due to the vertiginous changes that are taking place in this digital age. Developing countries do not have the technological elements to carry out this process, so teachers must design teaching strategies to facilitate learning through information and communication technologies (ICT). In Argentina in 2010, the Government presented an initiative focused on recovering and valueing public education, with the objective of democratizing access to technological resources and to statistical knowledge. In the university context, the situation is different. They have mostly computer rooms with computers for teachers to develop their classes of Statistics. Most of them have virtual campuses and web pages. We believe that the recent development of digital technologies in Statistics makes it possible to expand traditional learning environments with a virtual space that, when used appropriately, allows time and space to be added to face-to-face learning environments.

Keywords

Information and communication technologies (ICT); Web pages; Virtual campuses

1. Introduction

Nowadays, it is not enough to be well informed, it is necessary to be able to reach the levels of competencies necessary to function at work and in society. Knowledge is the main source of developing and Information and Communication Technologies (ICT) are the most effective tools for its production and dissemination (Castells 2001). The inclusion of ICT with virtual learning models in the educational world (Zambrano, 2006) has turned out to be a complex process, involving the generation of actions in favor of the training and development of digital skills for both the teacher and the student.

In the field of education, necessary literacies are proposed to develop knowledge in society (Voogtz and Knezek, 2008): literacy of multimodal information processing, to deal with multiple media; in navigation, to know when and why there is a need for information; in interpersonal communication; visual, which encodes, evaluates, uses and creates images; hyperalphabetization, which manages representations of nonlinear

knowledge; literacy that deals with personal management; literacy to assimilate the complexity of digital tecnologies, for the responsible use of the networks. Here the role of Statistics.

Delors (1996) indicates that competences are given from four areas: being, knowing, doing and living, which represent the pillars of education. Each area contributes to the competence of the tasks, finally converging into a single one, since there are multiple points of contact between them.

The teaching of Statistics must take place, as we see, in a new social context, due to the vertiginous changes that are taking place in this digital age.

Developing countries do not have the technological elements to carry out this process, which is why teachers must design teaching strategies to facilitate learning through information and communication technologies (ICT)

The National Statistical System has carried out various operations aimed at obtaining statistics on Information and Communication Technologies (ICT). The Census conducted in 2010 investigated the totality of households in the country about the possession and use of computers, fixed telephony and mobile phones in the last three years. The National Survey on Access and Use of Information Technologies and the Communication (ENTIC) in 2011, provided new data, updated the information obtained in the 2010 Census and proposed new indicators for the availability and use of them. In the last decade, international organizations have considered the need to favor the access and use of ICT in society and to analyze them as a factor of economic development and social inclusion. Sharing this initiative, the INDEC (National Institute of Statistics and Census) takes an active part in the measurement process, working for it with the organizations that are leaders in the theme and its harmonized measurement.

In Argentina until 2010, computer equipment in schools was insufficient to meet the expectations of teachers and students, since only the contribution of parents and the community involved was available for purchase. That same year the National Government presented an initiative focused on recovering and valueing public education, in order to reduce the digital, educational and social gaps in Argentine territory. The program called "connect equality" developed digital content usage in different didactic proposals, especially in Statistics and worked on the processes of teacher training, with the intention of transforming models, processes and paradigms of learning and teaching, contemplating both the use of portable equipment in the school environment as in the home, looking for an impact at a social level that transcends the educational field.

The availability of information on ICT accessed at home and its use, allows obtaining key estimates for the analysis of digital inclusion in Argentina. In this sense, it is relevant to know the number of households and people in households that have and use computers and the Internet: their frequency of

use and activities for which they are used. No less important is the information regarding those who access radio, television and telephony, fundamentally in a context of State policies that prioritize the distribution of speech and audiovisual communication. The survey of information used an interactive digital questionnaire developed by the INDEC Institute that was administered through personal interviews, using electronic tablets. The National Survey on Access and Use of Information Technologies and Communication 2015 (ENTIC) was administered to 3.804 households, belonging to a probabilistic sample of private homes in the 31 urban agglomerates that have more than 26.8 million people. It was administered to people who were 5 years and older, for which its representation reaches 8.4 million households and more than 24.7 million people of these ages. Among its indicators are: access to radio, households due to availability of ICT goods, according to estimation domain. It is observed at a national level that 68% used a computer, while the Internet was use by 66.1%. With respect to cell phones, 72.8% use them. With regard to the possession of computers, 67% have them, 61.8% have internet and 89.6% have cell phones. This fact emphasizes the need for new programs so that computers are available in schools from kindergarten. This way, those who do not have access to computers in their homes can learn from them at schools.

Its objective was to make society literate in ICT, democratizing access to technological resources and information, thus enabling ordinary citizens to interpret statistical information, awakening the critical spirit, without social discrimination, economic or spatial, reaching the entire country; implement accompanying actions for the incorporation of technologies in educational practices and in institutional management. In 2016, the program was suspended.

Currently, work has begun on the cell phone as a teaching tool in Statistics, since it is accessible to the vast majority of students, who can share with their parents at home.

The INDEC (National Institute of Statistics and Census) inaugurated in 2017 a program of Statistical Literacy, with a talk in which more than 60 secondary students of the City of Buenos Aires participated. The program is aimed at publicizing the Institute's production and promoting the proper use of statistical data.

The talk was in charge of INDEC technicians who explained the process of production of statistical information, where thematic, operational and methodological aspects of the next National Household Expenditure Survey and the current Consumer Price Index of national coverage were exposed. . It also detailed how to perform processing, searches and analysis in databases.

The INDEC thus inaugurates a new channel of communication with young users of statistical information, with the intention of expanding the statistical culture of the population in the stages of construction of citizenship.

In the university context, the situation is different. They have mostly computer rooms with computers for teachers to develop their classes of Statistics. Most of them have virtual campuses. The presence of the ICT makes the contact with the students and teachers more fluid, facilitating the understanding of the contents of this subject, but sometimes internet does not work at universities.

Another alternative is web pages, especially of Statistics. Adell (1997) argues that they facilitate teaching and learning processes, understanding this virtual resource as a tool that helps the student to continuously carry out a metacognition exercise.

Web pages, especially Statistics, tend to improve the quality of teaching, through a methodological change in teaching and learning processes in the Statistics Chair. This change makes it possible to expand the traditional learning spaces thanks to the use of digital communication and information technologies supplied by digital networks. Teachers adopt a methodology where the student plays an active role in the construction of knowledge through observation, recording, comparison, analysis and synthesis of what is observed in a series of programmed activities.

Another option in the digital age is to introduce a combination of non-face-to-face classes or elearning and blended learning or blended learning classes; being the face-to-face activities carried out in the theory and practice of the Statistical Classes, while the remaining, non-presential ones, are carried out through the web pages, developed by the Statistics Chair, with free support from the Wix company.

2. Methodology

In the Faculty of Veterinary Sciences of Casilda, dependent on the University of Rosario, the Chair of Biostatistics has developed a website that acts as a link, beyond the face-to-face classes, between teachers and students in order to improve the learning process.

The website is https://bioestadisticavete.wixsite.com/cursada

The construction of this website was made based on evaluation criteria related to the technological aspect, the quality of the information and the design. These aspects were focused as follows: Technology: accessibility to the site and its contents was prioritized, making sure that it was always available and that its elements could be consulted from the widest variety of equipment and systems. For its analysis, the following items were taken into account: ease of location of the site, access speed, functioning of the links and quality of the multimedia elements.

Quality of Information: A review of all the teaching material available in the chair was made as well as update and adaptation with multimedia support. Theoretical elements, resolved and proposed practices were arranged. Communication channels were opened between teachers and students. The items considered in this variable were: distribution of the classes, understanding of theoretical contents, interpretation of practical slogans and interpretation of the postulated solutions.

Design: A page with a high contrast design, large fonts, fixed menu bars, sections associated with units of the study program was made. The icons that present the downloadable material are indicative of the type of file and its format, so that the user knows what software they will need for viewing. The items considered were: structural design (colors, funds and sources) and location and understanding of menus.

In order to evaluate the quality of the development site, its efficiency was studied in terms of the use that the students made of it.

The concept of use is an attribute of quality that measures the ease with which the user uses the interface, because the first capacity of the system must be to respect the physical and psychological processes of the person who interacts with it.

There are two general methodologies to study the use of an educational web site. These are the expert analysis and the tests with the users (students).

In addition to the Wix website, the chair of Biostatistics has a virtual campus site where the daily information on the chair is presented, PowerPoint of the theoretical classes, applications on veterinary medical practices, resolved problems and a space to interact with the students among them, classes of virtual consultations.

The website is https://fveter.unr.edu.ar/

3. Results

In the case of the study of the use of the web site developed by the Chair of Biostatistics, it was decided to use the tests with the users, since this instrument is intended to sustain and improve the learning process.

The students' tests were made for the construction of a checklist based on the evaluation criteria that was used as the basis for the development of the page.

A total of 10 aforementioned analysis factors was postulated, which were turned into a questionnaire that was voluntarily completed by the students.

Each of these factors was evaluated using a Likert scale of 4 levels (Very Good, Good, Fair, and Poor, although for simplicity of presentation of results were summarized in 2 categories: Positive (Very Good and Good) and Negative (Regular and Bad).

The 45 participating students signed a consent where they are assured of the confidentiality of the use of the data and allow the same to be used in the present investigation. A table was drawn up showing an acceptance and general approval of the use of the page measured through the positive opinions of the students.

The Technology factor received high marks in ease of location and functioning of the links, while the lowest notes refer to access speed, something that depends on many factors and not only on the design and content of the page, and multimedial elements, where very personal and subjective tastes and uses interact.

The Quality of Information factor received in general almost 90% of positive qualifications, except in the item Understanding of theoretical contents, something that may be related to the nature of the Subject and its own difficulties, rather than with the material included in the website.

The Design factor received more than 90% of positive grades in the two items that comprise it.

4. Conclusion

These results, which indicate a high acceptance by the students of all the work carried out, constitute an incentive for future extensions, the improvement of the website and the inclusion of new features in it.

We believe that the recent development of digital technologies in Statistics makes it possible to expand traditional learning environments with a virtual space that, when used appropriately, allows time and space to be added to face-to-face learning environments.

The teacher, in our case of Statistics, is the one who has to apply didactic strategies so that in the digital age all the tools of information and communication are implemented in the teaching from the initial level to prepare citizens with a critical spirit that develops competences for their insertion in this vertiginous society in which we live.

References

1. Adell, Jordi (1997, noviembre). EDUTEC Tecnología Educativa. Tendencias en educación en la sociedad de las tecnologías de la información. Recuperado el 05 de abril de 2019 de http://www.uib.es/depart/gte/revelec7.html
2. Castells, M. (2001). La era de la información. Madrid. 3ª edición. Vol. 3 Fin de milenio. España: Alianza Editorial.
3. Córdoba, O.; Terán, T. (2018). Construcción de un sitio web educativo a partir de criterios de evaluación de páginas web. XIX Jornadas de divulgación Técnico Científicas, Facultad de Ciencias Veterinarias. Casilda

4. Delors, J. (1996). La educación encierra un tesoro. Madrid: Santillana, ediciones UNESCO. Madrid.
5. ENTIC (2011). Recuperado el 5 de abril de 2019 de https://www.indec.gov.ar/uploads/informesdeprensa/entic_10_15.pdf
6. Fernández E.; García, J; Tornero, I.; Sierra, A (2011). Evaluación de la usabilidad de un sitio web educativo y de promoción de la salud en el contexto universitario. ISSN 1135-9250. Revista electrónica de tecnología educativa (37)
7. García-Valcárcel, A., Basilotta, V. & López, C. (2014). Las TIC en el aprendizaje colaborativo en el aula de Primaria y Secundaria. Comunicar, 21(42), pp. 65-74.
8. INDEC (2010). Recuperado el 5 de abril de 2019 de https://www.indec.gov.ar/nivel4_default.asp?id_tema_1=2&id_tema_2=41&id_tema_3=135
9. INDEC (2015). Recuperado el 5 de abril de 2019 de https://www.indec.gob.ar/
10. Martinez Aldanondo Javier (2004). Blended Learning o el peligro trivializar el aprendizaje http://www.gestiondelconocimiento.com/
11. Peñaloza Castro Eduardo (2010). Modelo Estratégico de Comunicación Educativa para Entornos Mixtos de Aprendizaje: Estudio Piloto". Pixel-Bit. Revista de Medios de Educación (37). Pp. 43 –55
12. Salinas, Jesús (2004). Innovación docente y uso de las TIC en la enseñanza universitaria. Revista de Universidad y Sociedad del Conocimiento (RUSC).
13. Voogt, J. Y Knezek, G. (2008). International Handbook of Information Technologies in Primary and Secondary Education. New York: Springer.
14. Zabalza, Miguel A. (2007).Competencias Docentes del profesorado universitario. Calidad y desarrollo profesional. Narcea S.A España.
15. Zambrano, W. (2006). Modelos de aprendizaje virtual en la educación superior MAVES basada en tecnologías Web 2.0" (tesis doctoral). Universidad Pontificia de Salamanca. España. Recuperado el 20 de abril de 2019 de: https://www.ecoeediciones.com/wpcontent/uploads/2015/08/Modelos-de-aprendizaje-virtual-para-la-educai%C3%B3n-superior1ra-Edici%C3%B3n.pdf

Integrating Agricultural Censuses and Surveys for optimal sectoral data collection

Jairo Castano, Oleg Cara
Food and Agriculture Organization (FAO)

Abstract

In countries with less developed national statistical systems (NSS), agricultural censuses (ACs) and sample surveys are not conducted regularly. This means that both structural data (sourced from censuses) and current statistics (sourced from sample surveys) are not readily available or up-to-date for informed decision-making on agricultural and rural development. In such countries, because of the sheer needs, when a census of agriculture is planned, stakeholders exert pressure on the census agency to collect both structural and non-structural data (atypical for a census), overburdening the census questionnaire and ultimately jeopardizing the quality of the census operation. An increasing number of countries make efforts towards better integrating statistical activities. The FAO World Programme for the Census of Agriculture 2020 (WCA 2020) advocates the development of an integrated multi-year programme of statistical operations involving AC, current surveys and other data collection operations. By integrating these operations, the AC can focus on collecting essential structural items, while regular agricultural sample surveys and administrative registers can focus on collecting non-structural data needed more frequently.

Finally, the results are used to discuss the concept and measurability of dependents contractors in the relevant economic setting. Besides evaluating the suitability of the tested questions for capturing the group of dependent contractors, additional questions that could be relevant for measuring the phenomenon in a Danish – or similar – context are considered.

Keywords

Agricultural Census; World Programme for the Census of Agriculture; Integrated census/survey modality; Cost-effective data collection; Census items

1. Introduction

The Sustainable Development Goals (SDGs) have presented new demands for more data and challenges in terms of monitoring and reporting progress towards their achievement. There is a need for cost-efficient methodologies, modern tools for data collection and better data integration. While some progress on accessing existing information has been made thanks to open

data, critical gaps on data production still remain in many countries. This is partially due to lack of adequate coordination of data collection operations and mapping of these operations with country data needs. To meet these data needs and fill-in the information gaps in a cost-efficient way, an integrated agricultural statistics system is of crucial importance. Such a system involves a multi year programme of agricultural surveys articulated with the agricultural census.

In this integrated system, on one hand, structural agricultural data, such as size of holdings, land use, crop areas, livestock numbers and agricultural inputs are collected at the lowest geographical level through censuses of agriculture (every five or ten years). The AC is the backbone of the agricultural statistics system and has the widest (usually complete) coverage of agricultural holdings. The AC gives a snapshot of the structure of the agricultural sector in a country and, when compared with previous censuses, provides an opportunity to identify trends and structural transformations of the sector, and points towards areas for policy intervention. Census data are used as a benchmark for current statistics and their value is increased when they are employed together with other data sources. Data producers rely on the census to update the frame for current agricultural sample surveys (FAO, 2015).

On the other hand, data such as crop and livestock production, food consumption, cost of production, production prices and production methods are collected through regular sample surveys and/or administrative reporting systems to provide in-depth, more frequent and timely agricultural statistics. The current statistics are needed to monitor ongoing agricultural and food supply conditions and to support decision-making in the short term.

2. Regularity of the agricultural census

According to the WCA recommendations (FAO, 2015), a country should conduct an AC at least once every ten years, providing key structural items (23 essential items) and frame items for intercensal sample surveys. Although country participation in the census rounds has increased steadily since the 1990s until the 2010s (90 countries and territories in the 1990 census round), the number of countries that conducted an AC in the 2010 round (2006-2015) was relatively small. In fact, 128 countries and territories conducted an AC in the 2010 round, compared to 214 that conducted a population and housing census (PHC) in the same period (see Table 1).

Table 1. Number of countries and territories that participated in the 2010 census rounds

Region	Population and housing census	Agricultural census
Africa	**49**	**22**
North and Central America	**36**	**17**
South America	**14**	**11**
Asia	**44**	**31**
Europe	**47**	**35**
Oceania	**24**	**12**
TOTAL	**214**	**128**

Source: FAO and UNSD records, 2019.
Note: The 2010 round of PHCs covered the period 2005-2014 while that of ACs covered the period 2006-2015.

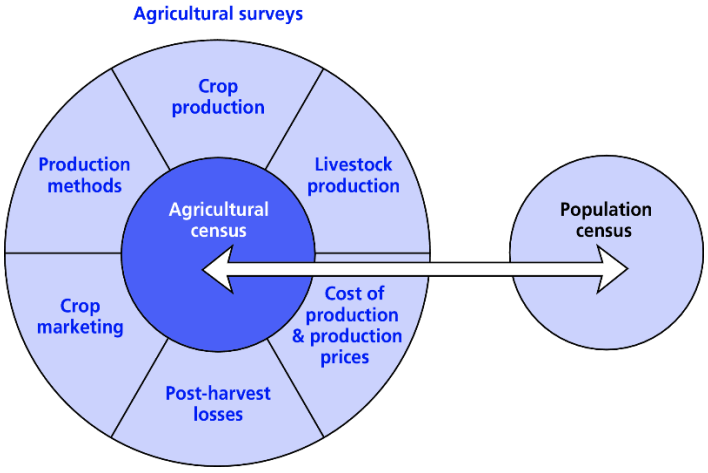
Countries that do not conduct regular ACs and do not have access to other reliable sources to bring up to date the sampling frame for their agricultural surveys, often conduct these surveys using outdated frames due to the long-time gap with the latest AC. This could have an important impact on the reliability of currents statistics emanating from these surveys. In such circumstances, when a new AC is conducted, serious discrepancies could be observed between two data sets: the series released before the AC and that based on the new AC.

3. Integrating Agricultural Censuses and Surveys

A good coordination between the AC and sample surveys means that statistics are based on standard concepts, definitions and classifications, preventing duplication of statistical activities, avoiding the release of conflicting statistics, excessive response burden and waste of resources, and, ultimately, contributes to better understanding and use of statistics by users. In this way, the AC is not overburdened with a wide range of items that may affect the quality of the collected data. Instead, the census can focus on a coherent and manageable set of items, assuming that other (non-structural) data needed more frequently are available in a comparable form through regular agricultural sample surveys and other sources.

A schematic representation of the system of integrated ACs and surveys is shown in Figure 1. It illustrates the links among agricultural surveys and the AC, and between the latter and the PHCs (for the household sector).

Figure 1. The system of integrated Agricultural Censuses and Surveys



Source: FAO, 2018

In order to increase synergies between the AC and other statistical operations when moving towards a more integrated system, the WCA 2020 recommends some actions, no mutually exclusive:

1) Identifying the specific role of the AC.
2) Focusing the content of the census of agriculture on structural items.
3) Integrating the census with periodic rotating surveys.
4) Using data available from administrative sources.
5) Better integrating the agricultural and population censuses.

1) Identifying the specific role of the Census of Agriculture

In countries where the agricultural statistics system is not so well-established, the AC (when conducted) is often an isolated one-off operation mobilizing a large amount of resources in a short period of time, followed by several years of data discontinuity. Identifying the specific role and objectives of the AC as a component of the system of integrated agricultural censuses and surveys is the first major step in preparation for an upcoming AC. A good strategic plan (like the Strategic Plan for Agriculture and Rural Statistics (SPARS), mainstreamed into the National Strategy for the Development of Statistics (NSDS) process) is of crucial importance, especially in countries without a well-established agricultural statistics system. This plan ensures that both the AC and surveys complement each other and together generate the required statistics with the appropriate frequency (FAO, 2015).

2) Focusing the content of the Agricultural Census on structural items

Identifying the content of the census and that of the sample survey programme is a challenging activity and should be undertaken in strong cooperation between producers and users of agricultural statistics. This activity might be difficult when statistical activities are under the jurisdiction of different government institutions. For instance, the national statistics office might be responsible for the AC, whereas ongoing sample surveys such as agricultural production surveys are conducted by the line ministry. In these circumstances, establishing coordination between all agencies involved in the agricultural statistics' production is paramount (FAO, 2015 and 2018).

As highlighted in the WCA 2020, including too many items in the census questionnaire would be counterproductive. The AC should include only key structural data, but the characteristics which are better collected through sample surveys (needed also on a more frequent basis) should be included in the sample survey programme. In order to help countries to identify the census content tailor made to country specific situation and ensuring international comparison, the WCA 2020 classifies the items to be included in the census into three categories:

(i) essential items (23 items in total);
(ii) frame items (15, of which 6 are also essential items); and
(iii) additional items (96).

However, the WCA recommends that an AC includes all essential items to enable national and international comparison. Frame items for census modules or follow-up surveys can also be included[1]. Thus, the WCA 2020 recommends up to 32 essential and/or frame items to be used as a starting point in defining the content of the AC in a country. However, the actual list of country-specific census items should be established by each country in close cooperation with the stakeholders.

3) Integrating the census with periodic rotating surveys

Countries apply different modalities for carrying out the ACs, depending on country statistical capacity, national preferences and the availability of resources and data sources. The changing environment and data requirements have led some countries to conduct a detailed review of their agriculture statistics programme, including alternative options for conducting the AC. In addition to the classical and modular approaches, the WCA 2020 introduced two new alternative modalities for cost-efficient

[1] If a community survey is conducted together with the AC, countries can select from 34 community items recommended by FAO for the community survey to complement holding-level items and other available data sources (i.e. statistical and administrative data).

census, namely the *integrated census and survey modality* and the *use of registers as sources of census data*. The former is discussed here while the latter is discussed further down.

The integrated census and survey modality aims at producing in a cost-effective way a wide range and regular flow of data by rolling out the collection of thematic data over the period separating two ACs (usually ten years). An example of this is the FAO Agricultural Integrated Survey (AGRIS) programme. AGRIS comprises: i) a census core module (which could be even lighter than in the modular approach) to be carried out on a complete enumeration basis and ii) an annual production module and several rotating thematic modules (such as "economy", "labour", "machinery and equipment", and "production methods and environment") to be conducted on sample basis between two censuses. The AGRIS production module (conducted on annually, or even quarterly), covers crop production and livestock production. Between two ACs, the AGRIS production module and one or more of the rotating modules will be implemented each year. In this integrated census/survey modality, the core census module together with the AGRIS rotating modules should cover all essential items. The frequency of the modules will depend on the countries' agricultural statistics systems and data demand priorities (FAO, 2015). Table 2 below illustrates a possible implementation of an AGRIS plan between two censuses.

Table 2. Example of implementing an integrated census and survey modality

		Years									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Agricultural Census (core module)		•										•
AGRIS Annual Production Module (AGRIS Core Module)[2]	Crop and livestock production (and other key variables)		•	•	•	•	•	•	•	•	•	•
Rotating Module 1	Economy		•		•		•		•		•	
Rotating Module 2	Labour			•				•				
Rotating Module 3	Production methods and environment					•				•		
Rotating Module 4	Machinery, equipment and assets		•				•					

Source: Global Strategy, 2017 (adjusted)

The content of the core and rotating modules, and their frequency are defined by countries depending on their data demand priorities and taking into account cost viz-a-viz human and financial resources

[2] EFE

available.

The countries without a well-established census and survey programme may find this modality as an important step towards the creation of a system of integrated agricultural censuses and surveys. This modality is proposed to provide tools for a system that ensures a continuous flow of data instead of concentrating all resources on a single census operation.

4) Using data available from administrative sources

In recent years, a growing number of statistical offices, particularly in developed countries, are moving towards more use of data from administrative sources in the statistical data production process. Making greater use of administrative data is a way to reduce the burden on respondents and generate more frequent data with reduced costs by not collecting data that are already available through the administrative process.

For example, an assessment made by Statistics Canada in 2012 concluded that the AC remains relevant and necessary for the country[3]. However, in order to render the program more efficient, some changes were needed, such as increasing the use of administrative data to gradually replace survey data and increasing utilization of remote sensing.

The WCA 2020 introduced the use of administrative registers as a source of census data as an alternative census modality. This modality is relevant for countries with well-developed administrative registers, suitable for statistical purposes and requires the access of statistical agencies to administrative data (individual records). When implementing this modality, the census agency needs to decide which items will be collected directly from available administrative data sources (and thus excluded from the census questionnaire) and those to be obtained through census enumeration.

5) Integrating the agricultural census and the population and housing census

The integration of the AC and the PHC, as well as other censuses (such as the economic census) constitutes another important aspect of the integration of statistical operations in the NSS. The relationship between the AC and the PHC, for instance, can take several forms - from coordinating aspects of the two censuses to including key agricultural

[3] The assessment used criteria such as relevance, accuracy, coherence, timeliness, interpretability, accessibility, respondent burden, cost and acceptability by user and respondent community.

items in the PHC, and even joint data collections. The relationship between the two censuses can cover (FAO & UNFPA, 2012):

- Coordinating aspects of the two censuses in terms of use of common concepts, definitions and classifications; sharing field materials; building enumeration areas which suit both censuses; organization of fieldwork.
- Using the listing of the PHC as a starting point for the frame for the household sector of the AC;
- Collecting agriculture-related data in the PHC to identify the households engaged in own-account agricultural production households (either through few basic items or adding an agriculture module).

There are many country examples of integration of the AC and the PHC. Some 60 countries included agriculture-related items in their PHCs in the 2010 census round. Sri Lanka conducted the census of the agricultural sector jointly with the Economic Census 2013/2014. It is anticipated that in the upcoming censuses more and more countries would be looking for better linkage between these censuses. Some Pacific island countries, particularly those composed of scattered atolls, have included or are planning to include an agriculture module in their PHC to deal with high fieldwork costs and logistical challenges.

The integration of data collections within the NSS requires, in many countries, to improve the legal and institutional framework and to build statistical capacity across the different institutions concerned, as well as the support of the government to optimize the data collections in line with statistical plans and programmes and secure budgetary allocations.

4. Discussion and Conclusion

Growing user demands for relevant, reliable and coherent data, and the need to improve cost-efficiency, require additional efforts in many countries towards achieving better integration of statistical collections within the NSS.

The AC, as the backbone of the system of integrated agricultural censuses and surveys should not be overburdened with a wide range of numerous items that may affect the quality of collected data. Instead, the census should focus on a coherent and manageable set of items, assuming that other (non-structural) data needed more frequently are available in a comparable form from regular agricultural sample surveys and other sources.

The alternative modalities of census data collection constitutes important ways to better integrate and improve cost-effectiveness of data collections. The use of registers as sources of census data, which is relevant for countries with well-developed administrative registers, would contribute to a better correlation

of data and to reducing response burden. The other new modality - integrated census and survey modality-would assist countries with underdeveloped agricultural census and surveys programmes to move towards a fully integrated approach.

References

1. Eurostat (2015). Strategy for agricultural statistics for 2020 and beyond. https://ec.europa.eu/eurostat/documents/749240/749310/Strategy+on+agricultural+statistics_Final/fed9adb7-00b6-45c5-bf2c-2d7dcf5a6dd9
2. FAO (2015). World Programme for the Census of Agriculture 2020 Volume 1: Programme, concepts and definitions. FAO. Rome. (also available at http://www.fao.org/3/a-i4913e.pdf)
3. FAO (2018). World Programme for the Census of Agriculture 2020 Volume 2: Operational guidelines. FAO. Rome. (also available at http://www.fao.org/3/CA1963EN/ca1963en.pdf)
4. FAO & UNFPA (2012). Guidelines for linking population and housing censuses with agricultural censuses. Special Issue of the FAO Statistical Development Series. Rome.
5. FAO (2017). Regional Roundtable on the World Programme for the Census of Agriculture 2020 (WCA 2020). 6-10 November, Nadi, Fiji, http://www.fao.org/index.php?id=85857
6. Global Strategy to improve Agricultural and Rural Statistics (2017). Handbook on Agricultural Integrated Survey (AGRIS). Rome. (also available at http://gsars.org/wp-content/uploads/2017/12/AGRIS-HANDBOOK.pdf)

The integration of the Census of Agriculture with the Business Statistics Program: The keystone for the next generation of Censuses in Canada

Étienne Saint-Pierre
Statistics Canada

Abstract

The approach used by Statistics Canada for the quinquennial Census of Agriculture consisted of contacting directly the entire farm population to obtain all of the data using a questionnaire as collection vehicle. With the 2021 Census of agriculture, this approach will start evolving towards a new business model as a response to an evolving context in the production of agriculture statistics.

In Canada, the farms are now predominantly complex operations with structures that are more aligned with the business sector than the household sector. The key issues affecting the agriculture sector go well beyond the primary sector. Data users, policy makers and farmers want to get their hands on high-quality, real-time information to make informed decisions. Simultaneously, the increased availability of alternate sources of information combined with the refinement of large datasets processing techniques open a realm of opportunities for the Census of Agriculture.

The article describes how the integration of the Census of Agriculture with the Integrated Business Statistics Program is a critical element in the short-term to position strategically the Census Program for the future.

Keywords

Census of Agriculture; Data Integration; Concepts Harmonization; Statistical Frame, Integrated Business Statistics Program

1. Introduction

Up to the most recent Census of Agriculture conducted in 2016, the approach used by Statistics Canada for this program consisted of contacting the entire farm population every five years to collect all of the data using a questionnaire as the collection vehicle. Unique concepts, systems, tools and methods were used throughout the statistical process.

The context is rapidly evolving in regards to the data landscape in the agriculture sector, offering both challenges and opportunities. Data users, policy makers and farmers want to get their hands on high-quality, real-time information to make informed decisions. The ultimate goal of the Agriculture Statistics Program (ASP) is to produce the near-real time granular information with minimal contact with respondents by exploiting alternative sources of

information such as satellite imagery, administrative data collected by various organizations or in other data collections and using the latest leading edge-methods to build performing models. In order to adapt to this rapidly changing context, it is recognized that practices and methods used in the Census of Agriculture and its level of integration with the rest of the statistical systems need to undergo a profound transformation. The integration of the Census of Agriculture (CEAG) - backbone of the Agriculture Statistics Program (ASP) - with the Integrated Business Statistics Program (IBSP) is an essential first step.

The evolving context affecting the production of agriculture statistics will be described in the second section of the paper. A brief description of the IBSP will be covered in the third section. The fourth section will cover the fundamental changes and opportunities for the Census of Agriculture 2021 induced at the different stages of the survey cycle resulting from the integration to the IBSP.

2. A rapidly evolving context

Since 1956, the Census of Agriculture has used a collection model based on the complete enumeration of farms and data obtained directly from respondents. As it is the case in many sectors of the economy, the agriculture sector is undergoing a rapid and profound transformation leading statistics providers to change their traditional business model to provide the information required by the data users.

i) Consolidation and complexification of the structures

Results from previous censuses show a massive consolidation of farms in the agricultural sector in Canada. The number of farm operations declined by 30.0% from 1996 to 2016. Operations have evolved to become larger and increasingly integrated and complex. In 2016, 25.1% of all agricultural operations reported being incorporated, compared with 2.2% in 1971. A growing number of agricultural operations reported more than one business, where traditionally all agricultural activity was reported as a single business entity. As agricultural operations move toward incorporation as a business practice, the level of complexity associated with agricultural operations has increased as well. The separation of an agricultural venture into legally separate business units adds to the logistical challenges of handling data collection and processing used in the traditional survey approach. Farms are businesses that can now be better handled in the Statistics Canada's business survey processing infrastructure.

ii) Need for more horizontal and timely information

The expectations of data users are rapidly changing. Data users need current, detailed and easily accessible data and analysis. The broad policies and issues related to the agriculture sector go well beyond the primary sector. The Census of agriculture needs to be adapted accordingly.

Whether it is to transport efficiently the agricultural goods from farms to markets by getting real-time production and capacity data, to assess the impact of climate change on current practices or to support the development of sustainable ones, to develop new food products or export to international markets, to measure the food availability and security and their impact on health outcomes, stakeholders expect the statistical agency to produce more frequently high-quality, granular agriculture data fully integrated with data from other economic sectors to support evidence-based decision making.

iii) Proliferation of data and refinement of large datasets processing techniques

Increased access to administrative and transactional data and access to high-quality satellite images, supported by refined techniques to process large datasets, make it possible to consider new ways to produce timely relevant statistics with minimal contact with respondents.

There has been an increase in the availability of alternative sources of agricultural data. So far more than 300 data sources are available to be used at different stages of the survey cycle in the ASP. These data sources include sources from the Canada Revenue Agency (e.g., tax data) and supply-managed sectors (including dairy, chicken, eggs and turkey), where datasets include quota and production figures.

Crop insurance data, which detail what crops have been planted and insured, as well as their yield at the field level, represent a great source of ground truth information that can be combined with satellite images, agro-climatic data, and advanced modelling techniques to produce frequent high quality yield and crop area estimates at a very granular level of geography without having to contact farmers. Access to animal traceability database, tracking movement and typical approaches in machine learning can be used to create animal level daily movement data and real-time pork inventory from the aggregate movement information. Leveraging these new techniques and these alternatives sources of information will definitely transform the way the Census of Agriculture was traditionally conducted.

Migration of the Census of Agriculture to the corporate platform used for the production of Economic Statistics (IBSP) is a first essential step to adapt the program to the rapidly evolving context.

3. The Integrated Business Statistics Program

In 2014, Statistics Canada launched the Integrated Business Statistics Program (IBSP) in order to have a more effective model to produce economic statistics. The IBSP provides a standardized and generic processing framework from collection to dissemination and a set of common tools and systems for a large number of heterogeneous economic surveys conducted at Statistics Canada. By 2021, nearly 130 economic programs (annual, sub-annual, industry specific, economy-wide, activity based, financial) will be integrated into this harmonized framework, including all the ASP and the Census of Agriculture 2021.

4. The Census of Agriculture as an integrated component of the Economic Statistics Program: Changes and opportunities

In this section, the most important elements of the migration of the CEAG to the IBSP and their opportunities are highlighted.

i) Use of Statistics Canada's Business Register as a common frame

As the Census of Agriculture program moved toward the integration of alternate sources of information, the use of the Business Register (BR) as a frame is an essential element. The BR is the common frame for all economic surveys. The BR is an essential tool that helps to maintain and profile the increasingly large and complex agricultural enterprises and their operations. This database is kept up-to-date by combining administrative sources (including tax data), survey feedback and results from direct contact with business respondents.

In addition to being constantly up-to-date, the BR contains many variables that are required to have efficient record linkages from different administrative sources (business number, legal name, operating name, address, name of the farm operator, geographic coordinate, etc…) and statistical programs. The efficient linkages of multiple alternative datasets with the BR is a central element to move toward the production of near-real time data with minimal contact with respondents. The use of the BR as a frame, the timeliness of its updates and the sharing of data with the tax agency has resulted in the production of an annual census of farm operators and families to estimate a range of financial variables without contacting respondents. These variables will no longer need to be collected from respondents in the Census program every five years.

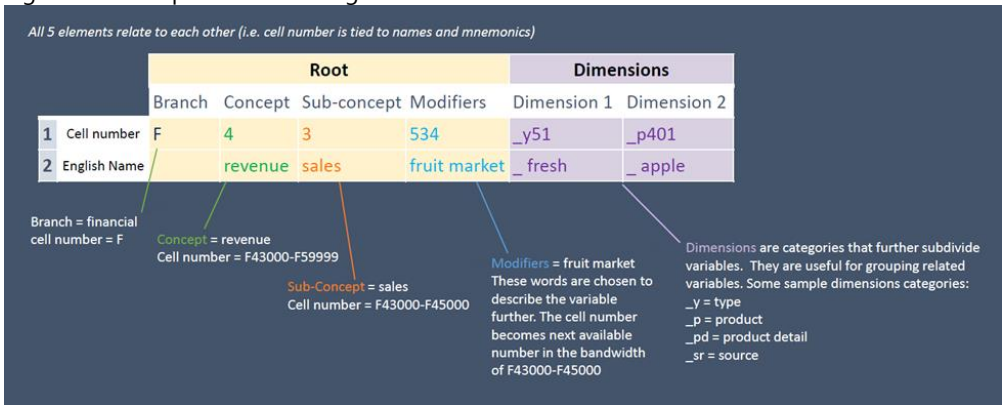
In the last Census of Agriculture 2016, the frame consisted of two universes: 1) Businesses involved in agriculture identified on the Statistics Canada's Business Register (BR) which comprise all units in all economic sectors involved in economic production in Canada and 2) Households identified through the Census of Population questionnaire who could be involved in agricultural

activities. For a better alignment with the concept used in the economic statistics program and the BR, the concept to define the farm population used in the Census of Agriculture is being reviewed. The second component of the frame coming from the self-identification in the Census of Population has been dropped. The vast majority of operations identified via the Census of Population were already covered in the Business Register (86%). Those missing were mostly operations having no real economic (0.4% of the value of sales) or policy impact (ex. hobby farms). The identification, collection and processing of these units is very expensive with very little impact on the various aggregates. With the Census of Agriculture 2021, it is proposed to define a farm or agricultural holding as a unit producing agricultural products and reporting agricultural revenues or expenses for tax purposes. Signals from the tax agency are used to keep the BR up-to-date. The convergence of the definitions will basically mean that the farm population used in the Agriculture Statistics Program will be kept up-to-date in real-time and will not see a degradation between Censuses.

ii) Harmonized concepts and content

With the migration to IBSP, all the Census of agriculture variables are organized according the IBSP Variable Naming Framework. In the IBSP, the naming framework is applied in a consistent, coherent, and logical manner that can be replicated for all statistical programs. All variables are assigned a cell number logically generated from variable names and contain semantic meaning. With the integration of a survey into the IBSP, all the variables are scrutinized, harmonized, decomposed. The Figures 1 shows an example of the naming framework.

Figure 1 – Example of the naming framework in the IBSP



Source: Statistics Canada, Quick Click Reference Guide for the IBSP Metadata

As an example, the question of revenues coming from *sales of fresh apple at the fruit market* is decomposed and harmonized with the IBSP Naming

Framework. In the example, the concept used is 'revenue'. In the IBSP, for all statistical programs, the concept of revenue is identified by cell numbers ranging from F43000 to F59999. The next component is the sub-concept. For all IBSP programs, the sub-concept of revenue named 'sales' is identified by cell numbers between F43000 and F45000. The third part call the modifiers provide a description of the variables. In the example here, the modifier is fruit market. The dimension component categorize the variables. The Census of Agriculture 2021 has adopted the corporate naming framework used in the IBSP. The 630 distinct variables of the Census of Agriculture cover 12 concepts, 21 sub-concepts, 86 modifiers and 15 type of dimensions and 331 unique dimensions. The Census of Agriculture shared common cells with 13 different programs.

This structure provides a fully coherent framework facilitating the integration of other programs data in the 'collect once-use multiple time' strategy or in the use of other programs data for validation or horizontal analysis. With this harmonized Naming Framework, the production of in-depth and timely analysis required by users to make informed, evidence-based decisions about increasingly complex issues that affect sectors bordering agriculture (e.g., food processing transportation, environment, rural affairs, international trade, and agricultural price indexes) is easier. There is no need to build complex transformations or concordances between variables in different programs or data sources for the integration of alternative data.

iii) Mandatory use of corporate tools and generalized systems

By adopting the corporate tools and systems used by the Economic Statistics programs, the Census of Agriculture will benefit, from the most state-of-the-art statistical leading-edge approaches introduced to improve methods, data quality and efficiency without having to invest locally to compensate for the obsolescence of systems. Whether it is for the design of the collection vehicle or for the management of the frame, to link efficiently datasets, to impute missing data, to tabulate estimates, to visualize and analyse data, to manage disclosure avoidance activities, to generate quality indicators and to publish results, the Census of Agriculture will now make use of all corporate harmonized systems and tools available.

The next Census of Agriculture will benefit directly from these advances in corporate systems:

a. The integration of the variance due to imputation methods into the generalized estimation system has made it more accessible. Its adoption by the IBSP as a standard output means that the Census now has a way to quantify a non-sampling error notion that was always known but never measured. Users will be better informed about the overall quality of the estimates.

- The development of agency-wide measuring quality of estimates that come from non-direct data sources (administrative data, satellite images, models, etc.) with methodologically acceptable approaches.
- The introduction of innovative approaches (ex. random tabular adjustment) to manage disclosure avoidance other than data suppression to enable a greater amount of data to be published, but without the need for complicated data perturbation of microdata records.
- The introduction of improved methods of data collection and case prioritization: The iterative derivation of a set of quality indicators to support actively the management of collection and follow-up activities is a new procedure replacing the more subjective static methods of the past to identify what records should be given priority for follow-up, when collection should cease in certain domains or even overall.
- A more standardized approach to pre-processing data prior to record linkage to ensure that the data sources that need to be linked are handled in a consistent and high-quality manner, thus improving the odds for successful linkages. New research in ways that record linkage can be undertaken without requiring manual verification, but still control the level of false results is being undertaken.
- The Corporate initiatives which can be used to coordinate work and knowledge transfer in machine learning, artificial intelligence among different parts of the agency. This also ensure a set of approved common methods, enhancing the coherence of the Census of Agriculture with other statistical programs.
- In addition, the use of common tools and databases facilitates the transfer and sharing of resources between programs. There are currently around 750 users of the IBSP tools and systems at Statistics Canada. Given the cyclical nature of the Census of Agriculture and the high demand of resources for a short period of time, this is an important element to ensure well trained staff is available when needed.

5. Discussion and Conclusion

The Census of Agriculture Program must adopt innovative, state-of-the-art methodological and operational approaches supporting the integration of multiple alternative data sources in its business model to take advantage of the new data landscape. It is essential to respond adequately and in a timely manner to data user needs. The use of the traditional approach of administrating a questionnaire to every member of the farm population to address the growing information needs is obsolete and has to be reviewed.

Statistics Canada is implementing the ambitious AG-Zero initiative to integrate information from alternate sources that will provide the data quality and details the agriculture sector needs with minimal contact with farmers.

The integration and the harmonization of the Census of Agriculture Program with the IBSP is an essential first step to position the Census of Agriculture strategically for the future. The use of common frame, concepts, methods and tools will improve the capacity to integrate alternative sources of data and the efficiency to produce outputs and analysis that extends beyond the primary sector. As well, the Census Program will benefit directly from the investments done at the corporate level in the development and implementation of leading-edge infrastructures, methods and tools, a non-negligible source of efficiency for the program.

The new harmonized processing framework introduced with the IBSP is anchored on the use of multiple standardized Corporate Services. The large number of actors involved in the multiple CEAG processes under the IBSP makes the coordination of all activities and the decisions making process slightly more complex than when all CEAG processes were managed locally. A strong governance enforcing the promotion of corporate goals over local preferences, well- defined service level agreements, performance and quality metrics for the various service centers are key to success.

References

1. Lee, J., Martineau, S. (2015). *Developing Metadata Standards in an Integration Projects at Statistics Canada*. Paper presented at the UNECE Workshop International Collaboration for Standard-Based Modernization. Geneva, Switzerland (5-7 May 2015).
2. Saint-Pierre, E. (2015). *Redefining roles and responsibilities in a new harmonized statistical production process: opportunities and challenges*. Paper presented at the UNECE Work Session on Statistical Data Editing. Budapest, Hungary (14-16 September 2015).
3. Statistics Canada. (2017). *Farm and Farm Operator Data*. Statistics Canada Catalogue no. 95-640-X. Ottawa.
4. Statistics Canada. (2015). *Integrated Business Statistics Program Overview*. Statistics Canada Catalogue no. 68-515-X. Ottawa.
5. Statistics Canada. (2019), *Quick Click Reference Guide for the IBSP Metadata*. Internal Document. Ottawa.
6. Thomassin, M. (2018). *The Migration of the Canadian Census of Agriculture to an Integrated Business Program Without Contact with Respondents*. Paper presented at the Fifth International Workshop on Business Data Collection Methodology. Lisbon, Portugal (19-21 September 2018).

The Philippine Census of Agriculture and Fisheries as part of the integrated agricultural statistical system

Minerva Eloisa P. Esquivias, Erma A. Aquino, Joyce Anne Marie M. Ruiz
Philippine Statistics Authority, Quezon City, Philippines

Abstract

Statistics have been increasingly recognized as a key element in crafting, monitoring and evaluating a country's development policies and programs. With the constant changes in global and local economy, greater demands for more responsive and relevant statistics are inevitable for all sectors. For one, the agriculture sector is important for food security, thus, a good national agricultural statistical system is imperative in all countries. This presentation will describe the census of agriculture as part of the integrated agricultural statistical system of the Philippine Statistics Authority (PSA). It aims to present the advantages of the integrated system in terms of cost and operational efficiency, as well as the issues and challenges that may arise.

As mandated by the Philippine Statistical Act of 2013, the PSA conducts census, surveys and other statistical activities to generate official statistics on agriculture, aquaculture and fisheries, including national accounts. The Census of Agriculture and Fisheries (CAF), which can provide information on the structure and characteristics of agricultural farms, aquaculture farms and fishery operations at the lowest geographical level, is considered as one of the key pillars of a national statistical system by the Food and Agriculture Organization. CAF can also cater to data needs related to the Sustainable Development Goals and Global Strategy to Improve Agricultural and Rural Statistics and to the Philippine Development Plan (PDP) 2017-2022 and the Ambisyon Natin 2040 in particular.

Consistent with the 2020 World Programme for the Census of Agriculture (WCA), the Philippines' CAF is a vital source of core data, and sample frame for the more frequent and more in-depth agriculture, aquaculture and fishery surveys of PSA. These surveys include Palay and Corn Production Survey, Backyard Livestock and Poultry Survey, Commercial Livestock and Poultry Survey, and Aquaculture Production Survey, among others. CAF addresses demand for indicators resulting from emerging developmental concerns and climate change as outlined in the Philippine Statistical Development Program 2018-2023, which was formulated under the direction of the Inter-Agency Committee on Agriculture and Fishery Statistics. To complement data from census and surveys, the PSA gathers administrative information from data producing agencies in the agriculture and fishery sectors through linkages and close coordination. Given the decennial conduct of the census, the PSA

conducts listing activities to update the statistical frame for the periodic surveys. Survey results can also serve as input data for the subsequent census.

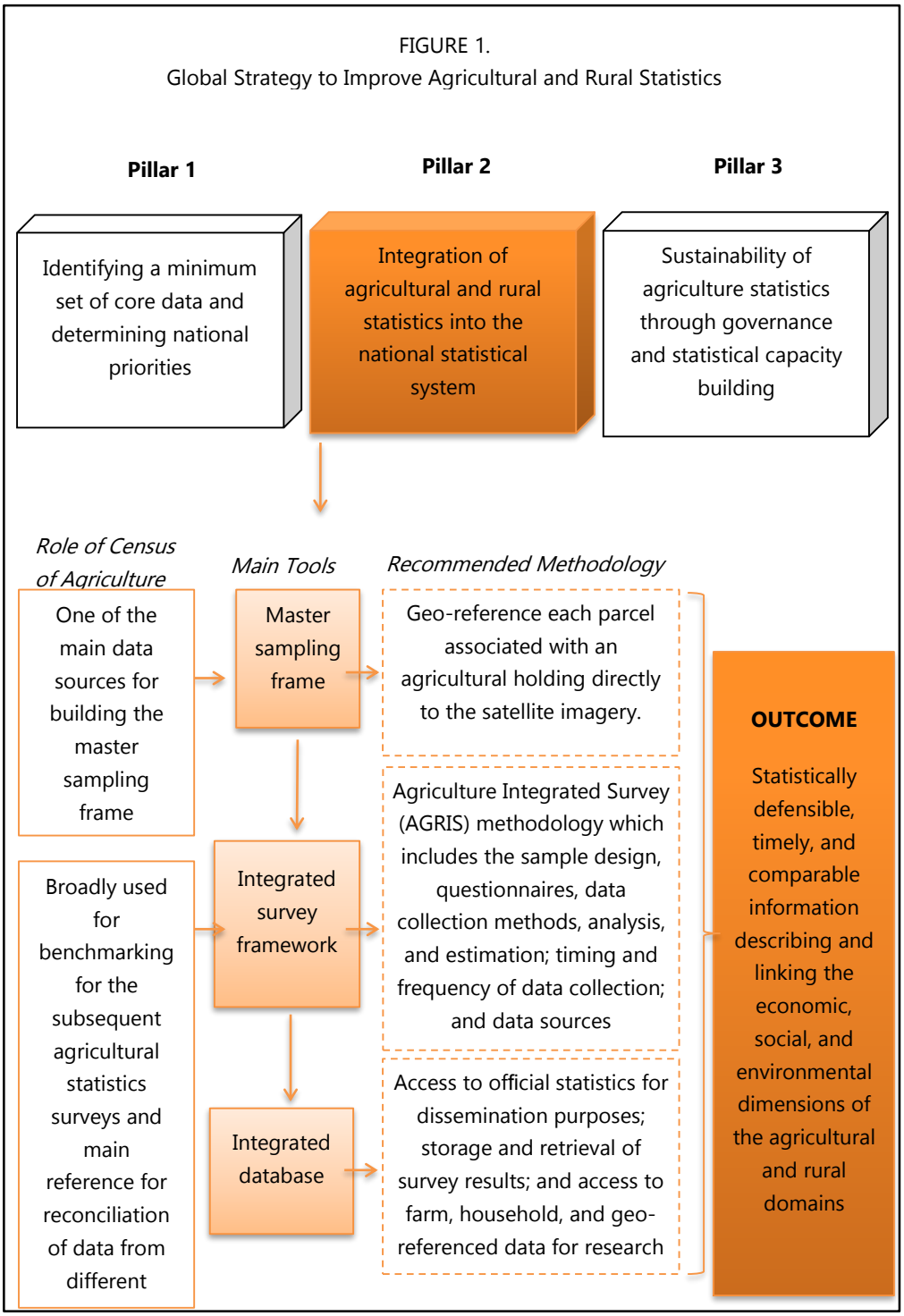
Keywords
Census of Agriculture and Fisheries; Integrated agricultural statistical system; Census and surveys

1. Introduction

The process of improving agricultural statistics will begin with the integration of agriculture into the national statistical system. In recent years, increasing efforts have been made towards achieving better integration of statistical activities. Integration, in a statistical sense, means that each statistical collection is carried out, not in isolation, but as a component of the national statistics system[1]. This is the groundwork of the second pillar and the most significant recommendation in the Global Strategy to Improve Agricultural and Rural Statistics (GSIARS) - to integrate agriculture into the national statistical system.

[1] Food and Agriculture Organization (2010). Global Strategy to Improve Agricultural and Rural Statistics.

FIGURE 1.
Global Strategy to Improve Agricultural and Rural Statistics



Sources:

1. *Food and Agriculture Organization (2012). Action Plan of the Global Strategy to Improve Agricultural and Rural Statistics.*
2. *Food and Agriculture Organization (2017), World Programme for the Census of Agriculture 2020 Vol. 1.*

The development of a master sampling frame (MSF) for agriculture is viewed as an essential element of this integration, highlighting the census of agriculture as one of the main data sources for building the MSF. One of the important methodological developments under the Global Strategy is the elaboration of an Agriculture Integrated Survey (AGRIS) methodology and toolkit for cost-effective approaches for collection of relevant agricultural and rural data on a regular basis. This will assist countries with underdeveloped agricultural census and surveys programmes to move towards a fully integrated approach. In an integrated census and surveys programme the census of agriculture is broadly used for benchmarking for the subsequent agricultural statistics surveys and is a main reference for reconciliation of data from different surveys and sources.

Further, in an integrated statistical system, the connection between the population and housing census and the agriculture census is strong, and as such it is useful to look at ways to strengthen the relationship between the two censuses. At the operational level, linking the population and housing census with the agriculture census is more suitable for countries where both censuses are carried out as a household enquiry[2].

Advantages of integration include production of coherent and manageable set of items knowing that related data are comparable with other sources. Through integration, concepts, definitions and classifications used in the different statistical activities can be made compatible, hence, will facilitate data analysis across sectors/collections. On the administrative side, integration will avoid duplication of statistical activities, prevent the release of conflicting statistics, ensure the best use of resources (human, financial, etc.), reduce the burden of response, and enable agriculture to be an integral part of statistical planning and budgeting processes. WCA 2020 promotes the development of an integrated statistical system to respond on resolving these issues.

GSIARS provides the framework for integration of the agricultural statistical system into the preparation of the National Strategy for Development of Statistics (NSDS), putting an emphasis on the census of agriculture as a main pillar of the national programmes. Countries are encouraged to design a Strategic Plan for the Development of Agricultural and Rural Statistics (SPARS) mainstreamed into the NSDS process. An integrated agricultural statistics system involves a multi-year programme of statistical activities, including an agricultural census and agricultural surveys, to provide all the data required.[3]

[2] Food and Agriculture Organization (2017), World Programme for the Census of Agriculture 2020 Vol. 1.

[3] Food and Agriculture Organization (2017). World Programme for the Census of Agriculture 2020 Vol. 1.

This paper aims to present the current structure and progress of integration of agricultural statistical system in the Philippines and role of CAF in the integration.

2. Results

Agricultural Statistical System in the Philippines

The Philippine Statistical System (PSS) is a decentralized system that ensembles all statistical organizations at all administrative levels, its personnel and the national statistical development program. One of its main components for an effective and efficient national statistical system is the management and coordination mechanism within the government. It requires strategic and coherent planning, bringing together various stakeholders towards achieving the shared goal of a better statistics, better policies and better lives. The PSS formulates the Philippine Statistical Development Program (PSDP) which consists all statistical activities to be undertaken in response to the requirements of development planning and policy formulation. The PSS takes into consideration the PDP 2017-2022 to respond to the new and emerging statistical requirements of the government and private sector. The PDP 2017-2022 envisions to expanding economic opportunities in the agriculture sector and to increasing access to economic opportunities for small farmers and fisherfolk.

Challenges in Previous Structure

Before September 12, 2013, two separate agencies in the national government were responsible for the production of agricultural and fisheries statistics:

1. Bureau of Agricultural Statistics (BAS), an attached agency of the Department of Agriculture (DA)
   - focal agency for the conduct of agriculture and fisheries surveys
   - served as the central information source and server of the National Information Network of the DA
   - provided technical assistance to end-users in accessing and analyzing product and market information and technology
2. National Statistics Office (NSO), an attached agency of the National Economic and Development Authority
   - responsible for the conduct of censuses of agriculture and fisheries that provide structural data on agriculture and fisheries and frame for agriculture surveys
   - provided government planners and policy-makers with data on which to base their plans for the country's development

Another agency, the National Statistical Coordination Board (NSCB), served as policy-making and coordinating body in the PSS. NSCB was responsible for

the compilation of the production and expenditure accounts, including of agriculture sector, on a quarterly and annual basis.

In 2008, Special Committee of experts who conducted an external review of the PSS reported that there are a number of structural limitations in the previous setup of the PSS. These limitations, coupled with constraints on financial, physical, human, and other resources, have hindered the PSS from responding quickly to users' requirements and criticisms, especially regarding the vast need for statistics for local development planning. This led to the recommendation of reorganization of the PSS so that there will be a central statistical authority in charge of coordinating the country's data activities.

Financial limitation was evident in BAS which had a centralized budget from the Department of Agriculture (DA) allocated to the Technical Divisions and Regional/Provincial Operation Centers for the agricultural surveys and other statistical activities. Statistical programs of BAS used to compete with DA programmes and projects, that were higher in priority and resulted to resource constraints.

Current Agricultural Statistical System in the Philippines

On September 12, 2013, Republic Act 10625 or the "Philippine Statistical Act of 2013" created the Philippine Statistics Authority (PSA) through a merger of the following four major statistical agencies in the government:

- Bureau of Agricultural Statistics (BAS)
- National Statistics Office (NSO)
- Bureau of Labor and Employment Statistics (BLES)
- National Statistical Coordination Board (NSCB)

The PSA is at the forefront in providing quality statistics on agriculture, livestock and fishery sector. As mandated by RA 10625, the PSA conducts censuses, surveys, special studies and other statistical activities, generates key indicators and serve as the source of official agricultural statistics.

Considering that the sector is a dynamic one where new developmental concerns continue to emerge, statistical development programs are proposed to be implemented in the medium term by the PSA. The statistical programs and activities are designed to provide vital information support for the Medium-Term Philippine Development Plan (MTPDP) thereby putting an orderly direction towards sustained improvement in the agricultural statistical system. The plans were crafted in consultation with the Inter-Agency Committee on Agriculture and Fishery Statistics (IACAFS) which is represented by the key institutions with stake in the agriculture, livestock and fishery sector. Primarily, it serves as a forum for the exchange of views and expertise to resolve technical issues and problems arising from the production, dissemination, and use of agriculture and fishery statistics. [4]

[4] Philippine Statistical Development Program 2018-2023

Philippine Census of Agriculture and Fisheries and Role in Integration

Over the years, the CAF has been a source of comprehensive statistics on agriculture and fisheries for the use of the general public, government, business industry, and research and academic institutions.

FIGURE 2. Contribution of Census of Agriculture and Fisheries (CAF) in the Agricultural Statistical System

<table>
<tr>
<td rowspan="2">Statistics on the distribution of agricultural lands, structure of agricultural, aquaculture, and fishing operations, and other information which shall serve as framework for the country's development programs;</td>
<td>Inventory of the country's agricultural and fishery resources,</td>
<td rowspan="2">Statistics for local area planning: barangay, city/municipality, provincial, and regional levels;</td>
<td>Sampling frame for the various surveys in agriculture and fisheries;</td>
</tr>
<tr>
<td>Benchmarks for various statistical series which are designed to measure progress in the agricultural and fishery sectors;</td>
<td>Data for monitoring the progress of the country towards attainment of the Sustainable Development Goals (SDG) and Ambisyon Natin 2040</td>
<td>Information needed by the United Nations Organizations for international comparability of data and monitoring the world food situation.</td>
</tr>
</table>

Source: 2012 CAF Enumerator's Manual, Philippines

TABLE 1.
Role of CAF in Major Agricultural Statistical Activities in PSA (2017-2019)

	Activities	**Role of 2012 CAF/Use for 2022 CAF**	**Alignment to GSIARS**
Completed	<u>Updating of frame</u> To provide reliable statistical frame for the conduct of agricultural and fisheries surveys, the following updating/listing activities were conducted. • 2017 Listing of Farm Households (LFH) • Updating of List of Aqua Farms (ULAF)	**2012 CAF** provided the basic frame and was used as basis for updating frames for the following surveys: • 2017 LFH was used for Palay/Corn Production Survey (PCPS) and Backyard Livestock and Poultry Survey (BLPS) • ULAF was used for Aquaculture Survey	

Completed	Enhancement of current design of surveys and proposal of new design of surveys for PCPS, BLPS, Inland Fisheries Survey, Municipal Fisheries Survey, and Aquaculture Survey.	**2012 CAF** was used as benchmark and frame in the redesigned agriculture and fisheries surveys	
On-Going	Rebasing of national accounts including the national accounts to enable provision of "more comprehensive" data set to capture emerging developments in the economy	**2012 CAF** will be utilized as basis for updating benchmarks for national accounts as inputs to agricultural production, inventory of permanent crops, inventory of livestock and poultry, and validation of household and establishment farm activities.	
On-Going	Ongoing geo-tagging of building structures attempts to bridge the integration process between geospatial and statistical information from census/survey.	Geo-referenced structures will be used in the **2022 CAF** operations. The development of the sample frame for agriculture surveys will be made using the 2022 CAF data.	Geo-referenced households with 2020 CPH data will be used as frame for **2022 CAF** and will be linked to digitized/satellite images of agricultural parcels. This complements the global strategy to develop the master sample frame for agriculture.
Early Stage	Pilot study with Asian Development Bank (ADB) on development of digitized agricultural parcels/area frame and area/yield estimation using satellite images/remote sensing.	The ADB collaboration will capacitate PSA for the **2022 CAF** in terms of determining crop area	
On-Going	Learning sessions on review of 2020 WCA participated by staff of Agriculture and Fisheries Census Division (AFCD), Census Planning and Coordination Division (CPCD), Crops Statistics Division (CSD), Livestock and Poultry Statistics Division (LPSD), Fisheries Statistics Division (FSD), Agricultural Accounts Division (AAD), and Expenditure Accounts Division (EAD)	To introduce the recommended data items in 2020 WCA and harmonize concepts, definitions, and classifications on structural characteristics of agriculture, aquaculture and fishery operations to the personnel of PSA in-charge of conduct of **agriculture and fisheries census** and surveys.	Standardized concepts, definitions, and classifications across **agricultural and fisheries census** and surveys will facilitate comparability of statistical data in agriculture and fisheries

Another approach to integration as recognized by Busan Action Plan to increase reliability and accessibility of official statistics focuses on the synergies between survey, census data, administrative data and vital statistics.

Continuous collaboration with regulatory agencies [5] through forging of Memorandum of Agreements (MOAs) and bilateral meetings increased the use of administrative-based data as supplement and/or alternative sources of data on agriculture and fishery.[6]

3. Discussion and Conclusion

The recent PSS organizational structure, with one statistical office mandated to produce most data on agriculture and fishery statistics, is an advantage towards the fulfilment of an integrated agricultural statistical system. Financial, human, and material resources, as well as operational procedures and concepts, would be more efficiently managed. Having a PSA budget with an AgriStat component is significant for the production of primary statistics in agriculture since there is no further competing claims from other 'non-statistical' programmes upon budget release by the Department of Budget and Management.[7] The continuously increasing demand to deliver more timely, comparable and relevant agriculture statistics, however, still poses risks on fund sources for the redesign and further innovation efforts for agricultural and fisheries census and surveys, including acquisition of required information technology infrastructure and human resource complement.

Moreover, while management of PSS is better in the new structure, generation of agriculture and fisheries statistics are challenged by the following factors:

- Increasing demand to deliver more timely, granular and relevant agriculture statistics
- 'Exodus' of some trained and experienced senior statisticians from Major Statistical Agencies
- Availability of new technology
- Budget approval process
- Communicating agriculture statistics to policymakers and stakeholders for better appreciation and utilization

Despite the challenges, efforts towards integration of agriculture statistical systems continues to progress. Geo-referencing to master frame is given the utmost importance which will enable linking of census of population with CAF

[5] Regulatory agencies include the National Meat Inspection Service (NMIS), Bureau of Animal Industry (BAI), National Dairy Authority (NDA), Philippine Carabao Center (PCC), Philippine Fisheries Development Authority (PFDA), Bureau of Fisheries and Aquatic Resources (BFAR), National Food Authority (NFA), Philippine Coconut Authority (PCA), Sugar Regulatory Administration (SRA), National Tobacco Administration (NTA), Philippine Fiber Development Authority (PhilFiDA), Local Government Units (LGUs)

[6] Philippine Statistical Development Program 2018-2023

[7] Africa, T.P. (2015). Agriculture and Rural Statistics in the Philippine Statistical System presented at the Workshop on Strategic Planning for Agricultural and Rural Statistics, Bangkok, Thailand, 17-19 March 2015.

and subsequently to agriculture and fishery surveys. Further, capacity building activities and initiatives for application of remote sensing to crop area estimation will enable PSA staff to adopt another geospatial technology in the production of reliable agriculture statistics. PSS likewise envisions to streamline and to create integrated framework for surveys on agriculture and fisheries and thereof integration of database, as these will improve coherence of agricultural data, lead to optimal use of financial resources and conform with the recommendations of GSIARS. In terms of delivery of agricultural data products, capacity building on communicating the importance of agriculture statistics is included in the preparation of 2022 CAF in hope to strengthen evidence-based policy making and to improve project design and implementation in the sector of agriculture. Ultimately, the integrated agricultural system aims to provide some critical and lower-level disaggregated data for monitoring SDGs and Ambisyon Natin2040 particularly towards the Philippines becoming a prosperous, predominantly middle-class society where no farmer or fisherman is poor or hungry.

ANNEX A

Indicators from Agricultural Censuses and Surveys

Census/Survey	Statistics/Indicators
Census	
Census of Agriculture and Fisheries (CAF)	A. Agriculture Characteristics of Operators, Size of Farm/Holding, Main Use of Land, Tenurial Status of the Farm/Holding, Area with Irrigation Facility, Largest Area Planted by Major Crop, Inventory of livestock and poultry raised B. Aquaculture Characteristics of Aquafarm Operators, Type of Aquafarm, Size of Aquafarm, Species Cultured in the Aquafarm C. Fisheries Characteristics of Fishing Operators, Commercial and Municipal Fishing Operators, Highest Gross Tonnage of Fishing Boats/ Vessels Used, Number of Fishing Gears by Type D. Community Proportion of barangays with agriculture/fisheries facility in the barangay by type of facility
Crops Surveys	
Palay and Corn Production Survey (PCPS)	• area planted /harvested and production by ecosystem (palay) and croptype (corn); • monthly distribution of production and area harvested; • farm household disposition/utilization of production; planting intentions indicator; • area with standing crops;

Census/Survey	Statistics/Indicators
	• use of seeds, fertilizers, and pesticides; and • awareness and availment of program interventions.
Monthly Palay and Corn Situation Reporting System (MPCSRS)	• estimates monthly palay and corn based on standing crops • estimates monthly palay and corn based on planting intentions
Palay and Corn Stocks Survey (PCSS)	• stock level of rice and corn at the household level § estimates of the current stock of rice and corn in farm and non-farm households
Crops Production Survey (CrPS)	• quarterly volume of production covering 200 crops, with 19 as major crops. The major and priority crops by sub-group are vegetables, root crops, fruit crops, non-food and industrial crops, and ornamental plants • area harvested/planted • number of bearing trees/hills/vine
Agricultural Labor Survey (ALS)	• daily wage rates of palay, corn, coconut and sugarcane farm workers
Farm Prices Survey	• data on prices received by producers for cereals, vegetables and legumes, rootcrops, fruits, commercial crops, livestock, poultry and fishery
Livestock and Poultry Surveys	
Backyard Livestock and Poultry Survey (BLPS)	• inventory of animals by farm type, by age and by classification • chicken by type • production of livestock and poultry • volume of egg produced
Commercial Livestock and Poultry Survey (CLPS)	• inventory of animals by farm type, by age and by classification • chicken by type • production of livestock and poultry • volume of egg produced
Survey of Slaughterhouses and Poultry Dressing Plants	• number of heads slaughtered/dressed
Dairy Production Survey	• milk production and disposition • average price of milk produced per liter
Fisheries Surveys	
Quarterly Commercial Fisheries Survey	• data on volume and value of production by species from commercial fishing
Quarterly Municipal Fisheries Survey	• data on volume and value of production by species from municipal fishing

Census/Survey	Statistics/Indicators
Quarterly Aquaculture Survey	• quarterly volume and value of aquaculture production
Quarterly Inland Municipal Fisheries Survey	• generates data on volume and value of all species from inland fishing

ANNEX B

Milestones in the Philippine Agriculture Statistics[8]

1. Enhancement of agricultural and fishery surveys in terms of survey design, coverage, data collection, processing and data management.
    a. Re-design of the quarterly surveys – Palay Production Survey, Corn Production Survey, Backyard Livestock and Poultry Survey, Aquaculture Survey, Inland Fishing Survey, Municipal Fishing Survey
    b. Updating/improvements of the processing systems
    c. A Study of the Integration of the Survey of Food Demand (SFD) for Agricultural Commodities into Family Income and Expenditure Survey (FIES)
2. Conduct of frame updating activities
    a. Conduct of 2017 Updating of List of Aqua Farms (ULAF)
    b. Conduct of 2017 Listing of Farm Households (LFH)
    c. Preparation for Conduct of Updating of List of Landing Centers
3. Conduct of Survey on Costs and Returns of Tomato Production
4. Conduct of SFD for Agricultural Commodities in 2018
5. Strengthening of advocacy on understanding and rational use of agriculture and fishery statistics and indicators
    a. Livestock and Poultry Information Early Warning System (LPI-EWS)
6. Use of administrative-based data systems as supplement/alternative sources of data on agriculture and fishery
7. Conduct of capacity-building on agricultural statistics data system, data review/validation.
8. Updating of technical conversion ratios/parameters for production estimation through the conduct of the following activities:
    a. Updating the Food Balance Sheet (FBS) parameters for livestock and poultry parameters
    b. Updating of the Milling Recovery Rate (MRR) of Rice
9. Conduct of orientations, capability development and feedback sessions for the Provincial Lead Implementors of the IT-enabled Maturity

[8] Philippine Statistical Development Program 2018-2023

Assessment (ITeMA) Program at the Agrarian Reform Beneficiaries Organization (ARBO) level

10. Conduct of capability development on the use of the following monitoring and reporting systems: Land Tenure Improvement (LTI) Operational Tool (Optool), Legal Case Monitoring System (LCMS) and Program Beneficiaries Development (PBD) Monitoring and Evaluation System

ANNEX C

References

1. Africa, T.P. (2015). Agriculture and Rural Statistics in the Philippine Statistical System presented at the Workshop on Strategic Planning for Agricultural and Rural Statistics, Bangkok, Thailand, 17-19 March 2015.
2. Ericta, C.N. (2012). Restructuring the Philippine Statistical System in Response to New Challenges: Redefining the Role of the National Statistics Office in the System presented at the 13th East Asian Statistical Conference, Tokyo, Japan, 5-7 November 2012.
3. Esquivias, M.E.P. (2016). Strategic Changes in the Agricultural Statistics System in the Philippines, Thimphu, Bhutan, 15-19 February 2016.
4. Food and Agriculture Organization (2017). World Programme for the Census of Agriculture 2020 Vol. 1.
5. Food and Agriculture Organization (2010). Global Strategy to Improve Agricultural and Rural Statistics.
6. Philippine Statistical Development Program 2018-2023
7. Philippine Statistics Authority (2012). Census of Agriculture and Fisheries Enumerator's Manual.
8. Recide, R.S. (2010). Food and Agricultural Statistics in the Philippines presented at the Twenty-Third Session of the Asia and Pacific Commission on Agricultural Statistics, Siem Reap, Cambodia, 26-30 April 2010.
9. The World Bank Group in the Philippines (2010). Assessment of the Philippine Statistical Development Program 2005-2010 Volume I: Main Report.

A technique for outliers detection in linear functional relationship model for circular variables

Abdul Ghapor Hussin[1], Nurkhairany Amyra Mokhtar[1], Yong Zulina Zubairi[2], Mohd Iqbal Shamsudheen[3]
[1]National Defence University of Malaysia
[2]University of Malaya
[3]University College London

Abstract
The occurrence of outlier may be due to error, or part of the phenomena under study. This paper discusses on outlier detection methods are discussed using difference mean circular error cosine statistic for circular variables. Here, we focus on a model with linear functional relationship model in which the variables are considered with equal concentration of their error terms. The cut-off equation for outlier detection is obtained by using row deletion approach and it is then tested to detect the outlier in a simulation study. The power of performance of this method increases as the concentration parameters of the errors and the level of contamination for the outlier increase. The applicability of this method is illustrated by using real wind direction data.

Keywords
Outlier detection; Circular variables; Row deletion; Power of performance; Simulation study

1. Introduction

Directional data arises quite frequently in many natural and physical sciences. The directions may be in two-dimensional or in three-dimensional. Observations on two-dimensional directions can be referred as circular data meanwhile the observations on three-dimensional directions can be referred as spherical data (Jammalamadaka and Sengupta (2001)).

An example of circular data is the data of wind directions. The distribution of the directions may arise either as a conditional distribution for a given speed, or as a marginal distribution of the wind speed and direction. The Von Mises distribution is said to be the most useful distribution on the circle (Mardia and Jupp (2000)). Fisher (1987) noted that the Von Mises distribution is a symmetric unimodal distribution and characterised by a mean direction μ and concentration parameter κ. The probability density function of the distribution is

$g(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)}$ where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero, which can be defined by $I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi}e^{\kappa\cos\theta}d\theta$ for $0\le x<2\pi$, $0\le\pi<2\pi$ and $\kappa>0$ where μ is the mean direction and κ is the concentration parameter.

In this paper, the circular data is described in a linear functional relationship model given by $Y=\alpha+X(\mathrm{mod}\,2\pi)$ for the rotation parameter α. In this model, both *X* and *Y* variables are subject to random errors δ_i and ε_i, respectively (Caires ans Wyatt (2003)).

However, the existence of outlier in the data may lead to a different model. This paper discusses on outlier detection method using difference mean circular error cosine (*FDMCEC*) statistic. Section 2 describes the methodology and Section 3 describes the result for the power of performance of the method.

2. Methodology

The *cosine* statistics is known as the functional mean circular error (*FMCEC*) in which the statistic is given by $FMCEC=1-\frac{1}{2n}\left[\sum_{i=1}^{n}\cos\left(x_k-\hat{X}_k\right)+\sum_{i=1}^{n}\cos\left(y_i-\hat{Y}_i\right)\right]$ where *n* is the sample size, $\hat{Y}_i$ is the estimated value of y_i and $\hat{X}_i$ is the estimated value of x_i, under the parameter estimation of the unreplicated LFRM, depending on the case either equal or unequal error concentration.

The approach of row deletion is applied in this method in which we find the absolute difference of the mean circular error when an observation is deleted one after another. Thereafter, $FMCEC_{(-i)}$ denotes the removal of i^{th} observation. The absolute difference between the value of full data set and the reduced data set is as given by $FDMCEC_{(-i)}=\left|FMCEC-FMCEC_{(-i)}\right|$.

The existence of outlier in *x* and *y* will give a large value of the $FMCEC_{(-i)}$ statistics. An i^{th} observation is defined as an outlier if the value of $FDMCEC_{(-i)}$ exceeds the cut-off equation. To detect oulier, we need to determine the cut-off equation as the indicator for a particular observation to be remarked as the outlier. Hence, a Monte Carlo simulation study is carried out with different values of sample size and error concentration parameter.

In doing so, we set the number of simulation *s* = 500 (Ibrahim et al. (2013)). Without loss of generality, the variable *X* is generated from the von Mises distribution and we set the value of α with $\alpha=\frac{\pi}{4}=0.7854$. The values of the concentration parameters of the error term used in this study are

$\kappa = 5, 10, 15 \text{ and } 20$. For each value of κ, the sample size n =20, 30, 50, 70, 130 and 150 are considered for the simulation. With the assumption of $\kappa = \nu$, the procedures are described below.

Step 1: Generate the values of X variable from the von Mises distribution of $VM(2,3)$ and for the size of n = 20, 30, 50, 70, 100, 130 and 150; and $\kappa = 5, 10, 15 \text{ and } 20$, respetively. Find Y according to the generated X based on the model $Y = \alpha + X \pmod{2\pi}$.

Step 2: The variables X and Y are considered with generated random error terms of where $x_i = X_i + \delta_i$ and $y_i = Y_i + \varepsilon_i$ for $i = 1,2,...,n$. The error terms are $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\nu)$, respectively where $\kappa = \nu$. The variables are fitted to LFRM with parameter estimation as described in Section 4.2.

Step 3: The values of functional mean circular error cosine (*FMCEC*) are calculated for all observations. The estimation of X for this equal error concentration case is given by

$$\hat{X}_{i1} \approx \hat{X}_{10} + \frac{\sin\left(x_i - \hat{X}_{i0}\right) + \sin\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)}{\cos\left(x_i - \hat{X}_{i0}\right) + \cos\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)}.$$

Step 4: Omit the i^{th} observation of the generated data, where i=1, 2, 3, ...,n to obtain $FMCEC_{(-i)}$. Repeat this step for all i observations to obtain the set of values for $FMCEC_{(-i)}$.

Step 5: Calculate the absolute difference between *FMCEC* and $FMCEC_{(-i)}$. Then, find value of $FDMCEC_{(-i)} = \left(\left|FMCEC - FMCEC_{(-i)}\right|\right)$ for all i.

Step 6: Repeat steps 1-5 for 500 simulations for each n and κ and note the values 5% upper percentiles of the $FDMCEC = \max\left(\left|FMCEC - FMCEC_{(-i)}\right|\right)$ to construct the cut-off equation based on the significance level of interest. These values of upper percentiles may be used as the cut-off equations in identifying the outlier for the unreplicated LFRM for equal error concentration parameters. Table 1 shows the values of *FDMCEC* based on 5% upper percentile.

Table 1 The values of 5% percentile of *FDMCEC* for equal error concentration

n	κ =5	κ =10	κ =15	κ =20
20	0.0256	0.0103	0.0068	0.0050
30	0.0158	0.0080	0.0050	0.0037
50	0.0198	0.0052	0.0034	0.0024
70	0.0122	0.0036	0.0023	0.0019
100	0.0119	0.0029	0.0019	0.0013
130	0.0091	0.0023	0.0015	0.0011
150	0.0081	0.0018	0.0013	0.0010

For *FDMCEC* of each value of κ, we find the best fit using the least square method to obtain the power series equation. Figures 1 to 3 show the power series graphs of 5% upper percentile for concentration parameters are 10, 15 and 20, respectively with their cut-off equation *y*. We consider 95% confidence level and thus the cut-off equation is to be at 5% significance level.

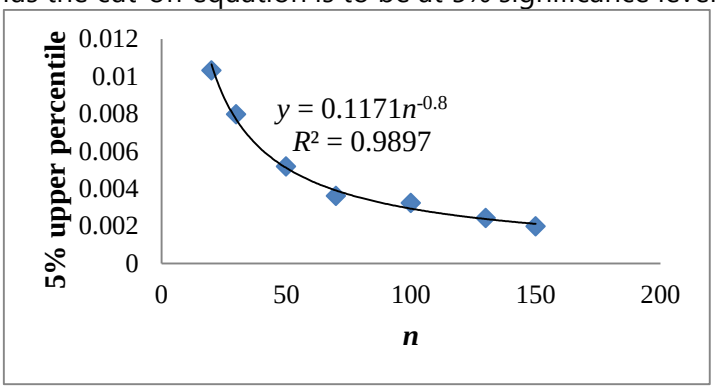


Figure 1 The power series graph for *FDMCEC* to determine the cut-off equation for $\kappa = 10$

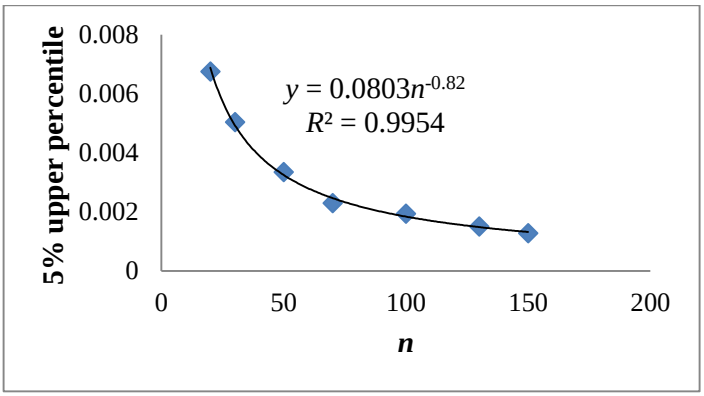


Figure 2 The power series graph for *FDMCEC* to determine the cut-off equation for $\kappa = 15$

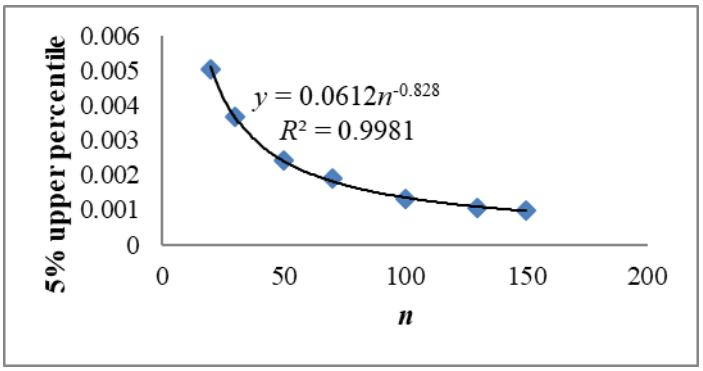


Figure 3 The power series graph for *FDMCEC* to determine the cut-off equation for $\kappa = 20$

3. Results

To assess the power of performance of the cut-off equations developed in the previous section, another simuation study is done with an outlier planted to the generated data set. The steps in the simulation are described and the

coding of the programming is developed using Tibco SPLUS statistical software in assessing the power of performance of *FDMCEC* statistic in detecting the outlier.

Step 1: The values of X variable are generated from the von Mises distribution and in the size of n = 70, 100 and 130 and $\kappa = 10, 15 \text{ and } 20,$ respectively. An observation X_d^* is then contaminated with some levels of contamination ω where the level of the contamination are ω = 0.2, 0.4, 0.6, 0.8 and 1, respetively. The formula of contaminating the observation is as follow:

$$X_d* = X_i + \omega\pi \pmod{2\pi}$$

Step 2: Find Y according to the generated X. The variables X and Y are considered with generated random error terms of $\delta_i \sim VM(0, \kappa)$ and $\varepsilon_i \sim VM(0, \nu)$, respectively where $\kappa = \nu$. The variables are fitted to the unreplicated LFRM with parameter estimation as described Section 4.2.

Step 3: The values of functional mean circular error cosine are calculated for all observations.

Step 4: Omit the i^{th} observation of the generated data, where i=1, 2, 3, …,n to obtain $FMCEC_{(-i)}$. Repeat this step for all i observations to obtain the set of value $FMCEC_{(-i)}$.

Step 5: Calculate the absolute difference between *FMCEC* and $FMCEC_{(-i)}$. Then, find value of $FDMCEC_{(-i)} = \left(\left| FMCEC - FMCEC_{(-i)} \right| \right)$ for all i.

Step 7: Determine the values of $FDMCEC_{(-i)}$ that exceed the cut-off equations developed in Section 6.3. If they exceed, they are marked as outliers.

Step 8: Steps 1 to 7 are repeated for 500 simulation and the percentage of correct outlier detection is calculated as the power of performance. Table 6.7 shows the power of performance of *FDMCEC* in outlier detection**.**

Table 6.7 The power of performance of *FDMCEC* in outlier detection

n	ω	κ =10	κ =15	κ =20
70	0.2	6.80	9.20	13.80
	0.4	24.60	50.00	73.60
	0.6	74.60	96.20	99.20
	0.8	99.00	99.80	100.00
	1	99.60	100.00	100.00
100	0.2	5.60	6.40	14.80
	0.4	22.00	49.40	69.00
	0.6	72.00	93.80	99.60
	0.8	97.00	100.00	100.00
	1	99.80	100.00	100.00
130	0.2	6.40	9.00	12.00
	0.4	22.00	45.80	66.20
	0.6	65.80	91.40	99.80
	0.8	96.80	100.00	100.00
	1	99.60	100.00	100.00

For better understanding, Figure 4 shows the pattern of the power of performance when n = 100.

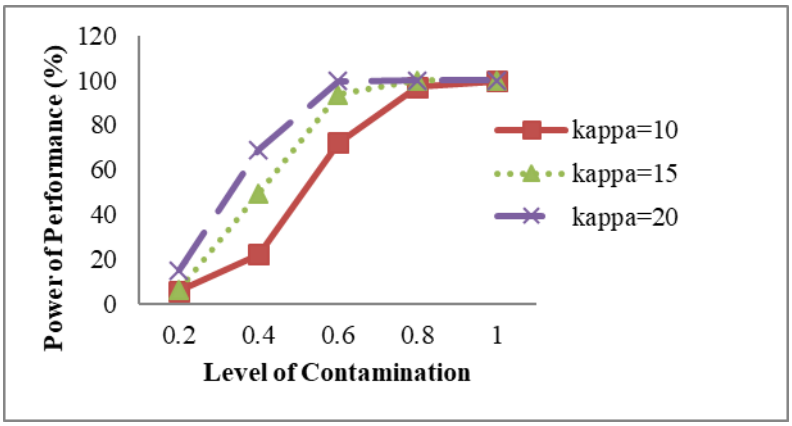


Figure 4 Power of performance for *FDMCEC* in detecting outliers for equal error concentration when n is 100.

4. Application to Real Wind Direction Data

The method above is illustrated using real wind direction data from Holderness coastline, Humberside Coast, UK, developed by UK Rutherford and Appleton Laboratories, with the sample size of 129. Variable x is the data measured by the techniques of HF radar system. It uses pulse radar and operates at frequency of 24.2-27 MHz. Meanwhile the variable y is measured

by using the technique of anchored wave buoy. Previous researchers of circular statistics such as Mokhtar et al. (2018), Abuzaid et al. (2008), Hussin et al. (2010) and Satari (2015) have used this data to illustrate the presence of outliers. It is worthwhile to note that the values of error concentration parameters of the variables x and y are assumed as equal. They have established that observations 38 and 111 as outliers of the data set. Figure 5 Values of FDMCEC for all 129 observations of the Humberside Coast wind direction data.

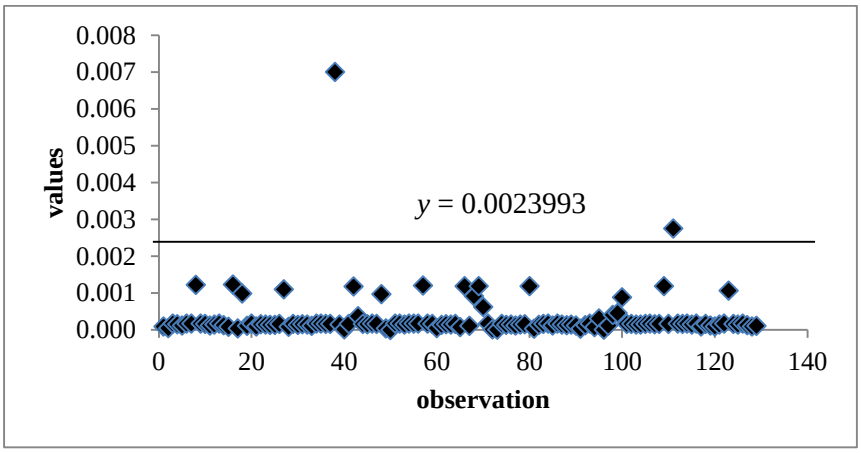


Figure 5 Values of *FDMCEC* for all 129 observations of the Humberside Coast wind direction data

5. Conclusion

To conclude, this paper discusses on outlier detection in linear functional relationship model of circular variables with equal. The functional difference mean circular error is proposed for the outlier detection. Simulation studies are carried out to obtain the cut-off equation for outlier detection and to obtain the power of performance of the method. The performance of the method increases as the concentration parameter and the level of contamination increase.

Acknowledgement

We would like to thank National Defence University of Malaysia and University of Malaya (grant number: GPF006H-2018) for supporting this work.

References

1. Abuzaid A. H. M. (2010) Some Problems of Outliers In Circular Data, PhD Thesis, University of Malaya.
2. Caires, S. and Wyatt, L. R. (2003). A Linear Functional Relationship Model for Circular Data with an Application to the Assessment of Ocean Wave Measurement. *American Statistical Association and the Internal Biometric*

Society Journal of Agricultural, Biological and Environmental Statistics: 8 (2),153-169.

3. Fisher, N. I. (1987). Problems with the Current Definitions of the Standard Deviation of Wind Direction. Journal of Climate and Applied Meteorology: 26, 1522-1529.
4. Jammalamadaka, S. R. and Sengupta, A. (2001). Topics in Circular Statistics, World Scientific Publishing.Fisher, N. I. (1993*). Statistical Analysis of Circular Data*, Cambridge University Press.
5. Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*, John Wiley & Sons.
6. Mokhtar, N. A., Zubairi, Y. Z. and Hussin, A. G. (2018). A clustering approach to detect multiple outliers in linear functional relationship model for circular data. Journal of Applied Statistics, 45.6: 1041-1051.
7. Satari, S. Z. (2015). Parameter Estimation and Outlier Detection for Some Types of Circular Model, PhD. Thesis, Universiti Malaya.

Use of big data analytics in targeting right customers for sustainable and social finance

Liza Mydin, Jamal Arif Jamaluddin, Zulfadzli Zaini
Maybank Islamic

Abstract

According to the Islamic perspective, a society can only thrive if all individuals are given access to resources and equality of opportunity to achieve a decent standard of living. These individuals become active market participants and consume, produce or trade and the resulting income gained are considered justly earned. However Islam ordains for a portion of this income to be distributed through a "redemption of rights" of those within the population who are unable to actively participate in earning an income due to the circumstances of their lives such as disabilities, injuries, illness, disabilities, bankruptcies or other factors. The redemption of rights of the less abled members of the society should be operationalised through redistributive risk-sharing social finance instruments such as zakat (alms-giving), waqf (endowment), sadaqah (direct payments to the less abled) and qard hassan (benevolent loan). Advances in the digital landscape can be utilised to develop sophisticated application of these instruments to address the existing challenges of poverty, inequality and unjust distribution of income. Big data, or information generated from digital sources as people go about in their daily lives has the potential to reveal trends, preferences, struggles and overall well-being. Making sense of digital data has led to useful insights into the multifaceted aspects of the human behaviour.

This paper explores how to turn the possibilities of big data analysis into meaningful means of operationalising sustainable and social finance. It assesses the landscape in which big data analysis can be helpful in identifying the Customers of these instruments through addressing the questions of "Who?", "Where?" and exactly "What" kind of financial aid is typically needed. Big data analytics may also be used to decipher the patterns of contributor of funds where Islamic Financial intermediaries could use this information to coordinate and promote sustainable and social finance. The purpose of this study is to paint a picture of the possibilities and challenges in how big data can be used to highlight priority of needs and sectors for sustainable and social finance.

Keywords

Islamic Finance; Social Finance; Big Data; Waqf; Financial Inclusion

1. Introduction

The teachings of Islam postulates that a society will prosper if all individuals are given equal opportunity and resources to attain a decent standard of living. A decent standard of living goes beyond the minimum standard of basic provisions for food, clothing, housing, medical care, and social services, as mentioned in article 25 (1) of the Universal Declaration of Human Rights (United Nations General Assembly, 1948). According to the Icelandic Human Rights Centre (2014), the minimum requirement of an adequate standard of living is the ability for all individuals to be a full participant in ordinary and dignified day-to-day interactions. It goes further to state that the conditions should not allow a person to have to satisfy their needs by degrading themselves or depriving their basic freedom. Notably, Iceland was ranked 2018's happiest country by World Happiness Report and was ranked 4th in 2019 (Helliwell, J., Layard, R., & Sachs, J. 2018, 2019). This is coherent to the Islamic perspective that a society can only thrive if all individuals are given access to resources and equality of opportunity to achieve a decent standard of living. As these individuals become active market participants and consume, produce or trade, the resulting income gained are considered justly earned. In order to achieve this, Islam ordains for a portion of this income of an able individual to be distributed through a "redemption of rights" of those within the population who are unable to actively participate in earning an income due to the circumstances of their lives such as disabilities, injuries, illness, disabilities, bankruptcies or other factors.

The process of redemption by the less abled members of the society should be operationalised through redistributive risk-sharing social finance instruments. Advances in the digital landscape can be utilised to develop sophisticated application of these instruments to address the existing challenges of poverty, inequality and unjust distribution of income. As technology becomes increasingly pervasive, information generated from digital sources as people go about in their daily lives has the potential to reveal trends, preferences, struggles and overall well-being. The amount of digital data being produced is at a rate of 2.5 quantillion bytes of data per day (Jacobson, R. 2013). Between the year 2000 and 2019, the global internet usage grew by 1,104% to an estimated 4.3 billion users, which is 56% of the estimated total global population of 7.7 billion people (Internet World Stats, 2019). This trend will likely continue as more businesses shift their Business Models into the digital space to gain a competitive advantage (Harry, B., Mark, R., and Shahrokh, N. 2017).

The prevalence of interactions and transactions in the digital realm has generated a superfluous amount of data, otherwise known as 'Big Data'. Big data is a term that refers to rapidly generated unstructured data. Big Data is gathered as a biproduct of a business and administration system, social

networks and the internet of things (Cornelia, L.H., Diane, C.K., and Gabriel, Q. 2017). Despite the lack of consensus on the definition of Big Data, its characteristics are commonly associated with the 3V's (Cornelia, L.H., et al. 2017; Devakunchari, R. 2014; Kshetri, N. 2014) which are Volume (Large data size in terabytes and petabytes), Velocity (The rate at which data flows in from sources), and Variety (Structured, semi-structured and unstructured data).

Despite the extensive existing research on Big Data, limited literature observes Big Data's impact on social finance for Islamic Financial Institutions. This paper attempts to deepen the literature by developing a descriptive framework for Islamic Financial Institutions to leverage off Big Data for social finance. Our contribution is to foster a better understanding of the possibilities behind Big Data's for sustainable and social finance. Thus, providing a basis for future studies on the areas where Islamic Financial intermediaries could use Big Data to better coordinate and promote sustainable and social finance.

2. Methodology

This study proposes a method that could be used to utilise financial data to develop a predictive model for cash Waqf beneficiaries. The data recommended for this would be the Islamic financial intermediary's customer information and financial history. The time period observed is suggested to span long periods however for the purpose of this paper the illustrative example will cover a one month period. The findings would be used as a basis for machine learning to develop a predictive model via a decision tree model. The following section sets out a significant part of this study which is qualitative, conceptual and aimed at building a theoretical framework.

3. Analysis

In order to gain insight into the potential characteristics that would be applicable to Big Data analytics, we had to first identify the relevant information. In accordance with the identified area of focus, the span of data was determined to a duration of a month. However, the decision to limit the span of data impacted the available sample size, which is considered in the observation of the results.

The primary data was the biproduct of the core business, which is readily available information captured to extend its services. Due to the continuous update process stemming from its business activities, it allows for a live study of customer behavior. On the spectrum of data types which consists of Unstructured, Multi-Structured, and Structured data, the data available would be mostly derived from structured data.

The data is first collected via forms followed by the information input of data into groups that will be key in assessing demographics. The vast amounts of data are then pooled into a data warehouse where each data set is assigned

a tagging number. This is critical in order to ensure there is no mismatching of information which would influence the validity of analysis moving forward. The data can then be categorized into sub groups to allow for a further detailed quantitative analysis. Hence, the process ensures that the produced data is clean and will provide a valid base of analysis.

4. Theoretical Framework

The approach begins with data filtering, with a machine learning-assisted filtering at the end. From the whole database of a bank's depositors, we propose to filter out all savings account depositors with income lower than a given threshold and with number of account (NOA) less than two. It is at this juncture that we apply a machine learning algorithm to help us identify the list of potential cash waqf beneficiaries. The analysis typically entails observing the movement of daily balances for all customer savings accounts because simple analysis of Average Daily Balances (ADB) or Monthly End Balance (MEB) would not yield the desired result as it would be skewed by the idle and secondary savings accounts.

As shown in Table 1 below, different accounts can have similar month end balance (MEB). However, we propose to only distinguish the potential beneficiaries from the inactive or secondary account by observing at least 1 week movement of their respective daily balance. We propose to further use variation metrics such as 7-days average daily variance, etc. as the validating signal whether a given depositor will be included in the beneficiary list.

Customer ID	**Payday Balance (RM)**	**Payday+ 1**	**Payday+ 2**	**Payday+ 3**	**… Payday + 7**	**Beneficiary Status**
000001	1500	500	300	200	200	Yes
000202	200	200	200	200	200	No (Inactive)
000310	100	200	200	200	200	No (secondary Savings Account)

Table 1. Sample table of customer savings account analysis

While we propose for IFIs to come up with the process to identify the beneficiary status, the tracking of analysing 7 days data for millions of savings account depositors would consume substantial time and storage, rendering it inefficient to do every month. This is where machine learning will step in to replace the repeating analysis.

Rather than analysing every depositors' daily balance, this paper postulated using a supervised machine learning algorithm to learn from past data and develop a predictive model, of which the latter will be used for future data. It

is expected for IFIs to only have to do one-time analysis of daily balance, calculate the validating signal, and let the algorithm learn from that data which customer attributes have significant relationship with the signal. This learned relationship would be the building block of the predictive model.

Name	ADB	MEB	Income	Race	Gender	Minimum Education	Other info…
000001	700	200	2000	Malay	Male	Not disclosed	…
000202	200	200	1200	Indian	Not disclosed	Degree	…
000310	110	200	2300	Chinese	Male	Degree	…

Table 2. Customer attributes

- Based on the attributes illustrated in the above table, IFIs could build a predictive model using the Decision Tree algorithm. We can then pass feed the month-end data with the attributes for a given depositor into the model, and the model will evaluate whether this depositor should be included in the beneficiary list. There are many similar algorithms available but Decision Tree provides a predictive model with visually understandable output, which in turn would allow a comprehensive and real time insights into the demography of the potential social finance beneficiaries. The predictive properties of utilising a decision tree based model has been a popular machine learning method throughout different industries as demonstrated in prior literature such as by Syed. S.H, Ismail. S, and Yap, B.W. (2018) in which the authors used a decision tree to develop a model that successfully predicts personal bankruptcy in Malaysia.

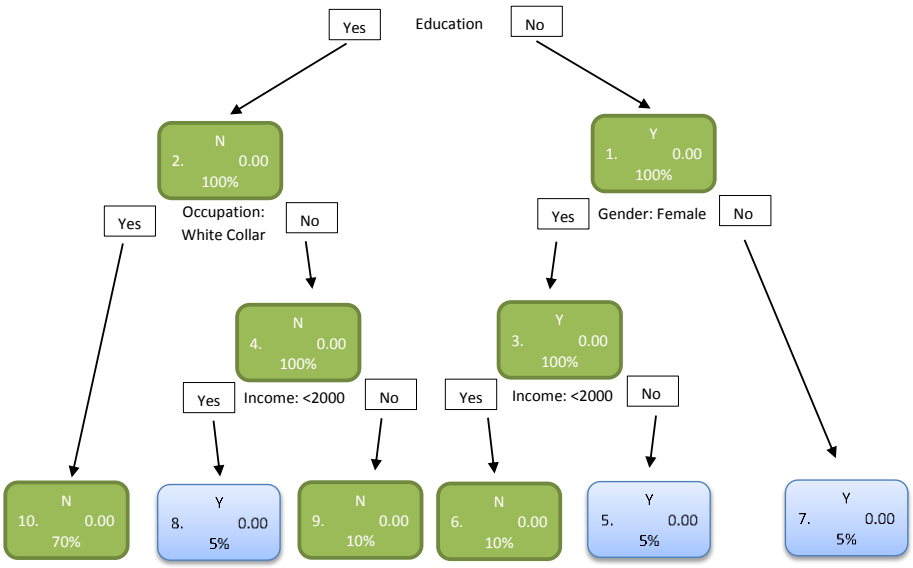


Figure 1. Possible sample of the Decision Tree

The outcome of the analysis is expected to point IFIs to segments that would require additional financial assistance. The cornerstone of Islamic belief is to ensure that "no one is left behind". Through a study of customers' account standing 7 days post end of the month, the analysis aims to theorise that the movement after month end (typically post salary disbursement) would indicate for those in need a reducing balance. The analysis could be repeated for another two 7 days cycle in which accounts that do not recover in balances highlight that the group is indeed the target segment of customers that require cash waqf assistance.

From Figure 1, we have illustrated what the results could look like. In the depicted example, the process begins with the basis of analysis being based on education profile, where the pool of beneficiary could be those that did not complete Tertiary education. From there, the analysis could establish those who are in white collar jobs and then determine the segment based on a limit of below RM 2,000 a month income level. Concurrently, it conducts an analysis of the different genders that make up the segment. Finally, the depicted analysis carries out the process to determine the income level that is associated with observation criteria of interest.

4. Challenges in Implementation

It is worth noting the identified challenges of operationalising this framework. Most pertinently would be the issue of data privacy of customers and the "stereotyping" nature of the algorithm. Currently, consumer behaviour towards sharing personal data is still in its nascence. Enhancement of data security is becoming ever more relevant in order to protect consumer data. Moreover, the computing power required to store and process the data in the initial phase of building the Decision Tree model still poses a challenge on its own. Both of the challenges would incur high cost due to the required technological upgrade and upskilling of staff.

5. Conclusion

This paper built a theoretical framework for Islamic financial intermediaries to leverage off Big Data using a decision tree model to enhance social financing via improved identification of the beneficiaries. This framework postulates that Islamic Financial intermediaries should qualitatively analyse the data available to arrive to a descriptive output. The output can be used to build a model that uses Big Data to allow a deeper insight into the financial behaviour of potential social finance beneficiaries. This allows for better planning of social finance policies and programmes that are able to improve their standard of living. The contribution of this study provides a clear perspective on the potential uses of Big Data usage by Islamic Financial intermediaries. The development of this framework was limited by the

confidentiality of customer information and limited computing power available. A further area of study would be to test the theoretical framework and observing the validity of results generated by the proposed decision tree model.

References

1. Jacobson, R. (2013), *2.5 Quantillion Bytes of Data Created Every Day, How Does CPG & Retail Manage It?* [online]. New York: International Business Machines, Available from: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/ [Accessed 13 March 2019]
2. Internet World Stats (2019), *World Internet Usage and Population Statistics March 2019* [online]. Colombia: Miniwatts Marketing Group, Available from: https://www.internetworldstats.com/stats.htm [Accessed on 13 March 2019]
3. Harry, B., Mark, R., and Shahrokh,, N. (2017) *The impact of Digitalization on Business Models: How IT Artefacts, Social Media, and Big Data Force Firms to Innovate Their Business Model* [online]. Kyoto: International Telecommunications Society (ITS). Available from: https://www.econstor.eu/bitstream/10419/168475/1/Bouwman-Reuver-Nikou.pdf [Accessed on 13 March 2019]
4. Devakunchari, R. (2014) *Analysis on Big Data Over the Years*, International Journal of Scientific and Research Publications, Vol. 4 (1), January 2014
5. Cornelia, L.H., Diane, C.K., and Gabriel, Q. (2017) *Big Data: Potential, Challenges, and Statistical Implications*, Ney York: International Monetary Fund
6. Icelandic Human Rights Centre (2014) *The Right to an Adequate Standard of Living* [online], Reykjavík: Icelandic Human Rights Centre). Available from: http://www.humanrights.is/en/human-rights-education-project/human-rights-concepts-ideas-and-fora/substantive-human-rights/the-right-to-an-adequate-standard-of-living [Accessed on 14 March 2019]
7. Helliwell, J., Layard, R., & Sachs, J. (2018). World Happiness Report 2018, New York: Sustainable Development Solutions Network.
8. Helliwell, J., Layard, R., & Sachs, J. (2019). World Happiness Report 2019, New York: Sustainable Development Solutions Network.
9. United Nations General Assembly (1948) *The Universal Declaration of Human Rights* [online], Paris: United Nations General Assembly. Available from: https://www.un.org/en/universal-declaration-human-rights/ [Accessed on 14 March]

10. Kshetri, N. (2014) *The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns*, Big Data and Society, Vol. 1 (2) [online], Greensboro: Sage Publication. Available from: https://libres.uncg.edu/ir/uncg/f/N_Kshetri_Emerging_2014.pdf [Accessed on 14 March 2019]
11. Syed. S.H, Ismail. S, and Yap, B.W. (2018) Personal bankruptcy prediction using decision tree model, Journal of Economics, Finance, and Administrative Science [online], Bingley: Emerald Publishing Limited. Available from http://doi.org/10.1108/JEFAS-08-2018-0076

Investment Account (IA) in Islamic Banking: Analysis of perceptions, knowledge and acceptance of corporate consumers on IA concept using structural equation modelling

Zuraeda Ibrahim[1], Hamim Syahrum Ahmad Mohktar[2], Zafiruddin Baharum[1], Muhammad Syahmi Mohd[3], Shariza Abdul Ghani[2], Azren Rizuani Aziz[2]

[1]Faculty of Accountancy, Universiti Teknologi MARA, Selangor, Malaysia

[2]Central Bank of Malaysia, Kuala Lumpur, Malaysia

[3]International Shari'ah Research Academy for Islamic Finance, Kuala Lumpur Malaysia

Abstract

The strategic efforts of Bank Negara Malaysia to ensure financial stability of the Islamic banking sector include the process of redefinition and reclassification the fund raising instruments of Islamic banks into deposits and investment. This is to correctly reflect the underlying *Shariah* requirements for each instrument separately. Lack of understanding and unfamiliarity of the underlying *Shariah* concept may deter a majority of Islamic banking consumers from placing funds in the Islamic investment accounts with Islamic banks and this may impact negatively on the Islamic banks as well as the financial stability of the sector in general. Hence, investigating and analyzing the stakeholder's perception, knowledge and acceptance on this new concept of investment account under the Islamic Financial Services Act (IFSA) is of paramount importance to Bank Negara Malaysia. A survey has been conducted; and the questionnaires have been distributed to the corporate consumers of all the Islamic banks in Malaysia, which includes businesses and government institutions. In this study, usable data were collected from a total 141 respondents; however only 44 responses have been further analysed. The respondents of this survey consist of public listed companies (38.6%), non-listed companies (25.0%), government-linked companies (18.2%) and 13.6% are government ministries or agencies (13.6%). The data set were subjected to analysis using structural equation modeling (SEM) based on the PLS approach. Results from the SEM analyses on the corporate consumers demonstrated that only one (1) factor, consumers' acceptance, has a significant relation with understanding of the investment account concept. Consumers with better understanding of the concepts have a higher level of product acceptance. Results also indicate that consumers' perception, knowledge and refusal are not related, and nor do they influence consumers' understanding of the investment account concept.

Keywords

Islamic investment accounts; Corporate Consumers behaviour; Islamic Banks

1. Introduction

An investment account is distinguished from an Islamic deposit in that an investment account is defined by the application of *Shariah* contracts with a non-principal guaranteed feature for the purpose of investment. In other words, when a consumer places money in an Islamic bank through an investment account, the Islamic bank may share the profit generated from the investment, but it is not obligated to compensate the consumer if the investment is not successful. There may be a good chance that a consumer who places money in an investment account may be able to earn more than if money is placed in a deposit account, but there is also no guarantee by the Islamic bank that a consumer will receive all of his investment back. The concept of no risk, no return is embedded in the concept of the investment account. In Malaysia there are two types of profit sharing investment accounts (PSIAs) that have been widely offered by Islamic banking institutions in Malaysia; namely general investment accounts (GIA), and specific investment accounts (SIA). *Mudarabah* is a profit sharing contract that has been widely used to structure the GIAs, SIAs and other profit sharing instruments or products in Islamic banking. Notwithstanding this, the IFSA 2013 provides an adequate legal basis to support the further strengthening of investment account operations in that it provides appropriate protection to investment account holders whilst ensuring financial stability of the Islamic financial system. The priority of payment for investment accounts upon liquidation of the Islamic financial institution is treated separately from Islamic deposits, in accordance with the rights and obligations accrued to the Investment Account Holder (IAH) that are also provided. Islamic banking has emerged as a feasible banking system over the past two decades and this has consequently resulted in growth in the size and numbers of Islamic banks (Metawa & Almossawi, 1998). Since its emergence, a considerable body of literature has examined consumers' perception, awareness and satisfaction towards Islamic banking products, services, practices and operations (Akhbar, Ali Shah & Kalmadi, 2012; Karim, 2012; Unegbu & Onuoha, 2010). To date, no research has examined consumers' perceptions, knowledge and acceptance of Corporate Consumers towards investment accounts in Malaysia, although prior research has identified many factors that may influence satisfaction, behaviour and awareness of products and services of Islamic banks. Therefore, this research examines consumers' perceptions, knowledge and acceptance towards the concept of investment accounts in Islamic banking.

2. Methodology

In order to enable this initiative to reach the target respondents, the study was conducted using both esurvey and hardcopy questionnaires as an instrument to collect data. The UiTM E-Survey system known as Perseus was

used for the e-survey and 1000 copies of questionnaire were distributed to all companies listed on the Main and ACE markets, identified through the Bursa Malaysia website in year 2015. The questionnaire is divided into five main sections. Section A is to identify characteristics and influential factors for acceptance of Islamic banking products, Section B is to indicate the level of awareness of respondents of Investment Accounts; Section C requests the respondents to indicate the extent to which they agree with the concept of Investment Account and their behaviour regarding investment. The respondents are required to specify factors that would influence their acceptance of Investment Accounts based on a 5-point scale, with 5 being 'Strongly Agree' and 1 being 'Strongly Disagree'. Section D requests the respondents to provide a demographic profile of their organisation; and finally, Section E requests demographic details of the respondents. Despite vigorous approaches done on the data collection part; this study could only managed to get responses from 141 respondents. However, only 44 responses have been analysed further because questionnaires with missing values to Likert scale items were removed from the analysis, since structural equation modeling (SEM) analysis is very sensitive to missing values for Likert scale items (Hair et al., 2014). The statistical program that was used was Smart-PLS 2.0 M3 (Ringle et al, 2004). This statistical program validates the psychometric properties of the measurement model and estimates the parameters of the structural model. Before considering the results from the structural model (parameters estimate), the quality of the measurement model was first reviewed. The measurement model (refer Figure 1) was assessed by examining convergent and discriminant validity, to determine the validity and reliability of the measurement items, where all the constructs were reflective type constructs.

3. Results

Table 1 below presents a summary of the results of the convergent validity for the measurement model. Three indicators were not above the recommended threshold of the loading, which is above 0.60, and were also non-significant loading ($t < 1.96$), whereas all items meet the recommended threshold loadings. In addition, the AVE for all constructs are more than the recommended value, which is 0.50, and vary from 0.566 to 0.634, indicating that the latent variables explain more than half of their indicator's variance. Moreover, the CR values for all constructs ranged from 0.714 to 0.947; thereby these values also exceed the recommended threshold of 0.70. These results indicate that the measurement model has demonstrated an adequate reliability of the grouped items. However, in terms of internal consistency reliability (i.e. Cronbach's alpha) test, the ICR values ranged between 0.537 and 0.934, which indicates that the internal consistency for the model was at the

poor level. Therefore, all indicators that fall below the minimum 0.60 loading score should be removed from the analysis, one-by-one.

Table 1: Summary of Results of Convergent Validity (Initial)

Construct	Items / Indicators	Outer Loading	t- statistic	AVE	CR	ICR
Perception	A4.1	.883	3.94**	.634	.873	.799
	A4.2	.682	2.37*			
	A4.3	.803	3.30**			
	A4.4	.803	4.68**			
Knowledge	C11.1	.431	0.98	.591	.714	.538
	C11.2	.998	2.21*			
Acceptance	C8.1	.689	5.66**	.599	.947	.934
	C8.2	.732	5.15**			
	C8.3	.785	5.12**			
	C8.4	.738	5.83**			
	C8.5	.807	8.73**			
	C8.6	.786	7.36**			
	C8.7	.743	5.10**			
	C8.8	.814	8.36**			
	C8.9	.841	8.63**			
	C8.10	.844	6.71**			
	C8.11	.696	4.98**			
	C8.12	.790	9.37**			
Refusal	C9.1	.773	3.65**	.566	.885	.849
	C9.2	.614	1.62			
	C9.3	.672	2.08*			
	C9.4	.830	3.59**			
	C9.5	.891	3.53**			
	C9.6	.698	2.25*			
Islamic Investment Account	B6.1	.584	2.02*	.586	.805	.537
	B6.2	.874	9.72**			
	B6.3	.808	7.19**			

Note: **the indicator loadings were significant at 99% confidence level if t-statistic >2.58 (p <.01); *the indicator loadings were significant at 95% confidence level if t-statistic > 1.96 (p <.05).

The table 2 below shows the final results of convergent validity after removing three items (i.e. C11.1, C9.2, B6.1) one-by-one from the model.

Table 2: Summary of Results of Convergent Validity (Valid)

Construct	Items / Indicators	Outer	t-statistic	AVE	CR	ICR
Perception	A4.1	Loading.886	4.04**	.624	.867	.799
	A4.2	.640	1.97*			
	A4.3	.760	2.91**			
	A4.4	.851	4.39**			
Knowledge	C11.2	One item Measurement				
Acceptance	C8.1	.678	5.66**	.597	.946	.934
	C8.2	.710	5.27**			
	C8.3	.761	5.27**			
	C8.4	.737	6.02**			
	C8.5	.813	9.11**			
	C8.6	.796	7.51**			
	C8.7	.737	5.03**			
	C8.8	.820	8.87**			
	C8.9	.850	8.99**			
	C8.10	.845	6.67**			
	C8.11	.702	4.99**			
	C8.12	.798	9.66**			
Refusal	C9.1	.805	4.26**	.602	.881	.827
	C9.3	.634	2.37**			
	C9.4	.845	5.05**			
	C9.5	.897	5.50**			
	C9.6	.664	2.81**			
Islamic Investment Account	B6.2	.878	11.73**	.815	.898	.755
	B6.3	.927	30.15**			

Note: **the indicator loadings were significant at 99% confidence level if t-statistic >2.58 (p <.01); *the indicator loadings were significant at 95% confidence level if t-statistic > 1.96 (p <.05).

It can be seen that all the items have outer loadings of above 0.60, hence indicating that all the items have a high degree of validity for the respective constructs and also that all indicators are statistically significant at least at the 95% confidence level. In addition, it was found that the AVE for all constructs are more than the recommended value of 0.50, with values varying from 0.597 to 0.815, hence indicating that the latent variables explained more than half of their indicator's variance. Moreover, the CR values for all constructs ranged between 0.867 and 0.946; these values also exceed the recommended threshold of 0.70. These results indicate that the measurement model demonstrates an adequate reliability of the grouped items. Since all the criteria, which are loading, AVE, and composite reliability, meet the recommended threshold of convergent validity, it can be concluded that the measurement model is valid from this aspect. With regard to discriminant validity, it was based on the cross loading method. The following table presents the results for the discriminant validity tests.

Table 3: Summary of Results of Discriminant Validity using Fornell and Larcker's (1981) technique

Construct	(1)	(2)	(3)	(4)	(5)
(1)	**.790**				
(2)	-.221	**1.000***			
(3)	.667	.064	**.773**		
(4)	.263	.086	.378	**.776**	
(5)	.279	-.183	.421	.331	**.903**

Note: (1) = Perception; (2) = Knowledge; (3) = Acceptance; (4) = Refusal; (5) = Islamic Investment Account; The value in the diagonal (bold) is a square root of the AVE of each construct and the element off the diagonal value is the inter correlation value between constructs; *Single item measurement.

Referring to Table 3 above, it is confirmed that the valid measurement model has adequate discriminant validity since all off-diagonal elements are lower than square roots of AVE for each construct. Results from discriminant validity using cross loading technique (not presented here) also demonstrated that all measurement items loaded higher against their respective intended latent variable compared to other variables. Therefore, it can be concluded that the measurement model has established its discriminant validity based on Fornell and Larcker's (1981) criteria and also on the cross loading assessment criteria.

As a conclusion, the reliability and validity of the measurement models were satisfactory and all items in this measurement model were valid and fit to be used to estimate the parameters in the structural model. The next section shows the assessment of the structural model of the measurement model. Next, after computing the path estimates in the structural model, a bootstrap analysis was performed to assess the statistical significance of the path coefficient in the structural model. The results are summarized in Table 4 below.

Table 4: Summary of the Path Coefficient and Hypothesis Testing for Direct Effects

Path	Path Coefficient	t-statistic	Supported
Perception → Islamic Investment Account	-0.131	0.62	No
Knowledge → Islamic Investment Account	-0.259	1.80	No
Acceptance → Islamic Investment Account	0.442	2.45*	Yes
Refusal → Islamic Investment Account	0.221	1.34	No

Note: * The path coefficient is significant at 95% significance level (t > 1.96).

Referring to Table 5 below, the R^2 value of Islamic Investment Account is 0.270 suggesting that 27.0% of the variance in Islamic Investment Account can be explained by Perception, Knowledge, Acceptance and also Refusal, and the relationship can be considered as substantial because the R2 is above 26% (Cohen, 1988). Besides that, the result indicated that only Acceptance variable (β = 0.442, t = 2.45) was found to have a positively significant relationship to investment account. It indicates that, if the level of the acceptance is high, then the possibility to invest in the Islamic account will be high, by controlling other factors. The other factors, such as Perception (β = -0.131, t = 0.62), Knowledge (β = -0.259, t = 0.80), and Refusal (β = 0.221, t = 1.34), were not found to having a statistically significant relationship with investment account.

With regard to predictive relevance (Q^2), Stone-Geisser's (Q^2) is the predominant measure utilized to measure the predictive relevance in order to assess a research model's capability to predict (Hair et al., 2014). Based on a blindfolding procedure, Q^2 evaluates the predictive validity of a model via PLS. Q^2 is generally estimated using an omission distance of 5-10 in PLS (Akter et al., 2011) and Hair et al. (2014) also stated that an omission distance between 5 and 10 should be used in most applications of this technique. Accordingly, Table 6 shows a summary of the predictive relevance for the endogenous construct under consideration in this research. The results show that all exogenous constructs have predictive relevance.

Table 5: Summary of Predictive Relevance ($\mathbf{Q^2}$)

Exogenous Variable	Endogenous Variable	Beta	R^2	Q^2	Predictive
Perception	Islamic	-0.131			
Knowledge	Investment	-0.259	0.270	0.216	Yes
Acceptance	Account	0.442*			
Refusal		0.221			

Note: * The path coefficient is significant at 95% significance level (t > 1.96); Omission distance 7.

To summarize, the measurement model was examined and the results showed that the model could be considered satisfactory from the evidence of adequate reliability, convergent validity and discriminant validity. Following the assessment of the measurement model, the structural model was examined. Table 6 summarizes the results of hypotheses that were tested through the structural model, while Figure 2 shows the results in graphical terms. In attempting to predict the relationships between the sets of variables as hypothesized, the bootstrapping procedure was applied to determine the significance of the relationships. Besides that, the predictive relevance of the model can also be concluded to have a predictive relevance to predict the exogenous variable in this study.

Table 6: Summary of Hypothesis Results

Hypothesis	Description	Results
H1	Perception is significantly correlated to the Investment Account	No
H2	Knowledge is significantly correlated to the Investment Account	No
H3	Acceptance is significantly correlated to the Investment Account	Yes
H4	Refusal is significantly correlated to the Investment Account	No

4. Discussion and Conclusion

The main objective of this study is to examine the relationship between perception, knowledge, acceptance, refusal and the understanding of the IA concept. The results indicate that only one (1) factor, consumers' acceptance, has a significant relation with understanding of the investment account concept. Consumers with better understanding of the concepts have a higher level of product acceptance, and that consumers' perception, knowledge and refusal are not related, and nor do they influence consumers' understanding of the investment account concept. We can conclude that the remarkable progress in the IFI sector, has not been matched by satisfactory knowledge and awareness of the market players. The depositors of Malaysian Islamic banks still do not fully understand the concepts behind Islamic financial products and services, specifically the principles and techniques. Therefore, in order to promote IFI products, especially to influence retail consumers to consider investing in investment accounts with Islamic banks, the banks themselves should take the initiative to create awareness and educate the public through various approaches, for example, seminars and workshops. The concept and expected benefits gained from IFI products and services should be clearly explained. Redefinition and classification of IFI products and services should be announced before any implementation or execution. Bank Negara Malaysia should also continuously monitor the investment activities of all banks so that the Islamic banks would not be disadvantaged by any changes introduced.

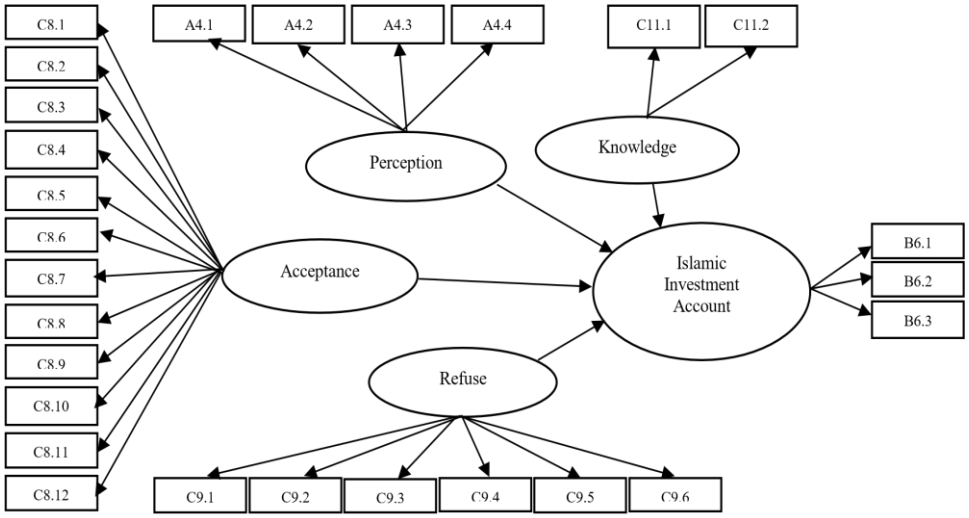


Figure 1: Assessment of the Measurement Model.
Structural Path Model

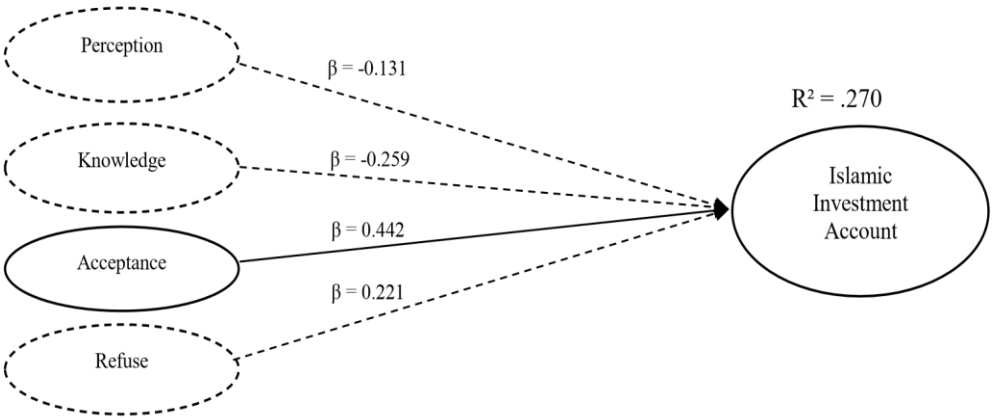


Figure 2: Assessment of the Structural Model.
Non Significant Path; Significant Path

Main Reference

1. Ibrahim, Z., Baharum, Z., Abdul Rasit, Z, Ahmad Mokhtar, H.S., Mohd, M. S., Mohamad, H., Aziz, A.R., Abdul Ghani, S., and Danbatta, B (2017). Investment Account Concept in Islamic Banking: Analysis of Perceptions and Behaviours of Stakeholders, UiTM and Bank Negara Malaysia, Kuala Lumpur

Other References

1. Akbar, S., Ali Shah, S.Z. & Kalmadi, S. (2012). An investigation of user perceptions of Islamic banking practices in the United Kingdom, *International Journal of Islamic and Middle Eastern Finance and Management*, Vol.5 No.4, pp.353-370.

2. Akter, S., D'Ambra, J., and Ray, P. (2011). An evaluation of PLS based complex models: the roles of power analysis, predictive relevance and GOF index. Proceedings of the 17th Americas Conference on
3. Information Systems (AMCIS), Detroit, USA. Retrieved from http://aisel.aisnet.org/amcis2011_submissions/151/
4. Cronbach, L.J. (1971). Test validation. *Educational Measurement, Issues and Practice, 2*, 443-507.
5. Fornell, C., and Larcker, D.F. (1981). Evaluating structural equation models with unobservable and measurement error. *Journal of Marketing Research, 34*(2), 161-188.
6. Hair, J.F, Hult, G.T.M., Ringle, C.M., and Sarstedt, M. (2014). *A Primer On Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Thousand Oaks: SAGE Publications.
7. Karim, F. (2012). Customer Satisfaction and Awareness of Islamic Banking Products and Services in Pakistan, *Interdisciplinary Journal of Contemporary Research In Business,* Vol. 4 No.4, pp.384-401.

Age of personal credit score

Jianhua (Klyment) Huang
WeChat Pay Malaysia Sdn. Bhd. Kuala Lumpur, Malaysia

Abstract

With a prosperous development of e-wallet in China, people are using their phones to buy almost anything. It is in this respect that the competition between e-wallet companies is entering upon a new phase, one that relying on big data and artificial intelligence to generate more values for customers based on the large transaction data. With a study of the credit profile system developed by the central bank, these commercial firms find it hard to apply for their small and medium merchants due to lack of data source and limitation of access. Therefore, they decide to set up a personal credit score based on the transaction big data along with other dimension sources. Moreover, now the personal credit score behind the e-wallet is a proven success in various payment scenarios. In this paper, we are exploring the origins of personal credit score in China and how does the score affect people's daily life both online and offline.

Keywords

China; Mobile Payments; Big Data; Data Modelling; Credit Reporting Model

1. Introduction

In China, everyone pays for everything with their phones. Starbucks, 7-Eleven and even street hawkers. With such tremendous transaction data, e-wallet companies are considering how to generate more value for customers. When they are looking for examples in western countries, these commercial firms find out that a more flexible and dynamic credit score of the individual could be an entry point.

Nowadays, when Chinese citizens have credit deals such as applying for credit cards, mortgage for purchasing realestate, banks will check their credit profile in the credit reference center powered by the People's Bank of China. However, the source of credit history is limited in banks channels; for instance, if a customer has overdue credit card bills, the bank will file a report to the credit reference center. We can tell there is a shortcoming in terms of this credit collection. Firstly the period of collecting credit history is a monthly basis, and there is no down drill of the default reason. Secondly, only banks and a few non-bank institutions are allowed to check customers' credit profile, but many small and medium business still required such information when

they have credit deals with the customer such as car renting, visa applying and so on.

It is in this requirements that e-wallet companies such as Alipay backed by Alibaba, WeChat Pay powered by Tencent, to start considering setup a personal credit score model for these merchants to evaluate customers credit history.

In this article, we are discussing three aspects of personal credit score including sources of data collection, credit score modeling methods, and applications of credit scores.

2. Methodology

The methodology we apply for conducting research is mainly public search and experts interviews. Due to confidential business reasons, some data we present in the illustration is non-disclosed.

3. Result

Before we dive into the personal scores' modeling, we shall have a brief understanding of credit reporting definition and industrial development.

3.1 Definition and Industrial Structure

Credit reporting is the credit profile of individual or enterprise set up by a professional and independent institution which can collect, store and analyze information obtained from individual and enterprise legitimately. The primary purpose of setting up this profile is to serve credit inquiry platform, helping them identify risks. A typical example of credit reporting and credit inquiry platform is the Credit Reference Center powered by The People's Bank of China.

The credit score is a simplified version based on credit reporting. It is a ranking indicating an individual's default rate and potential losses caused by him, which is evaluated by third-party credit inquiry platform based on debtors' capacity and willingness to repay the debt principal and interest.

The industry structure of credit reporting is simple and clear; and three significant parties are getting involved including data supplier, credit inquiry platform and credit score inquirer.

Data supplier mainly includes the bank, e-Commerce site, water, electricity, and gas provider, telecom operators, educational department, hospital and medical departments, public security bureau and so on. It almost covers every functional department tightly related to people's daily life.

After credit inquiry platform collects data from the above suppliers, they will process for data laundry and gives out a personal credit score based on certain calculation models. There are three types of credit inquiry platform including private credit inquiry institution, enterprise credit

inquiry intuition, and Financial rating agencies. Usually, depending on the business, personal credit and enterprise credit inquiry can be provided by the same platform.

Credit score inquirer is mainly from real-estate, car manufactures, peer-to-peer loans platform, a financial institution like banks. A majority request of personal credit score is for evaluating the risk of credit deals such as personal mortgage of buying real-estate and cars, personal micro-loans, enterprise loans and purchasing bonds.

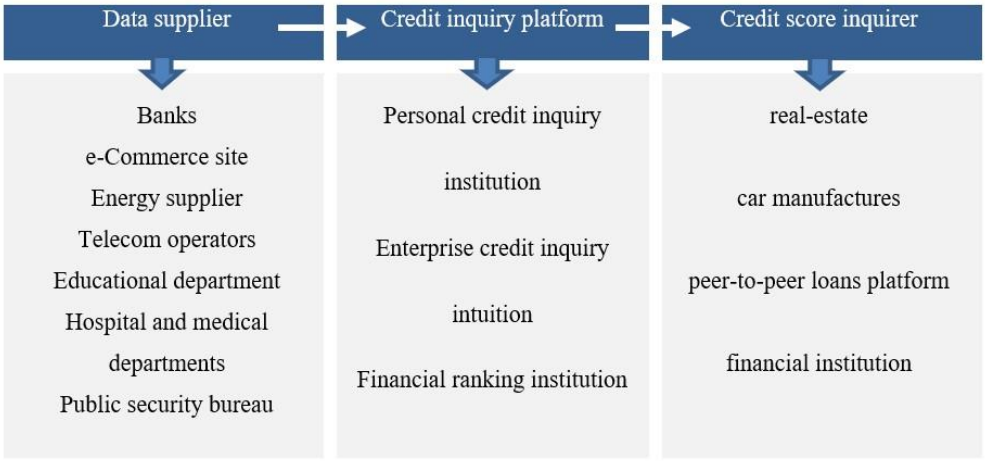


Figure 1: Credit Reporting Industrial Structure

3.2 Credit Reporting Development Status in China

Currently, China's credit reporting system is led by Credit Reference Center powered by The People's Bank of China (CRC), and jointly developed by local and private credit inquiry platform.

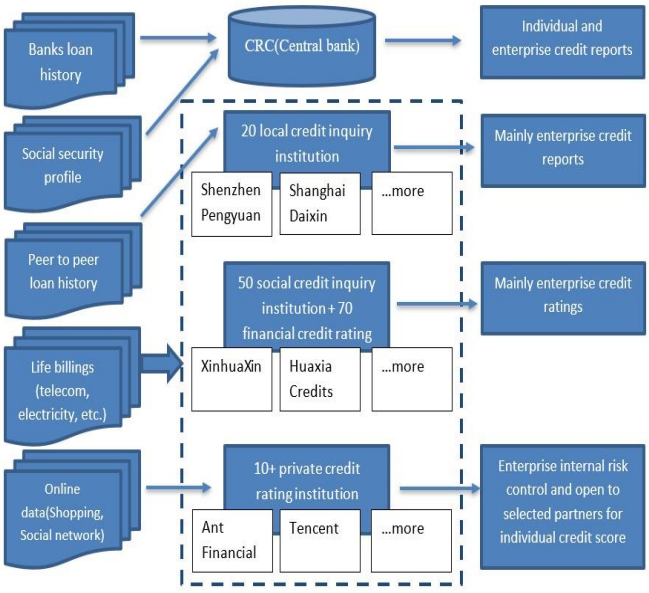


Figure 2: Operation Mode of Credit Inquiry Platforms in China

From the perspective of credit report service targets, the current market structure for large and medium-sized enterprise has been relatively stable. However, in terms of individual and small enterprise credit report service is still underdeveloped. There are two significant challenges for now.

3.2.1 Lack of quantitative ratings

The credit reports provided by CRC mainly includes credit bill overdue records, but there are no quantitative ratings which indicate an individual's credit level intuitively. While most financial institutions use credit rating model from FICO or the three big credit firms (Experian, Equifax, and Transunion), many medium and small-sized credit cooperation, peer-to-peer loan platform could not afford the high cost of technical development. Therefore, their risks management is weak, and reviews of loan applications are less productive.

3.2.2 Interoperable credit sharing is not sufficient among financial institution

Although the P2P Network Financial Information Sharing System (NFCS) have access to 957 institutions, less than half of the institution have reported credit data. The total individual who has a credit report in NFCS is only 4,548,149, which is a small number in a market of 160 million online debtors. Moreover, credit reports contributed by banks, and the big financial institutions are only available to private party or data exchanged partners. Therefore, it is quite difficult for small and medium financial firms and merchants to acquire complete individual credit reports.

3.3 The key technology and social network contribution for Internet credit score

With 54 percent of e-wallets penetration rate in China, dominated by WeChat Pay and Alipay, the companies behind these two wallets, Ant-financial and Tencent starts to study how to establish a credit score for its users, and allow its' ecosystem partners to have better risk control when offering credit trades. Typically, credit score modeling consists of four steps.

3.3.1 Data preparation

Build a database for data recording, cleaning, and filtering. The source of data comes from various dimensions including

- Personal identity, such as education, driving license, employer information
- Credit history, such as credit card repayment history, house mortgage
- Contact network, an individual's friends credit can affect his own as well
- Transaction behavior, such as shopping, borrowing devices, peer-to-peer money transfer

Before further processing the above data, necessary computation will be done such as normalization and noise rejection.

3.3.2 Metrics selection

For an individual credit score, different companies have different weights on the metrics of the rating model. Typically, credit repaying history will contribute the most, followed by transaction records in terms of volume and frequency.

3.3.3 Univariate analysis

Through univariate analysis, we can find out the projection maps between univariate and default rate. For example, a continuous overdue of credit card bills may be more likely to indicate the individual is a defaulter compared to his late confirmation of order upon received.

3.3.4 Model fitting and parameter estimation

For example, a logistic regression model using forward feature selection. The parameter estimation of the model usually adopts the optimization method such as the least squares method, maximum likelihood estimation, and maximum posterior probability.

Also, for a commercial enterprise like Tencent, the social network data is a featured source which they can apply in credit score modeling in two ways, feature development, and credit score correction.

Feature development consists of two essential metrics, self-network structure and friends network structure. Selfnetwork structure mainly refers to the user's network topology, which can be described using some features. For example, the user's reading behavior can reveal his latest interest which can be taken into consideration when issuing micro-loan.

The friend's network structure is a look-like analysis model which may reflect a user's potential behavior. For instance, a friend's packet money interaction and group package money interaction will conclude some relative effects based on their demographics.

3.4 Applications of Personal Credit Score

Generally, there are several mainstream services which personal credit score is applied including shopping, renting, accommodation, recycling, transportation, telecommunication, and micro-finance. In this paper, we mainly discuss the application of renting and micro-finance.

Since Chinese mobile phones users spend much time on the handset such as watching videos, playing games, a typical scenario is running out of battery when they hang out with friends. In the past, they may need to carry a power bank along which makes it inconvenient but now thanks to companies like JieDian; there are many power bank renting shops in shopping malls, restaurant, and cinemas. The process of borrowing a power bank is quite straightforward. Firstly, use an app like WeChat to scan the QR code on the power bank machine, then agree to the terms and conditions, pay the deposits

which they can withdraw later, and the power bank will pump out of the machine. Many people may not be willing to pay the deposit since they are afraid that they cannot get back the money. Similarly, Jiedian also has concerns that consumers may not return the power bank and lost it forever. That is how personal credit score plays its role. With users' authorization to check his credit score, Jiedian will waive the deposit if the score is high enough. Moreover, subsequently, credit score platform will check whether the customer returns the power bank and pay the renting fees. If not, the customer's credit score will be adjusted accordingly.

Figure 3: Borrowing a Power Bank and Authorize Merchants to Check Credit Score for Waiving Deposit

As for microfinance, a typical example is loans. Let us see an example of Weilefen, a product which cus customers can lend money to pay credit card bills with low interest. Weilefen is built within WeChat's ecosystem, through Official Account, a mini-portal that merchants interact with its followers, an individual can check his loan limit. For the first time users, he needs to authorize Weilefen to check his credit profile in Credit Reference Center powered by The People's Bank of China by providing IC number and full name. Then in the backend, Weilefen will calculate the loan limit based on his credit profile and the credit score model designed by Tencent for internal business. After the user gets the limit, he may proceed to pay the credit card bills by several installments right inside Official Account. By default, Weilefen will read the card information bound in WeChat Pay, but users can add another credit card as well. The loan will be released within 30 minutes after users' confirmation.

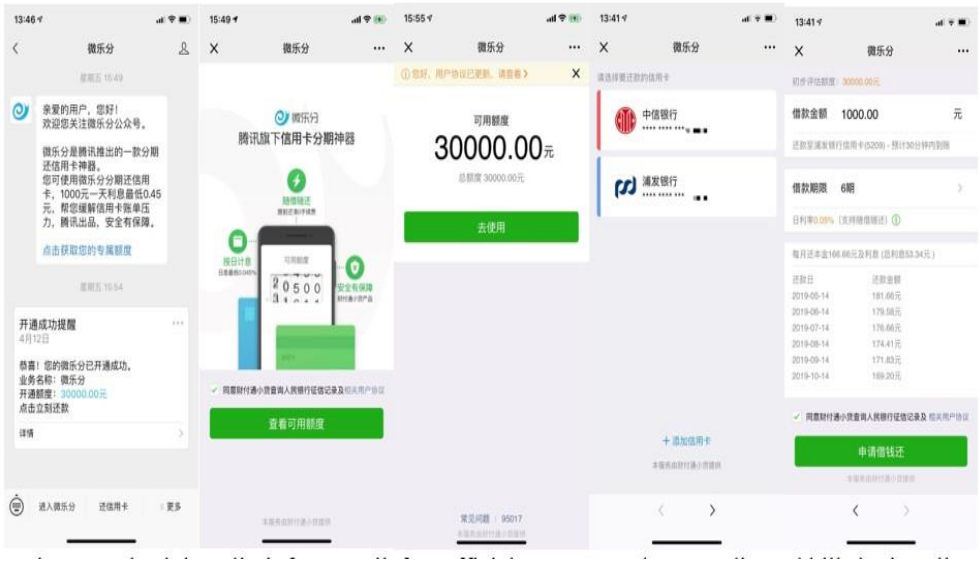


Figure 4: Check Loan Limit from Weilefen Official Account and Pay Credit Card Bills by Installments

4. Discussion and Conclusion

As personal credit score is getting mature, the usage scenarios are getting richer, hence providing much more convenience to customers since it not only offers an intuitive way for individuals to understand his credit level better but also a secure bridge for merchants to estimate default risk when offering credit deals. Nevertheless, the voice of privacy concerns is getting louder and louder as people realize more and more data is being collected.

In order to strike a balance between credit score applications and personal privacy protection, it is essential to have clear and comprehensive policy both from regulators and commercial companies. Firstly, regulators should improve personal information and privacy legislation protection system by integrating existing laws and regulations on personal information protection, privacy protection, credit information regulation, etc., and accelerate the establishment of special regulations such as the Personal Information Protection Regulations and the Personal Credit Information Management Regulations to protect Internet personal information and privacy. In the era of big data, the protection of privacy should pay more attention to the actual control of personal information by the information subject, fully respect the "informed consent" of the information subject and the right of correction, dissent, and deletion, and improve administrative responsibility and criminal responsibility. Based on this, a scientific and reasonable civil compensation mechanism is constructed to make the protection of privacy more comprehensive and efficient.

On the other hand, for commercial companies which provide credit scores, they should strengthen self-discipline and take the initiative to protect

personal privacy and safeguard the authority and social influence of themselves and the industry. Besides, commercial companies should establish corporate self-regulatory organizations, issue initiatives or conventions to form corporate protection personal privacy alliances, promote privacy protection certification, and play the role of industry organizations. At the same time, the company should sort out the risk points within the company that may reveal privacy, strengthen internal control and improve technical protection tools, and update the concept of technological protection.

"All data is credit" is becoming a reality. While data is quietly changing people's online and offline life, it is also reshaping the basic rules of credit reporting. While serving the economic and social development, personal credits scores need to protect and respect personal privacy and protect the freedom of individuals to free their data from the collection and analyze.

References

1. Jiedian Corporated with WePay Credit Score. Web material.In https://www.cnbeta.com/articles/tech/810043.htm. 2019
2. The Development Report of Global Crediting in Big Data Era. Web material.In https://cloud.tencent.com/developer/article/1102964. 2018
3. Research Report of Crediting Industry. Web material.In http://www.sohu.com/a/139027123_720186. 2017.
4. Zhibin Liu.The Industry Structure and Key Technology of Internet Crediting. Web material.In http://www.tisi.org/4254_49. 2017.

Innovative approaches in measuring financial inclusion - Linking survey and administrative data

Damola Owalade
Insight2Impact facility[1]

Abstract

Demand-side surveys are particularly popular in developing and emerging markets, where financial inclusion data is not readily available on a large portion of the adult population. This information plays a critical role in developing financial inclusion strategies, which aims to bring financially excluded adults into the formal financial system. This paper documents the data quality implications of linking administrative data with survey data to explain the use of digital financial services in Nigeria and credit behaviour in Zimbabwe. Merging demand side surveys with administrative data provides an opportunity to utilise a customer-centric approach in assessing the impact of financial inclusion. This methodology has challenges which include access to administrative data, the quality of the administrative data and the ability to recruit respondents to participate in survey that allows the construction of additional information such as demographics and behavioural attributes of customers.

Keywords

Financial Inclusion; Linked data; Financial Inclusion Measurement; Customer-centric

1. Introduction

The emphasis on using data for decision-making in financial inclusion (FI) has resulted in rapid growth in the number of demand-side surveys implemented globally (For example, FinScope, FinDex, FinAccess). Demand-side surveys are particularly popular in developing and emerging markets, where financial inclusion data is not currently available on a large portion of the adult population. This information plays a critical role in informing and measuring financial inclusion strategies, which aims to bring financially excluded adults into the formal financial system.

Financial inclusion is typically measured along the domains of access (access to financial touch points), uptake (take up of a financial product), usage (usage of products) and quality of services (value proposition to customers). Access is adequately captured when there is high quality location or spatial data on financial access points (e.g. agent location) which can be overlaid with

[1] Launched in 2015, i2i is jointly hosted by Cenfri and FinMark Trust and is funded by the Bill & Melinda Gates Foundation in partnership with The MasterCard Foundation.

population data with GPS coordinates. Uptake data can be measured when there is a unique identifier for customers using financial services (from regulated entities that report to a regulator like a central bank) which makes it possible to aggregate the number of unique customers registered with banks. Uptake can also be assessed through demand side surveys that are nationally representative where respondents are asked whether they are registered with formal financial institutions.

Moreover, usage data is usually based on the level of activity or recency of using of a financial product within a set period – usually – a time period of 3 months. This can either be determined through a demand side survey or administrative data. Using administrative data to quantify usage data should be more accurate since actual usage can be observed as opposed to a demand side survey where data is susceptible to non-sample errors such as recall error by the survey respondent.

Finally, quality of usage can be assessed using administrative data on complaints and customer feedback. It can also be assessed using a demand side survey module with questions designed to understand the customer value proposition of using financial services. This paper will be focusing on the usage measurement of financial inclusion highlighting data quality considerations of linking demand side and administrative data to understand financial inclusion specifically on credit and borrowing behaviour and the use of digital financial services in Zimbabwe and Nigeria.

Administrative data takes the form of government or financial service provider (FSP) records that one can link with survey data to form an integrated dataset. Examples of administrative data include payment transaction history, tax, health or employment records. This data is often more reliable as it is regularly collected and can reduce survey interview burden on the respondent because they do not have to report sensitive data (e.g. income) or remember historical health or employment events. (Sakshaug & Kreuter, 2012).

The linking of administrative data with demand side data is not a common practice in financial inclusion research due to lack of feasibility in developing countries when considering lack of data sharing culture, rigid implementation (or lack) of customer data privacy laws, and the unreliability in compiling information on unique customers. Therefore, this paper will be focusing on the usage measurement of financial inclusion highlighting data quality considerations of linking demand side and administrative data to understand financial inclusion specifically on credit and borrowing behaviour and the use of digital financial services in Zimbabwe and Nigeria respectively. These projects in Zimbabwe and Nigeria were implemented as part of the insight2impact's mandate in contributing to the discourse of financial inclusion measurement. The i2i financial inclusion measurement conceptual framework posits that financial needs (e.g. resilience needs) through

deconstructed use cases (having experienced an unexpected shock such as loss of employment or flooding of a farm) drives the usage of financial services. Accessing customers' actual history while administering a survey offers additional demographic and attitudinal dimensions – allowing for a more nuanced approach to understanding the usage of financial services be it credit (in Zimbabwe) or digital payment platforms (in Nigeria).

The next section provides some country context, the research design including sample design, and data collection methodology.

2. Methodology

As mentioned, this paper will focus on quality assessment of merging administrative and survey data. The following will be the key quality considerations.

- Quality assessment of supply side data[2] along the quality dimensions of relevance, accessibility, interpretability, coherence, accuracy and institutional environment
- Non-sampling errors based on the quality outcomes of securing interviews to merge administrative records with survey data.

Zimbabwe

Contextually, the Reserve Bank of Zimbabwe launched a National Financial Inclusion Strategy (NFIS) for the period 2016-2020 on 11 March 2016 to facilitate an inclusive, shared and broad-based economic growth. The NFIS aims to increase overall level of access to formal financial services from 69% in 2014 to 90% by 2020[3]. Credit plays a role in the financial lives of Zimbabweans creating opportunities for microfinance banks and microfinance institutions in supporting individuals and businesses mostly earning a living in the informal economy.

A description of the Zimbabwean study is as follows:

Research objective: Understanding of the borrowing behaviour, repayment behaviour and to assess potential financial distress or over-indebtedness.

Sample (demand side): Randomly selected adults in Harare, Bulawayo, Manicaland, Mashonaland-West and Midlands (n=700).

Sample (administrative): Customer loan repayment data from a commercial credit bureau from June 2012 to 2017 (n=750).

Sample (merged demand and credit bureau data): Completed interviews for customers sampled from the credit bureau (n=307).

[2] Data Quality Assessment Tool for Administrative Data, Iwig W, Berning M, Marck P, Prell M, February 2013 available at https://www.bls.gov/osmr/datatool.pdf

[3] https://www.rbz.co.zw/index.php/financial-stability/financial-inclusion/financial-inclusion-strategy

Data Collection methodology: Computer Assisted Personal Interviews (CAPI).

Nigeria

In Nigeria, the Central Bank of Nigeria launched its National Financial Inclusion Strategy in 2012 with a target of reducing financial exclusion to 20% of the adult population by 2020[4]. Digitising payments is a cornerstone of Nigeria's financial inclusion strategy, an objective that is hampered by low banking penetration and very low take-up of mobile money.

The next section provides the findings of the quality assessment of merging administrative and survey data to understand usage of digital financial services in Nigeria.

A description of the Nigeria study is as follows:

Research objective: To explore Nigeria Interbank Settlement System (NIBSS) data in combination with demand-side data and, where possible, create indicators of financial inclusion in Nigeria; Assess the potential value of transactional (administrative) data in supporting efforts to monitor progress on financial inclusion targets in Nigeria; and test the feasibility of conducting matched transactional and demand-side research more broadly and understand how it could be optimised in the future.

Sample (demand side): Randomly selected adults in Lagos and Kano (n=2,395).

Sample (administrative): 1 million records of interbank transaction records based unique Bank Verification Numbers (BVNs) from June 2016 to December 2017.

Sample (merged demand and credit bureau data): Completed interviews for customers (BVNs) sampled from NIBSS (n=611).

Data Collection methodology: Computer Assisted Personal Interviews (CAPI).

3. Results

This section is broken into two parts with each focusing on the quality assessment of the research design in Zimbabwe and Nigeria. The analysis was based on the key findings, technical reports (documenting lessons learned) and first-hand experience of the author who designed the research methodology and occupied a project management role in the two case studies.

[4] https://www.cbn.gov.ng/Out/2018/CCD/Exposure%20Draft%20of%20the%20National%20Financial%20Inclusion%20Strategy%20Refresh_July%206%202018.pdf

The data quality assessment[5] from the two case studies based on the methodology has discussed in section 2 are presented below.

Zimbabwe[6]

The data[7] was on loans from microfinance entities, department stores, other retailers and agricultural suppliers.

- Relevance

The data is relevant as a sample source, but it is limited to one of the four credit bureaus in Zimbabwe including the Reserve Bank of Zimbabwe's Credit Reference Bureau (CRB). Notably, the credit records do not include retail credit from commercial banks and mobile money operators. There is a dearth of publicly available data on the credit market in Zimbabwe but according to Zimbabwe FinScope Consumer Survey 2014, commercial bank credit was accessed by only 4% of the adult population compared to the non-bank credit which was accessed by 10% of the adult population. If it is assumed that the ratio of bank to non-bank credit, based on Zimbabwe FinScope 2014, still holds, then the relevance of the data will be limited to only a segment of the formal credit market. However, the data is adequate to inform policy design for credit users from microfinance banks and institutions.

- Accessibility

The data was provided to the i2i team in a spreadsheet. There was no engagement with the management information system of the credit bureau. A list of selected potential survey respondents were provided with names and contact details. Administrative data was analysed at the office of the credit bureau in Harare in adherence to data privacy and sharing protocols.

- Interpretability

The research team depended on the contact person from the credit bureau to explain the data as there was no data dictionary. In ensuring high quality data, it is required that interpretation of data is informed by institutional based definitions of the data points as outlined in a data dictionary.

- Coherence

The data is stable and there were no changes to the data template over the period of study. Given data privacy concerns and time availability, the data could not be matched with actual records from the credit providers.

[5] Data Quality Assessment Tool for Administrative Data, Iwig W, Berning M, Marck P, Prell M, February 2013 available at https://www.bls.gov/osmr/datatool.pdf

[6] The research project was conducted between July and October 2017. Fieldwork was conducted by Research Continental-Fonkon (RCF)

[7] The indicators provided by the credit bureau include Credit application, Accepted and rejected applications (reasons for rejection), Loan value, Credit provider type (Microfinance banks and institutions, retail stores, and agriculture input suppliers), Use case of credit (e.g. consumption or productive categories) and Repayment defaults

- Accuracy

In some cases, address and contact details are not available for customers. There is no data on gender and age categories limiting the ability to generate disaggregated analyses. There could be room for improvement in the data templates (for onboarding customers) used by microfinance banks and microfinance institutions, and the data sharing arrangement between the credit providers and the credit bureau.

- Non-sampling errors
  - Securing interviews to merge admin with survey data

Out of the 700 (which was subsequently increased) customer details provider, only 207 participated in the survey. This meant that intended sample cluster at a province level was not tenable. To improve the response rates would have required involving all the credit providers (which the credit bureau had data for) to reach out to their customers, informing them of the research. In this case, the credit providers did not participate in the project. A call centre was initially set up by the credit bureau to secure interviews which led to low levels of confirmations. However, the research house was more successful in securing interviews with respondents.

Other challenges in recruiting respondents with administrative data included unavailability of respondents (due to cold calls) while others were not willing to participate in the study. This resulted in high substitution rates.

Nigeria[8]

The insight2impact collaborated with the Nigeria Interbank Settlement System (NIBSS) to analyse data[9] generated through card and mobile phone-based transactions for interbank transaction. The aim of the collaboration is to understand and provide some stylised facts on user characteristics of payments made using digital financial platforms like banking apps and USSD[10]. The types of transactions include those conducted using include mobile phone, internet banking, use of cards through point-of-sale devices, use of bank branches and merchant payments. The data pulled considered transaction histories over a period of 18 months (June 2016 to December 2018), depending on the platforms under consideration and the analysis being conducted. Aside from transactions data, the data included some demographic data for each BVN, including age, gender and location.

[8] The research project was conducted between July and December 2018. The project team consisted of the i2i, 71point4 and AC Nielsen Nigeria that conducted the fieldwork

[9] Data provided by NIBSS included NIBSS Instant Payment (NIP), Point of Sale transactions (POS), Cheque transactions, NIBSS Electronic Funds Transfer (NEFT), CMMS transactions and mCASH

- Relevance

Given the focus on interbank transactions, the data does not include intra-bank or 'on-us' transactions. To achieve the required coverage, intra-bank data is required and there was no definitive information on the market split between interbank and intrabank transactions in the Nigerian payments system. Furthermore, only POS transactions provided information on merchant payments that are intrabank in nature. The consequence is that the data is not representative of digital payments in its entirety in Nigeria.

- Accessibility

The data provided was masked specifically for BVNs and bank account numbers of the sample provided. The research team was unable to access the data remotely so the extraction of data was conducted on-site.

- Interpretability

A data schema was provided which depicted how the indicators are organised in the NIBSS management information system. The team relied on NIBSS staff members to provide accurate description of the data due to the lack of a data dictionary.

- Coherence

The data was generally stable for the time period covered. One of the digital platforms (mCASH) was only introduced in 2016 and had limited usage levels. Therefore, it was not considered during the analysis.

- Accuracy

The data was directly pulled from an operational environment therefore it is likely to be accurate. Nevertheless, the quality of the data is only as good as the data provided by customers during the on-boarding process to use digital services such as when they open an account. Location information proved to be inaccurate in some instances as observed during the recruitment of respondents for the survey. Some of the customers on the list provided claimed that they no longer reside in the location they provided when they registered for the BVN. In addition, merchant location information (when receiving POS devices) might not match location of the actual terminal as POS devices are mobile and in Nigeria, they are currently not tracked. This makes analytics based on location prone to errors.

- Non-sampling errors

- Securing interviews to merge admin with survey data

Of the initial list of 4,710, a total of 611 respondents were recruited for the research, the majority of whom were from Kano (448). Similar to the case in Zimbabwe, sampling from the NIBSS database resulted in a low response rate (13% response rate). This is expected especially in a commercial city such as Lagos where people are guarded to avert the possibility of fraud.

4. Discussion

Measuring financial inclusion has been focused on a product-based lens with M&E framework of FI strategies mainly measuring the level of penetration of savings, credit, and insurance products. However, using financial products is a means to an end of paying for use cases such as getting an education or starting a business where value to livelihood is derived. Therefore, a customer centric approach to measuring financial inclusion is crucial to understanding how financial services can customer needs to fulfil non-financial outcomes such as improving living standards.

A customer centric approach to measuring financial inclusion requires multiple data sets to measure the four measurement domains based on ensemble and time-based dimensions. Demand side surveys have been the gold standard as one can assess the financial life of a unique customer across product types while drawing on behavioural attributes to explain the levels of access and usage. Demand side surveys are adequate in providing cross sectional data but they are usually expensive, time consuming and lack the dynamism to observe unique customer segments across time.

The merging of demand side surveys with administrative data makes robust FI data management possible to assess the outcomes of using financial services. The linkage of administrative data with survey data is a potential best practice in providing quality data to inform policy interventions to expand financial inclusion in order to reap the potential positive outcomes on micro-economic and macro-economic indicators.

The findings from the two case studies discussed in this paper show that improving the quality of administrative data on financial services can potentially reduce the cost and time taken to assess customer experience in using financial services. As discussed, there are challenges in collecting additional information from customers observed on administrative data platforms. Financial inclusion stakeholders such as the regulators and financial service providers are integral to the goal of improving the quality of admin data. This calls for a policy direction that looks to develop a financial inclusion data ecosystem with the requisite data sharing protocols hinging on sound consumer data privacy laws.

5. Conclusion

Financial inclusion is one of the developmental pursuits that can affect economic growth and the improve livelihoods. There has been a global emphasis on interventions that directly impact on development goals, starting with the Sustainable Development Goals (SDGs)[11] which prioritise certain

[11] https://www.uncdf.org/financial-inclusion-and-the-sdgs

objectives, with financial inclusion being a factor in meeting some of those objectives.

Measuring FI is an on-going process of experimentation and this paper aimed to show that there is credence to the advocacy of FI data best practices in developing a robust data management framework to monitor and evaluate country level FI policies.

This paper provides the findings from two case studies which involve experimentation with research design specifically on merging administrative data with demand side surveys. The methodology in question is not error free but given the concerted effort by regulators and FSPs in ensuring high quality data management, there is a possibility that merging administrative and survey data can become a viable model in explaining financial inclusion and potentially shedding some light on its effects on economic growth and livelihoods.

References

1. Insight2Impact facility (2019). Advancing financial inclusion ı Executive summary: Nigeria pilot study. Availabe online at https://i2ifacility.org/insights/publications/advancing-financial-inclusion-executive-summary-nigeria-pilot-study?entity=blog
2. Chamboko, R & Makuvaza, L. (2018). A needs-based approach to financial inclusion measurement in Zimbabwe. Available online from https://i2ifacility.org/system/documents/files/000/000/066/original/A_needs-based_approach_to_financial_inclusion_measurement_in_Zimbabwe_i2i_June_2018.pdf?1530184081
3. Owolade, D. (2016). Revisiting the building blocks: Getting the basics of financial inclusion demand side data right. Available online from http://access.i2ifacility.org/Publications/DQ_Innovation_Focus_Note.pdf
4. Sakshaug, J. W. & Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. Survey Research Methods, 6(2), 113-122.

Examining customer journeys at financial institutions in Cambodia

Robin Gravesteijn[1], Mayank Kumar Jain[1], Jonggun Lee[2]
[1]United Nations Capital Development Fund
[2]United Nations Pulse Lab Jakarta

Abstract

National financial inclusion surveys such as Finscope, Intermedia and Findex provide significant insights into people's access to finance, but the understanding of customer's long-term financial service use remains limited. This study uses readily available big data from four Financial Service Providers (FSPs) in Cambodia—covering approximately 21 percent of the loans and savings market—to examine how long customers stay with their financial institutions and what types of products they take up during their journey. Conducting 'survival analysis' and applying a gender lens, the study finds that although men and women have equal access to credit and saving services, the actual amounts of loans and savings mobilized are much lower for women than men, despite women generally staying longer in borrowing relationships for individual loans than men. Nearly 70 percent of customers had low-value or passive savings accounts with deposit balances below US$5 and women were more likely to have passive accounts (75 percent) than with men (59 percent). Savings mobilization remains a challenge in Cambodia, particularly outside Phnom Penh and for older people, as these depositors are more likely to have passive accounts. The majority of borrowers (78 percent) exit the FSPs within the first three years, implying there is a limited long-term borrowing relationship across the board. The study estimates that reducing passive savings accounts and borrower exit by 10 percent could add an additional $52 million to the deposit portfolio (10 to 33 percent for 2015 portfolio levels) and $304 million to the loan portfolio (24 percent) of the four FSPs as well as reduce operating expenses by $54 million. The paper offers business and policy recommendations for improving customer retention through better product development and recommends incorporating savings mobilization for women and youth into the National Financial Inclusion Strategy. [1]

Keywords

Financial Inclusion; Survival Analysis; Savings Mobilisation; Cambodia; Big Data

[1] The current short paper is an excerpt from the wider joint working paper of UNCDF and UN Pulse lab Jakarta (2018), drafted by Gravesteijn, R., M. K. Jain, and J. Lee (2018, July). Examining Customer Journeys at Financial Institutions in Cambodia. Using Big Data to Advance Women's Financial Inclusion UNCDF UN Pulse Lab Jakarta pp1-40.

1. Introduction

Over the course of the last decade, Cambodia has experienced rapid poverty reduction and economic growth. This has been accompanied by an accelerated rate of financial inclusion. Nearly 59 percent of adults in Cambodia have access to formal finance (FinScope, 2015) which is dominated by the domestic payment markets (to which 37 percent have access) and credit markets (to which 28 percent have access). Like most developing ASEAN nations, young, rural and low-income Cambodians have significantly less access to formal finance than the overall adult population. Interestingly, Cambodian women have slightly greater access to formal financial services (2 percent) than men because they receive more remittances, but they are underserved in terms of credit and savings.

Access to formal financial services does not necessarily imply efficient and active use of such services for personal consumption, investment, or business activities. As many as 22 percent of Cambodian adults have inactive saving accounts, which have had no deposits or withdrawals in the past one year (Findex, 2017), and the borrower exit rates from Cambodian Microfinance deposit institutions (MDIs) vary between 28 percent and 39 percent per annum (MIX Market, 2015). When financial services are appropriately designed and used effectively by the under-served population, they can contribute more strongly to sustainable development goals such as poverty reduction (SDG1), women's economic empowerment (SDG5), inclusive economic growth and decent work (SDG8) (see e.g. ILO, 2015; Buvinic and O Donnel, 2016; Banerjee, Karlan et al., 2015). This paper investigates for how long customers use a variety of financial products and services, how efficiently they use them, and whether those usage patterns differ by gender, age, and location (rural/urban).

2. Methodology

This paper examines customers' loan and savings mobilization patterns by using readily available management information system (MIS) data from four leading Cambodian FSPs: AMK; AMRET; Sathapana Bank; and WB Finance (formerly Vision Fund). The consolidated dataset contains around 5.4 million loan and savings records for 2.3 million customers, which represents almost a quarter of Cambodia's total adult population in 2015. The study covers around 60 percent of the depositors and 53 percent of the borrowers from the Cambodian microfinance sector, and around 22 percent of depositors and 38 percent of borrowers from the Cambodian banking sector. The dataset contains information on customers' gender, age, province, and their savings and loans products, allowing us to study long-term financial service usage for different groups of customers during the period 2010-2015. The big data study applies descriptive and survival analysis to measure the customer journey, looking at actual financial service usage patterns by sex, age and other

demographic indicators, from the point people enter the financial institutions to the point of their exit. Insights were further triangulated with Cambodia's National Financial Inclusion Survey (FinScope 2015) (n=3150) as well as feedback from over 80 FSP experts to identify why certain underlying patterns occurred.

3. Results

The study finds significant gender and youth gaps in average loans and savings mobilization in Cambodia. While Cambodian men and women have almost equal access to formal financial services, the loan and savings mobilization is higher among men (the gender gap in loan amounts is $825 and in savings is $658). This gender gap may be explained by existing gender inequalities in wages and incomes, access to assets and employment activities in Cambodia (see e.g. ADB, 2013) and the fact that women have a preference for informal savings schemes (FinScope, 2015) and group loans. Cambodian youth (defined as those between 18-25 for the purposes of this study) have 20 percent lower access to formal finance than older adults, and they also save and borrow less (the youth gap in loans is $567 and in savings is $215). Despite stronger customer loyalty, youth and women received smaller individual loans due to lack of credit history and collateral and a perceived lack of business skills and experience.

Nearly 70 percent of customers had low-value or "passive savings accounts"[2] with deposit balances below $5. Results of survival analysis[3] show that the level of a customer's savings is most likely to fall below $5 within the first year of the account opening (see figure 1a). Descriptive analysis showed that the percentage of passive accounts was higher among depositors with credit-linked savings (72 percent) than with voluntary accounts (45 percent), indicating that linking credit and savings accounts did not advance savings mobilization. Women (75 percent) and older adults (72 percent) have a significantly higher proportion of passive accounts than men (59 percent) and youth (47 per cent) respectively. Figure 2 below demonstrates that female depositors have a higher proportion of passive accounts than men across most provinces of Cambodia, as shown by the larger red slices on the lef-hand fgure.

Men mobilize savings better than women in most provinces (as shown by the larger yellow slices on the right-hand fgure). Similarly, Figure 3

[2] To analyse savings mobilisation of customers, we define accounts with savings below 5US$ as passive accounts. This threshold is a reasonable proxy for savings account dormancy as over 80 per cent of the dormant savings accounts had savings balances less than $5.

[3] Originating from medical science, survival analysis has recently been widely applied in the field of economics, finance and engineering, amongst others. Survival analysis for example is used to measure the survival-time (or death) of cancer patients after diagnosis or to measure the 'period of unemployment' after providing a training to a group of unemployed individuals.

demonstrates that older adults have a signifcantly higher proportion of passive accounts than youth across most provinces of Cambodia.

Among the key reasons for the large proportion of passive savings account are: customers' limited awareness of savings opportunities; low financial literacy; limited access points in rural areas; and the attractiveness and convenience of informal savings.

For example, nearly 60 percent of adults are unaware of formal saving methods, 54 percent cannot reach a bank within 30 minutes, and 33 percent of the adult population saves informally. Moreover only 10 percent of the people save to engage in business or farming activities, indicating that these activities deliver higher returns than savings in the bank (UNCDF Finscope 2015).

On the supply side, credit-linked savings accounts are often opened merely to function as repayment vehicles for loans. In addition, staff and customer incentive systems are geared towards improving savings access and not savings usage. Savings products are not linked to regular income streams, such as wages and pensions. In fact, 92 percent of people receive their income in cash, rather than directly into their bank account and it remains challenging for depositors to make payments or transfer remittance from their savings accounts, especially in rural areas.

On the credit side, survival analysis finds that customers have a limited borrowing relationship with FSPs; 39 percent of the borrowers exited the loan programme after the first year and 78 percent of borrowers exited within three years. Women, youth and rural customers are more loyal borrowers, yet they receive lower individual loans than men and older adults respectively. We also find strong differences in exit rates among FSPs, varying as much as 24 percent between the top performing and the weakest performing FSP, which indicate that borrower exit is at least partly under the control of the FSP and not just an issue of competition and market saturation (Mimosa, 2016). In some cases, borrowers may exit because they have become financially self-sufficient and do not require further loans. However, high exit rates also suggest that loan products are not tailored to the needs of borrowers (Churchill, C., and S. Halpern 2001; Copestake 2002).

There is a business case for FSPs to strengthen customer journeys, especially among women and youth. Depositors staying for five years with FSPs saved nearly 4.5 times more than the level of their average opening balances ($174 compared with $820 in savings account balances) and 1.8 times more than short-term depositors staying for a year ($453 compared with $820). Long-term borrowers who stayed with the FSP for three years also took up slightly larger loans ($613) compared with those who stay with the FSP for one year ($521). We estimate that reducing the number of passive savings accounts by between 10 percent and 30 percent, would mobilize an additional

$52 to $172 million of savings into the four FSPs. Likewise, reducing the borrower exit rate by 10 percent (from 39 percent to 29 percent) is estimated to contribute an additional $304 million to the loan portfolio of the four FSPs, equivalent to an increase of 24 percent in the portfolio. Improving borrower retention will help FSPs to better manage operational expenses. Assuming acquiring a new customer is at least five times costlier than retaining an existing one, it is estimated that a 10 percent reduction in borrower exits could further reduce the operational expenses of the four FSPs by around $54 million.

Key barriers to savings mobilization and factors contributing to the prevalence of passive savings accounts include the attractiveness of informal savings over formal savings, limited delivery channels for women and in rural areas, limited linkage between savings accounts and payments and income streams, and low financial literacy. Key barriers to borrower retention include weak customer assessments and limited customer loyalty programmes. Under the influence of increased customer data collection, big data analytics and digital finance movements, the potential to use consumer data for product development and policymaking has increased. Below we identify a selective set of financial technologies (fintech) and digital finance measures for FSPs and policymakers that can ease the transition from access to use of financial services.

Figure 1: Exit Rates by Gender for Depositors and Borrowers measured through Survival Analysis.

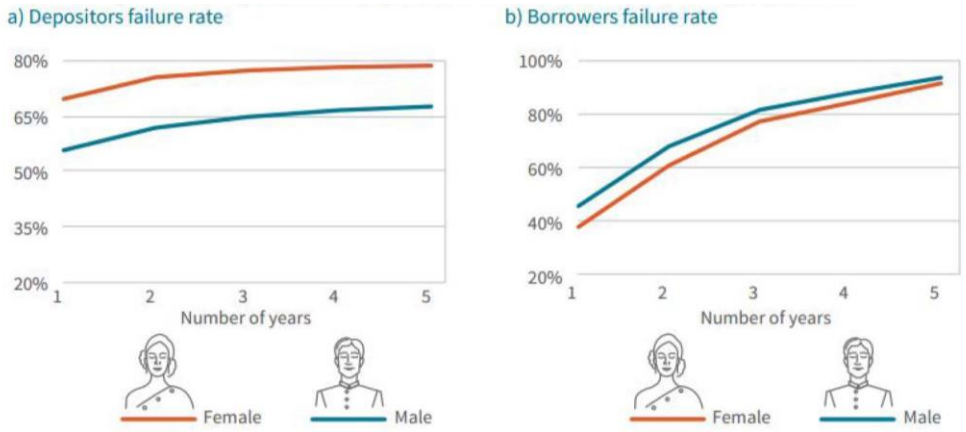


Figure 2: Gender Gap in Savings Distribution Across Different Provinces in 2015

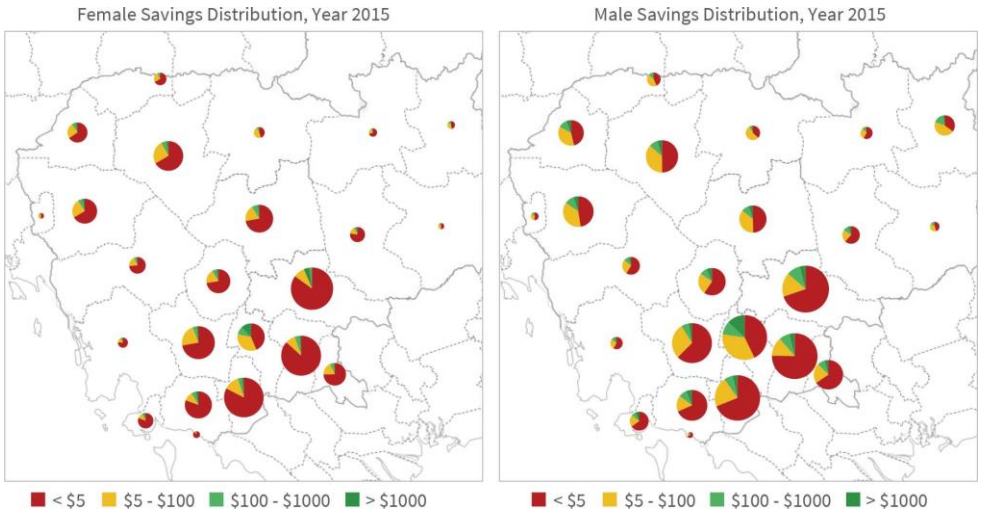
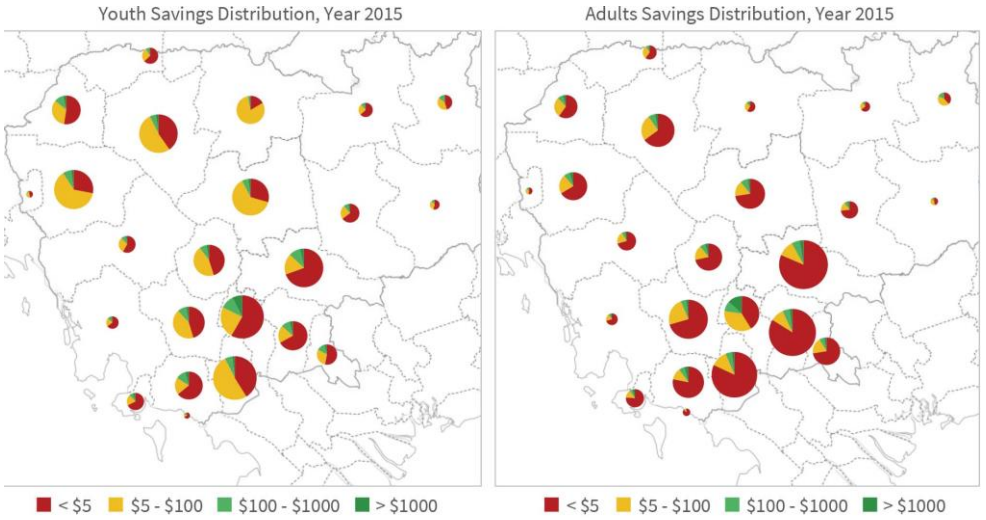


Figure 3: Youth Gap in Savings Distribution Across Different Provinces in 2015



Note: The pie charts show the distribution in terms of number of depositors with passive, small, medium, and large savings accounts. The larger the pie the more depositors in the province.

4. Discussion and Conclusion

Recommendations for FSPs

Promote mobile wallets to allow more convenient formal savings: Mobile wallets allow FSPs to move away from product-specific approaches towards a more unified customer-oriented approach by offering multiple products through one mobile phone touch point. They can offer convenient alternatives

to informal savings, as deposits can be made in relatively low and frequent amounts, mirroring the convenience of informal savings and mitigating some of the major barriers to using established banks for low-income customers. Examples of mobile wallets in Cambodia that are increasingly moving from providing payments to include other services are ABA Mobile, TrueMoney, PiPay and Wing.

Tailor delivery of savings services to female and rural customers to improve savings mobilization: In Cambodia, one specific challenge is that deposits often have to be made at the MDI branches, and users of mobile wallets and savings accounts still require agent networks and doorstep banking to penetrate rural areas. An example of tailoring delivery networks better to women customers is the LienVietPostBank in Viet Nam, which partnered with the national women's union to offer a distribution network with accompanying financial literacy workshops.

Develop delivery channels to provide better access to savings in rural areas: In Cambodia one specific challenge is that deposits often have to be made at MDI branches. Yet nearly 50 percent of adults in Cambodia take more 30 minutes to reach a bank or MDI (FinScope, 2015). As many users of mobile wallets and savings accounts still require cash, agent networks and doorstep banking are necessary in rural areas and provide an opportunity to help clients use all suitable available products. For example, Equity Bank in Kenya, which uses a rural agent network, mobilizes 20 percent of its deposits digitally. Another example includes the Pafupi savings accounts of NBS Bank in Malawi which are opened by mobile sales agents in less than ten minutes with local agents visiting women customers in rural communities. The account holders are also given an ATM card to use for withdrawals at agent locations and at any NBS Bank ATM (UNCDF MicroLead, 2018).

Link savings accounts to regular income streams such as wages and pensions and payment services to turn passive accounts active: An example is the AMRET family savings product, a digital e-wallet that aims to link the income streams of garment factory workers to their family savings and payment accounts. Likewise, mobile network operators (MNOs) and fintech providers can partner with banks and microfinance institutions (MFIs) to ensure payments and savings can be made by customers. An example is Paytm wallet in India, which was originally a payments application but now allows users to link their mobile wallet with their savings accounts and debit and credit cards from other FSPs to link savings and payments.

Enhance digital financial literacy and savings product awareness among customers: Fintech firms such as Juntos increase savings mobilization by sending personalized messages on mobile phones and social media to depositors with dormant accounts, either reminding them or making them aware of their dormant savings account. Another cost-effective approach to increase product usage is through digital financial literacy applications. For example, Wave Money in Myanmar is designing a financial gaming application where people can learn about savings, interest payments and insurance. There are even digital education tools tailored to children, including mobile piggy banks such as Ernit and Bankaroo. Another option is to design commitment savings accounts, whereby savings are held by the bank until a predetermined goal, set by the customer, has been met which have been effective in improving savings mobilization for women (Buvinic and O'Donnel, 2016)

Improve customer assessment and reward customer loyalty: Design customer loyalty programmes and reduce interest rates for long-term customers that take follow-up loans and mobilize deposits. Retained customers are more costeffective and take up larger loans and savings than new customers, yet the pricing models of banks do not always reflect this pattern. For example, although women and youth in Cambodia had longer customer journeys, they were offered similar pricing and received lower loan amounts than men. Customer data can be leveraged to develop alternative credit scoring models using mobile phone and transaction data not only to reduce collateral requirements, but also to improve on customer retention. Examples of fintech firms working on customer loyalty improvements include Retentionscience, Feedzai, and ZestFinance. Given that borrower exit is partly under the control of FSPs, good customer assessment matters. In addition to using insights from management information reports and exit surveys, FSPs can also use low-cost software such as R-Studio and Tableau to generate customer insights to develop more appropriate products.

Policy recommendations for regulators

Establish digital identity database: Harmonizing the use of national ID cards in the MDI sector can greatly reduce the time and cost of delivering financial services in Cambodia, where 95 percent of adults are reported to have such a card (FinScope, 2015). As an example, Aadhar in India was established to target delivery of financial services and government to-person payments such as subsidies, wages and pensions, combining these services with an "all-in-one" proof of national identity. Other countries such as Malaysia and Singapore also have multi-functional digital national identification cards for citizens (which serve the purpose of a driver's licence, an ID card, a health document and can serve as a digital wallet and a means of payment, amongst other functions),

which has helped increase financial inclusion. A more standardized identification database can further support the protection of customers and help to increase the monitoring of over indebtedness and cross lending for the FSPs and credit bureaux.

Explore regulatory technologies (RegTech): Technology can also help to monitor the increased number of transactions and necessary KYC regulatory compliance, while strictly complying with applicable data protection laws and regulations (e.g. General Data Protection Regulation in the European Union). Examples include IdentityMind global, Onfido, Ancoa, and AQMetrics which conduct Know Your Customer (KYC) and Anti Money Laundering (AML) fraud prevention checks. Another interesting example in the field of customer rights and consent to use data includes Trunomi, which provides customer data rights management technology to private sector companies that enables businesses to request, receive and capture customer consent to use their personal data. All this enables financial service providers to comply with regulations by putting in place auditable workflows to record and prove the lawfulness of processing of customer data.

Facilitate partnerships between banks and non-bank institutions: Mobilize small savings, especially among women and people in rural areas. Implement policies that allow mobile wallet providers to link their mobile money accounts with the savings accounts at Banks and MDIs. Allow interest bearing on savings wallets for the Cambodian FSPs to help to incentivize the promotion and usage of digital wallets. Allow MDIs to do interbank transfers with other FSPs to make savings products more convenient for account holders.

Examine the after-effects of the interest rate cap on financial inclusion: To offset the risks of an interest rate cap, FSPs are likely to offer higher loan sizes, lengthen the loan period, and charge higher upfront fees. This may then widen the gender gap in loan mobilization because rural and female customers often need lower loan amounts. The repercussions of an interest rate cap are relevant in the context of a national financial inclusion strategy which focuses on the financial inclusion of women and un- and underserved (rural) populations.

Incorporate financial service usage and customer-value insights into monitoring indicators for the National Financial Inclusion Strategy: In this regard, segmented customer data by sex and age can be used for design, implementation, and evaluation of policies aimed at improving financial inclusion and financial services usage in Cambodia.

References

1. Asian Development Bank (2013). Gender Equality in the Labor Market in Cambodia. Manila. Available from: https://www.adb.org/sites/default/files/publication/31193/gender-equalitylabor-market-cambodia.pdf
2. Banerjee, A., D. Karlan and J. Zinman (2015). Six Randomized Evaluations of Microcredit: Introduction and Further Steps. American Economic Journal: Applied Economics 2015, 7(1), 1– 21.
3. Buvinic, M., and M. O'Donnell (2016). Revisiting What Works: Women, Economic Empowerment and Smart Design. Washinton D C: Center for Global Development. Available from: https:// www.cgdev.org/sites/default/files/CGDRoadmap-Update-2016.pdf
4. Churchill, C., and S. Halpern (2001). Building Customer Loyalty. Available from: https:// responsiblefnanceforum.org/wp-content/ uploads
5. Copestake, J. (2002) Unfinished Business: The Need for More Effective Microfinance Exit Monitoring. *Journal of Microfinance*, 4(2).
6. Findex (2017). The Global Findex Database. The World Bank. Available from: http://datatopics. worldbank.org/fnancialinclusion/country/ Cambodia
7. FinScope, (2015). Results from FinScope Consumer Survey Cambodia 2015. UNCDF and FinMark Trust, Available from: https://www. fnmark.org.za/results-from-fnscope-consumersurvey-kingdom-of-cambodia-2015/
8. ILO (2015). Microfnance for Decent Work. Enhancing the Impact of Microfnance Evidence from an Action Research Programme. Social Finance Programme, ILO Manheim University.
9. *MIX Market (2015 & 2016). Microfinance Information Exchange, Inc. Available at: https://www.themix.org/mixmarket/countries-regions/cambodia
10. UNCDF MicroLead (2018). Pafupi Savings: Expanding Financial Inclusion to Rural Women.

State of financial inclusion in Malaysia

Zarina Abd Rahman
Development Finance and Inclusion Department
Bank Negara Malaysia

Abstract

Financial inclusion has improved significantly in Malaysia as a result of various initiatives to increase the access, take-up and quality of financial products and services. These improvements were largely driven by increased accessibility to financial access points across the country, more responsible usage of products and higher levels of satisfaction among financial consumers. A significant innovation in the Malaysian market is the development of agent banking, which has profoundly altered the access channels for financial services. At its core, agent banking enables consumers to obtain formal banking services by financial institutions through third party agents. While the progress in financial inclusion has been pronounced, it continues to be a strong policy priority, with more efforts planned ahead to improve outreach to address the last mile challenges. This includes, among others, greater usage of electronic payments, digital technology and cost effective solutions that are better able to transcend the unique boundaries facing this last segment. With further progress, it is expected that the unbanked population will be further reduced to 5% by 2020.

Keywords

Financial Inclusion; Measurement; Malaysia

1. Introduction

Financial inclusion is critical in reducing poverty and achieving inclusive economic growth. The usage of financial products and services provide opportunities for Malaysians to safely save and invest, borrow for productive activities and have safety nets against financial shocks. Malaysia has made significant strides in financial inclusion, as a result of intensified efforts over the years in promoting access and usage of financial services to all segments of the society.

To date, financial access points are present in almost all sub-districts, propelled by the introduction of agent banks since 2012. Meanwhile, supply-side data as provided by financial institutions show that 95% of Malaysian adults have at least one account at the formal financial institutions. In addition, Bank Negara Malaysia (BNM) periodically conducts a demand-side survey at the national level to supplement the supply-side data on financial inclusion, as well as to identify on-the-ground issues relating to financial access and usage. The latest survey conducted in 2018 indicated an improved take-up of financial products and greater usage of digital financial services.

2. Methodology

The financial inclusion measurement for Malaysia is represented by the Malaysia Financial Inclusion Index (MYFID Index). The information obtained to calculate the MYFIDS Index is based on supply and demand side data. It is made up of four (4) components:

- Convenient accessibility
- Responsible usage
- Product take up rate
- Satisfaction level

Table below is a more detailed description of the components that makes up the MYFID Index.

MYFID Index

Dimension	Indicators	Data 2018 (%)	Target 2020 (%)	Index of Each Indicator	Weight	Index of Each Dimension	Equal Weighted Dimension	Equally Distributed FII
Convenient Accessibility	% of mukim with at least 2000 population with at least 1 access point	96	98	0.98	0.5	0.99	0.25	0.91
	% of population living in mukim with at least one access point	99	99	1.00	0.5			
Take-Up Rate	% of adult population with deposit accounts	92	95	0.97	0.5	0.73	0.25	
	% of adult population with financing accounts	39	40	0.98	0.25			
	% of adult population with life insurance/takaful policies	17	40	0.43	0.25			
Responsible Usage	% of customers with active deposits	95	95	1.00	0.5	1.00	0.25	
	% of customers with performing financing accounts	98	98	1.00	0.5			
Satisfaction Level	% of customers who are satisfied –Overall financial services	73	80	0.91	1	0.91	0.25	
Index ranges from 0 – 1, with 1 being perfect financial inclusion							1	0 – 1.00

3. Results

The Overall Score for MYFID Index

The MYFID Index ranges from 0 to 100, with 100 representing the ideal state i.e. full financial inclusion. The overall MYFID Index for the year 2018 stands at 0.91, a marked improvement since 2011 by 0.2 point. The data also indicate an improved score in responsible usage, while the accessibility component remained stable. It is observed in 2018 that the scores the take-up of financial products improved further, particularly in financing products. Likewise, the satisfaction rating given for the overall financial services has also improved to 73%. The satisfaction component included in the MYFID Index looks into the satisfaction level of the financial product and services. The improved satisfaction level is contributed by positive experience given by Malaysians for the services received for savings, transaction and payment services.

Key Findings from the Demand-Side Survey 2018

The take-up rate focusses on adults aged 15 years and above that is holding the following financial products i.e. deposit accounts, financing

accounts and life insurance/takaful policies [1]. There appears to be no significant difference in ethnicity and gender for those that are in the financial system. Product holding is lowest amongst those in the rural areas and household income less than RM3,000 per month. The vast majority of the account owners owns the deposit account where 9 in 10 of the Malaysians will have at least a saving account. Owning an account is an important first step to financial inclusion. According the the Global Findex Database 2017, in order to fully benefit having an account, people need to be able to use it in a safe and convenient ways[2].

In 2018, the user base for both mobile and internet banking have grown since 2015. The usage of internet banking has increased at a much faster pace compared to mobile banking. Digital usage is also seen more obvious in the market centres and amongst those in the T20 segment, with a monthly income level of above RM7,000 per month. The usage of the newly added digital services, i.e. payment card and mobile payment is high even though they were only recently introduced to the market. Similar to mobile and internet banking, the user base of these newly added services are more commonly used by the male segment as well as those in the younger age bracket i.e. between the age of 20-44 years old. The common reasons mentioned by those using digital banking services are for the usage of digital services is for the purpose of money transfer and bill payment.

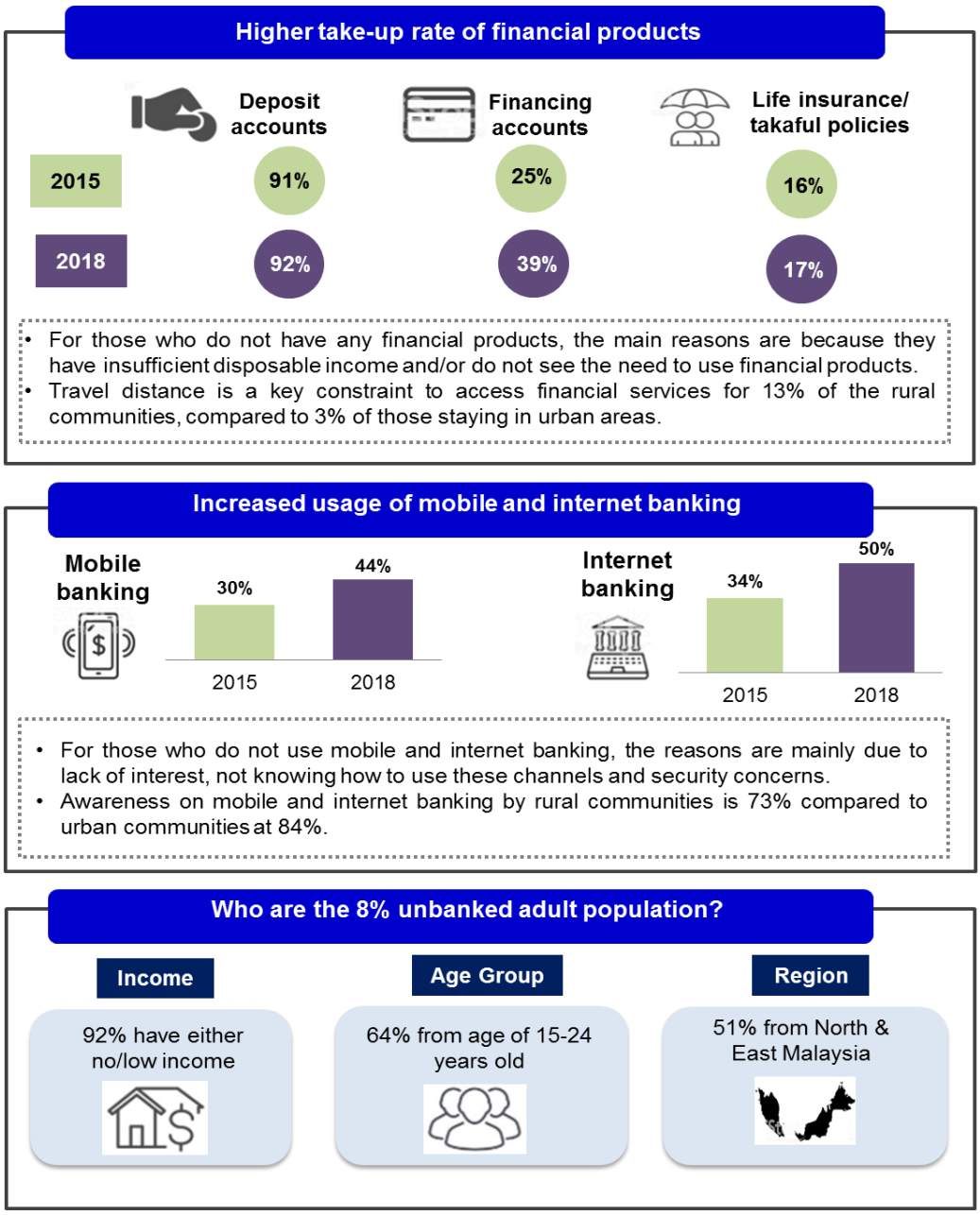
The proportion of the unbanked is 8%, maintained since 2015. These segment do not hold any financial products and services. The unbanked segment is highest amongst those between the age of 15-19 years old. The reason for not having a bank account when they were below 17 years old was because of not having the need for it and also that they do not have any money to put in the account.

In addition to young age segment, one's income level also contributes to the unbanked situation. It is observed that Malaysians with no income as well as those with monthly household income of less than RM3,000 i.e. B40, are also among the unbanked. Likewise, Malaysians residing in rural areas specifically Malaysians in the northern region, and in East Malaysia.

[1] Deposit Account products measured here comprise of all forms of deposit products namely saving account, current account, fixed deposit account. Financing Account products is comprised of all forms of financing such as secured, unsecured loans, mortgage, car etc., includes credit cards.

[2] The Global Findex Database 2017, Measuring Financial Inclusion and the Fintech Revolution.

Highlights from the Demand-Side Survey 2018

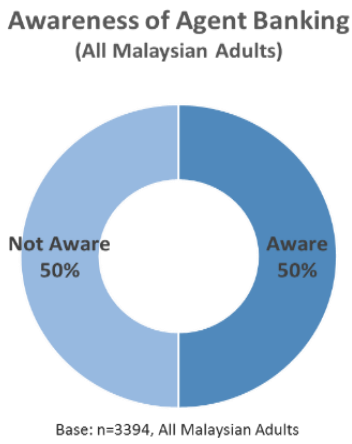


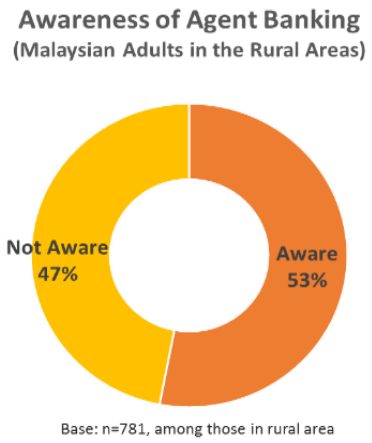
As for the barriers to participation in the financial system, the survey shows that more that 45% of Malaysians with no financial product/service have cited "having insufficient disposable income/money" to be the main reason, followed by not having the need for the financial services.

The Growth of Agent Banking

Agent banking was introduced with the aim to provide financial access to the population living in rural areas which have limited number of bank branches. Agent banking enables consumers to obtain financial services by licensed financial institutions through third-party agents such as retail outlets, petrol stations and post offices.

The agent banking regulation allows agents to facilitate online-real-time transactions with biometric identity verification. The agents can offer basic financial services to consumers on behalf of the financial institution, such as opening savings accounts, deposit or withdraw money, making bill and loan payments, as well as making domestic fund transfers.





The 2018 survey findings shows that about 5 out 10 Malaysians are aware of agent banking. The awareness is considerable high and more Malaysians are using the services since 2015. For those that do use agent banking, it is mainly for the purpose of bill payment, withdrawal and deposit. There is a mark increased in the usage of these three main services, amounting to more than 50 % of the total financial services provided by the agents.

It is envisaged that as people become more familiar with agent banking, it has the potential to promote active use of accounts and to become the distribution channel of a wider variety of financial services such as micro-financing and micro-insurance among the population in rural areas.
[1]

[1] World Bank 2017:The Malaysia Development Experience Series. *Financial Inclusion in Malaysia: Distilling Lessons for Other Countries*.

4. Conclusion

The inclusion level for Malaysia has improved tremendously since 2015. The findings from the supply-side and demand-side survey in 2018 shows that there is higher take-up in financial products in addition to high accessibility. It is also noted that there is a need to create more awareness of accessibility and targeting more responsible usage. It is estimated that the unbanked population will be further reduced to 5% by 2020.

The mechanics to greater responsible usage amongst Malaysians will be driven by improving the financial knowledge of Malaysians by helping consumers make better financial decisions. This is done by ensuring that Malaysians have a better grasp in the financial literacy and capability to promote more positive experience from their participation in the formal financial system. Furthermore, responsible usage will also be enhanced by correcting the financial behaviour and attitudes of Malaysians. This is done correcting the behaviour of money management; budgeting, spending, savings and long-term planning.

BNM views digital innovation as an important enabler to promote greater access, quality and responsible usage of financial services. Improving digital literacy and focus on effective education, support and protection for financial consumers is also required to improve responsible usage amongst Malaysians. This ensures that Malaysians are able to utilize the account owns and have better participation in the financial system.

References

1. Bank Negara Malaysia. March 2019. *The Financial Stability and Payment System Report 2018*. Kuala Lumpur: Bank Negara Malaysia.
2. Demirgüç-Kunt, Asli, Leora Klapper, Dorothe Singer, Saniya Ansar, and Jake Hess. 2018. *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution.* Washington, DC: World Bank.
3. Jose de Luna Martinez, lead author. May 2017, Knowledge & Research Hub: The Malaysia Development Experience Series. *Financial Inclusion in Malaysia: Distilling Lessons for Other Countries*. Kuala Lumpur: World Bank.
4. Bank Negara Malaysia. March 2016. *The Financial Stability and Payment System Report 2015*. Kuala Lumpur: Bank Negara Malaysia.
5. Zarina Abd Rahman. 10 May 2013. Central Banking Journal: *Bank Negara Malaysia's Approach to Developing a Financial Inclusion Index.* London: Central Banking Publications

New tools for data collection in Swedish surveys on use of fertilisers and animal manure and cultivation measures in agriculture

Ylva Andrist Rangel, Daniel Eiserman, Lena Otterskog, Anna Redner
Statistics Sweden, Örebro, Sweden, *ylva.andrist-rangel@scb.se

Abstract

Technology is making rapid progress in agriculture, and the typical Swedish farmer is more and more likely to have access to an array of different technological tools, ranging from PCs, smartphones, tablets, GPS and farm management software, to self-steering machinery etc. At the same time, many statistical surveys are struggling with high costs of data collection, declining response rates, and obligations to reduce the overall response burden. Therefore, to meet the increased demand for high quality agri-environmental statistics, smart and cost-effective strategies for data collection are essential. The aim of the present paper is to show examples from recent advances in data collection implemented in a number of agricultural surveys at Statistics Sweden. The methods were developed for the two sample surveys *Use of fertilisers and animal manure* and *Cultivation measures in agriculture.* An overall project was initiated and consisted of three phases: 1) investigating alternative data sources, 2) developing a new web system for data collection, and 3) increasing survey coordination. Pilot studies showed that several of the variables, previously obtained via data collection directly from farmers, could be extracted from a combination of different administrative registers, a Land Parcel Identification System, and from farm management software. A new web system for data collection was launched in late 2016, developed by IT experts in close collaboration with statisticians and agronomists at Statistics Sweden. Several agricultural surveys are coordinated within the same system, and new ones can easily be added as needed. A tool for data import of XML files from farm management software was designed and incorporated into the web system to make it easier for the farmer filling in the questionnaires. The system has now been running for three years. The overall achievements include more efficient data collection, lowered costs in the long term and decreased response burden for the farmers.

Keywords

administrative registers; mixed-mode; farm management software; response burden; LPIS

1. Introduction

Technology is making rapid progress in agriculture, and the typical Swedish farmer is more and more likely to have access to an array of different technological tools. There is also a generational renewal under way, in which young farmers transitioning in are more prone to be open-minded to new technologies and tend to take digital solutions for data transmission for granted. The array of tools ranges from PCs, laptops, smartphones, tablets, GPS and farm management software, to self-steering machinery and precision farming. At the same time, many statistical surveys are struggling with high costs of data collection, declining response rates, and obligations to reduce the overall response burden for companies, including agricultural holdings, i.e., the farmers. Therefore, to meet the increased demand for high quality agri-environmental statistics, smart and cost -effective strategies for data collection are essential. In Sweden, the responsibility for statistics in the area of agriculture, including agri-environmental statistics, is shared among three authorities: the Swedish Board of Agriculture, the Swedish Chemicals Agency and Statistics Sweden; a range of surveys are conducted on a regular basis. The aim of the present paper is to show examples from recent advances in data collection implemented in a number of agricultural surveys conducted at Statistics Sweden.

2. Methodology

The methods were developed for the two intermittent sample surveys Use of fertilisers and animal manure and Cultivation measures in agriculture. The main target characteristics in these surveys are: quantities of nutrients applied to crops and different measures of aspects of handling of manure, such as spreading technique and storage capacity, as well as age of fallow (set-aside) land, age of temporary grasses, tillage methods, catch crops and liming.

Before the project started, both surveys had been struggling with gradually increasing costs of data collection. The Use of fertilisers and animal manure survey, in particular, struggled with costs, as it was entirely carried out via telephone interviews. At the same time, there was a demand by data users for greater detail in the statistical output from these surveys. Attempts to obtain increased government funding had not been successful. Therefore, a major review of the overall process and the methodologies of data collection was unavoidable. The main data users and stakeholders were consulted. They were presented with two options: 1) a lower level of detail in the statistical output from the surveys, or 2) prolong survey intermittence from two to three years. There was consensus for the second option. At the same time, several opportunities for temporary additional funding arose and two pilot studies were initiated, aiming to reduce the costs of data collection in the long term, minimise the response burden and also, if possible, increase the level of detail

in the statistical output. The overall project was divided into three phases: 1) investigating alternative data sources, 2) developing a new web system for data collection, and 3) increasing survey coordination.

2.1 Investigation of alternative data sources

In the first phase of the project, all of the present target characteristics in the two surveys were thoroughly reviewed in close collaboration with the main data users and stakeholders. Target characteristics with low use or low relevance were marked and excluded from the next step, which involved assessing possible alternative data sources for the target variables, other than data collection directly from the farmers. Until then, preliminary crop areas, used as supporting variables in the data collection, had been obtained from the Integrated Administration and Control System (IACS) register, and final crop areas and livestock types and numbers, used in the final estimations, had been obtained from the Farm Register. Two new registers were examined as potential additional data sources: the Swedish Block Database – a Land Parcel Identification System (LPIS) – and the register of crop areas with subsidies for organic farming. The second alternative data source involved collection through automatic data transmission from professional farm management software, which was examined for the first time for this purpose. This first phase of the project resulted in a list of variables that could potentially be collected from alternative data sources (see Results, Table 1).

2.1.1 Land Parcel Identification System

The Swedish Block Database, which is owned by the Swedish Board of Agriculture, is a database that contains data from all utilised agricultural areas for which agricultural subsidies are applied. A "block" is defined as a surface with permanent boundaries such as roads, streams, forests or another farmer using the land. The database contains coordinates of the geographic position of the centre of each block, as well as data on area and crop grown. A new version of the database is set up every year. Statistics Sweden ordered copies of the Block Database for the years 2008–2014 to explore the possibility of collecting data from an LPIS for the age variables for fallow and temporary grasses, respectively. All blocks in the Block Database with temporary grasses in 2014 were matched with blocks from the 2008–2013 databases to investigate for how many consecutive years temporary grasses had been cultivated in the same block. However, one obstacle was that a block can change geographic coordinates due to changes in the boundaries. Another challenge was that a block can be divided into many parcels with the same coordinates. Therefore, a

decision was taken to only match blocks that consisted of a single parcel of temporary grasses. This meant that 94 percent of the total number of blocks with temporary grasses could be matched and the age of the temporary grass could be determined. However, when the same method was applied on blocks with fallow (set -aside) land, only 67 percent of the blocks consisted of one parcel. Therefore, further tests were carried out to find out whether it was possible to increase the number of blocks that could be matched. For example, in one test, block size was used as an additional matching variable. First, the exact block size was used, and then a deviation of +/-10 percent of the size was allowed to find matching block pairs among the different years. Another test included information on buffer zones, as the geographical positions of these parcels are constant over time. Finally, after concluding the tests, an optimal method was established. For fallow land, this method increased the number of blocks that could be matched to 85 percent.

2.2 Development of a new web system for data collection

In 2015, a project was launched to upgrade and modernise the web system that had been in use since 2005 in the surveys on crop production and autumn sown areas (Ländell et al., 2004). One of the main aims of this modernisation was to facilitate future additions of new modules, such as new surveys. This opened up the possibility for existing sample surveys in agri-environmental statistics to join the project. A group of IT experts, methodologists, statisticians, agronomists and responsible line managers was established at Statistics Sweden. An agile approach was used, based on Scrum™ (Schwaber & Sutherland, 2017), with sprints, demos and frequent meetings within the group to discuss issues such as definitions and harmonisation, code lists, product backlog items (PBIs), and prioritisation. The composition of the group could vary to some extent based on the purpose of the meeting. The aim was to fully replace the old system used in crop statistics, which was a mixed- mode data collection system with self-administered web questionnaires and an interface for telephone interviews, built -in tools for administration, questionnaire validation, data checking and editing. The new system had to additionally provide new functionality (see Results, Table 2), mainly based on suggestions provided by interviewers who had worked on the surveys in the preceding systems. One of these new functions was a tool for importing data from professional farm management software.

2.2.1 Tool for data import from farm management software

During the first phase of the overall project, an initial contact had already been taken with the two dominating companies offering farm

management software on the Swedish market. In the beginning, both companies were interested in the proposal, and discussions on the data transfer procedure were initiated. Several possibilities were discussed, but can be summarised by two main options: 1) a transfer function in the farm management software, where the farmer actively initiates the process of exporting data of a set of known variables into Statistics Sweden's new web system for agricultural statistics, or
2) the software providers allow Statistics Sweden access to their databases with all data from their clients (subject to the client's consent). There was consensus for the first option for a variety of reasons. Firstly, the farm management software clients (the farmers) were most likely not willing to give consent for unconditional secondary use of all their data (as in option 2). Secondly, option 2 would involve data quality issues, as farmers use the software for different purposes: from purely planning, a certain degree of follow-up, to a complete accounting of last year's result in terms of physical input and output, and dates for different measures. Hence, in option 2, neither the state nor the quality of the data would be known to Statistics Sweden. At this stage of the discussions, one of the software providers informed that they wanted to postpone possible cooperation. Hence the next step, which was designing an interface for exporting data in the farm management software, was only implemented by one of the software providers. At the same time, the tool for importing data was designed and incorporated into the new web system for agricultural statistics at Statistics Sweden. XML was settled on as the format for the data transfer.

2.3 Coordinating surveys

Since 2006, data collection for either the Use of fertilisers and animal manure survey or the Cultivation measures in agriculture survey had been carried out every year. The review included examining whether these two surveys could be merged into one. Issues that had to be considered were sampling frame, sampling method, number of strata, length of new/merged questionnaire, time series constraints, and response burden. In addition, the new web system for agricultural statistics was designed for coordinated administration and data collection of all ongoing surveys in the system, independently of agricultural domain, i.e., crop production or agri-environmental statistics (pesticides, fertilisers, manure and cultivation measures).

3. Results and Discussion

3.1 Alternative data sources

Several of the variables (Table 1) previously collected directly from farmers in the surveys on Use of fertilisers and animal manure and Cultivation measures in agriculture, via telephone interviews only, or mixed- mode, could instead be extracted from a combination of various administrative registers – including the LPIS – and from farm management software (Redner & Andrist Rangel, 2016). The review of the data collection process also showed that only a limited number of target characteristics could be excluded from future surveys, due to continued user needs.

Table 1. Results of the review of target characteristics and variables in the surveys Use of fertilisers and animal manure and Cultivation measures in agriculture.

Variables available from LPIS and from the register of crop areas with subsidies for organic farming	**Variables available from farm management**
Age of temporary grasses and grazings	Field identification and size
Age of fallow (set aside) land	Crop
Preceding crop (crop rotation)	Type of fertiliser applied (mineral/manure/ other organic fertilisers/soil amendments)
Organically cultivated fields (partly derived data)	Date of application
	Application rate (kg or tonnes per ha)
	Spreading technique for manures

Concerning the use of the Swedish Block Database (LPIS), the final method that was developed (Andrist Rangel et al., 2016) was implemented in the survey on *Cultivation measures in agriculture 2016* (Statistics Sweden, 2017), where it was used to collect data on age of fallow. However, it was only applied to the objects (farms) included in the sample, and thus the full potential of the register was not utilised in the 2016 survey. Nonetheless, the farmers' response burden was reduced, which was the original goal of the new approach. The reason why the whole register was not used was because other data related to fallow, which had been collected via the questionnaire, was to be cross- utilised with the data on age. However, in future use, or in use for other purposes, it would be preferable to use the whole register. The Swedish Block Database has very good coverage, as it includes information on more than 99 percent of the agricultural area (Swedish Board of Agriculture, 2018); this means that this approach would eliminate the sampling error present in the current design. The next step will be to investigate the methodological

aspects of a mixed design, i.e., a combination of a sampling design and a complete enumeration (via registers) in a single survey.

3.2 New web system for data collection

The new web system for data collection was launched in late 2016 and has now been running for three years. Response rates have varied between 81 percent and 96 percent, mainly depending on the survey, and the share of web responses has increased from roughly 30 percent to 40 percent during the period.

3.2.1 Work method

The agile approach, including the demos by the IT experts, provided an opportunity for the end users to give feedback on the ongoing development of the web system. This led to continuous amendments and additions to the original requirements. This approach resulted in a final product with new and enhanced functionality (Table 2) that met user needs very well and fulfilled the high expectations of the customers. For the agri-environmental surveys in particular, this was a major step forward, as one system would now replace four systems. The only potential drawback was an awareness of the high costs of frequent meetings in relatively large groups. However, the risk of using another work method, with fewer meetings, could have led to unexpected costs due to errors or misunderstandings. Also, during the project, optimal constellations were arrived at for different types of meetings, which led to a drop in the total number of hours spent on meetings.

Table 2. Functionality in the old and new web systems (for crop statistics 2005–2015; and agricultural statistics from 2016). x = existing functionality, xx = enhanced functionality compared to the old system.

Functionality	**Old (2005–2015)**	**New (2016–)**
Mixed-mode data collection for web and telephone	x	x
Role-based user interface (respondents, interviewers, administrators, editing staff, readers)	x	xx
Designed as a general platform (easy to add on similar surveys)		x
Support for historic data		x
One single login for respondents (farmers) and other users	x	x
Data collection administration	x	xx
Coordinated data collection for crop and agri-environmental statistics		x
Prefilled cells (e.g., data from administrative registers)	x	x
Tool for data import from farm management software		x
Transfer of data between surveys (to avoid multiple data collection for the same variable)		x
Tracking and re-creation of previously saved data in the questionnaire		x
Trace logs		x
Calculation assistance (e.g., conversion from volume to weight)	x	xx
Real time calculations (e.g., conversion of kg fertiliser to kg nutrient; summing of annual totals – for verification)	x	xx
Questionnaire validation (both for respondents and editing staff)	x	xx
Manual data checking	x	x
Automatic data checking		x
Data editing	x	x
Tabulated reports		x
Graphic reports (e.g., scatter plots)	x	xx
Exports to production database	x	xx

3.2.2 Implementation of a data import tool

The two options that were discussed concerning the transfer of data from the farm management software provider to Statistics Sweden were quite different in terms of the concept of data collection. In the first option, which was the selected approach, only farmers who are included in the samples of the surveys are encouraged to use the tool, even if it is visible in the farm management software to all users. When the data has been actively exported from the farm management software and imported by the respondent into the web system for agricultural statistics, the relevant cells in the questionnaire autofill with those data. Once the data is in the questionnaire, it has the same properties as data entered manually by the farmer or interviewer. It can be verified, validated, changed, deleted, or saved. In this way, the imported data will be checked for quality in the same way as any other data entered into the questionnaire. The second option could be described as a big data concept, since data could be potentially collected from all farms that have an active licence for the software, regardless of whether or not they are included in the sample, or even in the sample frame of the surveys. However, using option 2 as the only data collection method would lead to a bias in Swedish agricultural statistics, since the software users are not representative of the whole target population (see below).

The tool for data import was implemented in the *Use of fertilisers and animal manure 2016* survey, and the following year also in the *Use of pesticides 2017* survey. The information that could be imported was: *field* , *crop*, *type of fertiliser/pesticide, date of application* and *application rate*. The tool made it easier for respondents to fill in the web questionnaires, especially among those who had farms with many crops and types of fertilisers/pesticides, since a large part of the questionnaire could be autofilled. Unexpectedly, use of the tool was limited – about 5 percent of the total web response in both 2016 and 2017. The actual export function was not a new feature for the software users, as the farm management software already included a tool for data export, for example for the yearly applications to the Swedish Board of Agriculture for area-based agricultural subsidies via IACS. In Sweden, farms that use professional farm management software account for a major part of the agricultural area, but represent a smaller part of the total number of farms, as mainly larger farms invest in such tools. Hence, only a limited part of the sample is able to use the tool. Statistics Sweden was aware of this situation before the function was developed and implemented in the web

system. Despite this, Statistics Sweden had hoped for more extensive use of the tool. Therefore, ahead of the upcoming survey in 2019, efforts will be made to raise awareness about this statistical reporting tool among farmers and within the agricultural advisory service.

3.3 Increased survey coordination

A goal and a successful outcome of the overall project was increased coordination of surveys. Three separate groups of surveys, with different funding and different government agencies responsible for the statistics, are now coordinated throughout the data collection process. This leads to significantly decreased costs of administration and interviewing, enhanced possibilities to cross -check data, and a reduced response burden, both absolute and perceived. The number of contacts with the respondent can be minimised, and information obtained from a selected farm may be used in all ongoing surveys within the web system. In addition, the advantage of the new web system is that that new surveys can easily be added, which facilitates future coordination of more surveys. The merger of the surveys *Use of fertilisers and animal manure* and *Cultivation measures in agriculture,* together with the prolonged survey intermittence from two to three years, also made it possible to increase the sample size which, in turn, allowed for a greater level of detail in the statistical output.

4. Discussion and Conclusion

The investigation into the use of alternative data sources showed that several variables could be obtained from a combination of different administrative registers, an LPIS, and from farm management software. Systematic reviews of stakeholders' data needs and existing data sources should be made on a regular basis to avoid unnecessary direct data collection from agricultural enterprises and households.

There is strength in combining expertise and knowledge from different types of staff when working in large projects focused on developing new tools, such as a web system for collecting agri-environmental data.

For farms with many crops and many applications of fertiliser and/or pesticides, a tool for data import from farm management software reduces the response burden. Direct data transmission will probably have a great potential for the production of agricultural statistics, as the agricultural sector, and therefore also the relevant information, is becoming more and more digitalised.

The coordination of surveys, largely assisted by the web system for agricultural statistics, has led to a more efficient data collection process, with reduced response burden and lowered costs in the long term. This, in turn, has

made it possible to increase the level of detail in the statistical output, one of the most prioritised stakeholder needs.

Acknowledgements

This work was supported by Eurostat via an EU grant, the Swedish Agency for Marine and Water Management via a Swedish Environmental Emissions Data methodology project, and by direct government funding.

References

1. Andrist Rangel, Y., Fägerlind, K., Ländell, G., Otterskog, L., Redner, A. & Wahlstedt, G. (2016). Improvements in agri-environmental and grassland statistics in Sweden. Data collection from LPIS – Land Parcel Identification System – Excretion factors for livestock - Coefficients for harvested crop products and crop residues. Final report. PM RM/Lantbruksstatistik 2016:1.
2. Ländell G., Engström N.Å., Lagerson, N.G. & Sundqvist, J.O. (2004). Data collection with the aid of Internet. The implementation of a Tapas Action 2003. PM RM/L 2004:1.
3. Redner, A. & Andrist Rangel, Y. (2016). Metodutveckling för datainsamling i undersökningarna om gödselmedel och odlingsåtgärder. SMED Intern Rapport 2016.
4. Schwaber, K. & Sutherland, J. (2017). The Scrum Guide™. The Definitive Guide to Scrum: The Rules of the Game.
5. Statistics Sweden (2017). Cultivation measures in agriculture 2016. Set-aside land, temporary grasses, tillage methods, catch crops and application of lime on arable land. MI 30 SM 1703.
6. Swedish Board of Agriculture (2018). Use of agricultural land 2018. Final statistics. JO 10 SM 1802.

NASS Geospatial applications from the cropland data layer

Avery Sandborn, Rick Mueller, Claire Boryan, Dave Johnson, Zhengwei Yang, Lee Ebinger, Arthur Rosales, Patrick Willis, Robert Seffrin, Rachel Jennings, Matt Deaton, Hubert Hamer*
United States Department of Agriculture/National Agricultural Statistics Service

Abstract
For more than a decade, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) has produced the Cropland Data Layer (CDL), a geospatial crop-specific land cover product covering the conterminous US. The CDL is increasingly being integrated into NASS's programs. An early application was the development of remotely sensed based crop acreage and yield estimates, which are independent of the survey estimates. Numerous derivative products have been created from the CDL. The Cultivated Layer and Crop Frequency Layer identify, respectively, where and how often a crop is planted. The June Area Survey (JAS) sample is drawn from the NASS area frame, and the stratification of the frame is now based on the CDL. Information for imputation of crop type and acreage for the JAS now comes from the CDL. Most recently, the CDL has been used as a primary input into disaster assessments. Now it is being considered as the foundation for integrating the diverse data sources available to NASS. This paper will discuss the creation and uses of the CDL and its derivative products and then focus on their potential future uses within the Agency.

Keywords
CropScape; Cropland Data Layer; Cultivated Layer; June Area Survey; Disaster Assessments, Data Integration

1. Introduction

The mission of the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) is to provide timely, accurate, and useful statistics in service to U.S. agriculture. To help achieve NASS's mission, the Cropland Data Layer (CDL) utilizes remote sensing techniques to provide operational in-season acreage estimates to the NASS Agricultural Statistics Board and Regional Field Offices. The CDL is a 30-meter national raster, geo-referenced, crop-specific land cover classification product produced annually by NASS (Figure 1). The CDLs are published on the NASS CropScape web application (USDA/NASS Cropland Data Layer 2018). CropScape is designed to provide the public with open access to serve the CDL with interactive visualizations, data dissemination, geospatial queries, and

online analytics (Han et al. 2012). Within NASS, CDL products have been used in a variety of research and operational applications, including masking crop extent for yield assessments (Johnson 2012), disaster assessments (Boryan et al. 2018), area frame stratification (Boryan and Yang 2017), improving estimates for the number of farms at the state and national-levels, and June Area Survey (JAS) imputation. Monitoring U.S. agriculture is important for food security and the CDL program provides a consistent geographical extent and spatial resolution over the past eleven years serving that purpose.

The original purpose of the CDL program was to generate acreage estimates of major commodities to reduce sampling error at the state, agricultural statistical district, and county-levels for internal NASS use by the Agricultural Statistics Board (Allen and Hanuschak 1988). The CDL is a supervised land-cover classification utilizing a decision tree machine learning approach using optical satellites while leveraging ground reference data collected from the USDA Farm Service Agency (FSA), as well as ancillary data from the U.S. Geological Survey (Boryan et al. 2011). Medium resolution satellites such as Landsat 8, Disaster Monitoring Constellation Deimos-1 and UK2, Resourcesat-2 LISS-III, and Sentinel-2 are used to collect imagery throughout the growing season. The CDL leverages ground reference data and multiple image collections across the growing season to capture the varying crop phenologies and derive a crop-specific land cover classification of planted area. CDL uses and applications external to NASS have been identified (Mueller and Harris 2013), and best practices and recommendations on studies with the CDL dataset have been developed (Lark et al. 2017). This paper focuses on internally driven applications that leverage the CDL product for improvement of agricultural statistics and geospatial data products.

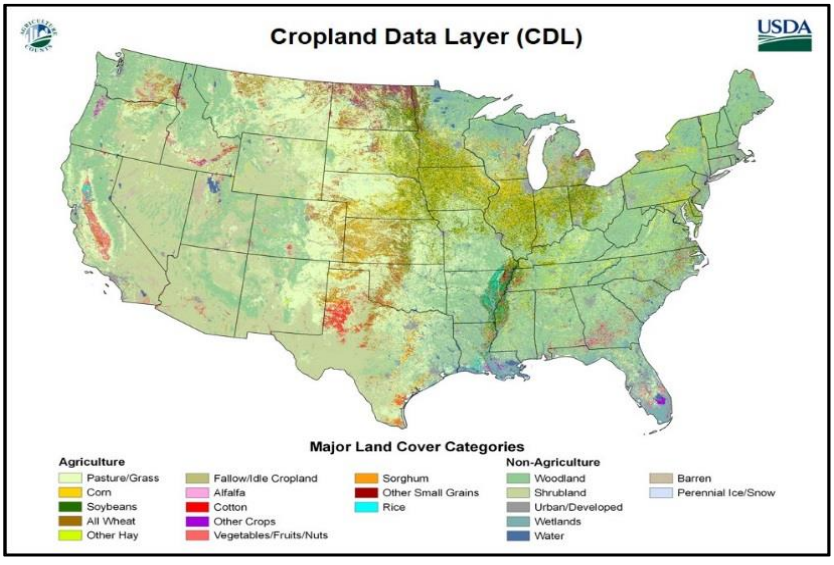


Figure 1: The Cropland Data Layer (CDL).

2. Current methods and products

Frequency Layers

The U.S. national scale Crop Frequency Layers, which are derivative products of the CDL, identify planting frequency or intensity at the 30-meter pixel-level for corn, cotton, soybeans, and wheat. The four national-level crop frequency layers were built and validated with ground reference data from the FSA. These layers provide indicators for future crop planting, which is valuable for improving agricultural survey estimates, agricultural production planning, water resource management, natural resource allocation, and conservation (Boryan et al., 2014a). These layers are available for visualization, analysis, and download from CropScape.

Cultivated Layer

Another important derivative product of the CDL is the Cultivated Layer (Figure 2). The Cultivated Layer is a highly accurate characterization of cultivated land across the continental U.S. Unlike the original CDLs, which include more than 100 different crop categories, the Cultivated Layer includes only two categories: cultivation and non-cultivation. The cultivated land cover classes include tilled/planted crops and does not include hay, other hay, pasture, or rangeland. For operational purposes, five years of CDLs are combined to create the national-scale Cultivated Layer, which helps to reduce errors in identifying cultivated land pixels (Boryan et al., 2012). The Cultivated Layer is defined as any pixel identified as cultivated two out of the last five years or identified as cultivated in the most recent year. The 2018 Cultivated Layer is produced using 2014-2018 CDLs and validated using 2014-2018 FSA data.

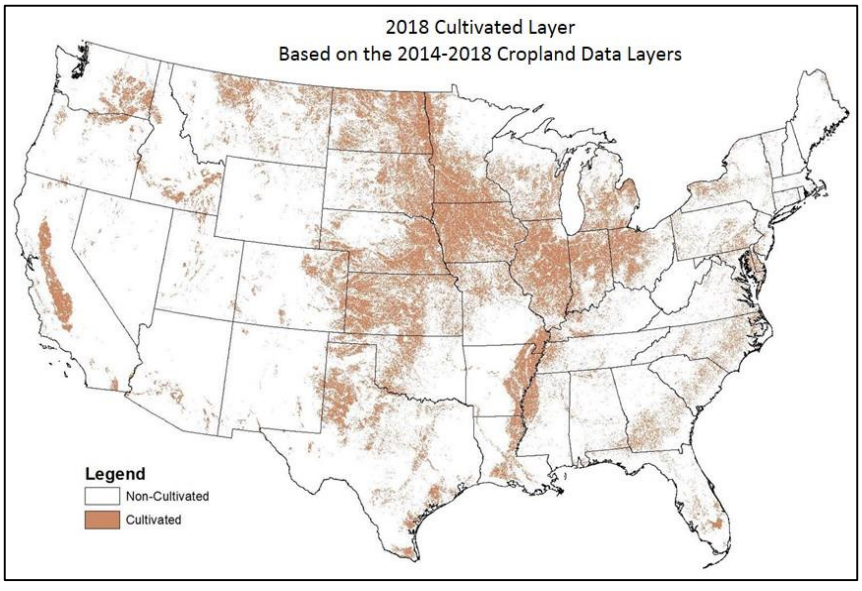


Figure 2: Cultivated Layer

Crop Yield Modelling

The CDL is also a foundational element for remotely-sensed yield estimation within NASS. Operational efforts to date have focused on corn and soybean yields within the Corn Belt at county and state-levels (Johnson, 2014). Research efforts are ongoing for other crops and areas of the U.S. (Johnson, 2016). Because crop progress, condition, and yield are dynamic, high revisit rate optical and thermal satellite imagery is required. Weekly observations are minimally sought and the best source for that type of information has been from the Moderate Resolution Imaging Spectroradiometer (MODIS), which has been in existence for nearly two decades. MODIS, while temporally ideal, is compromised spatially as it is only 250 meters in resolution, and is a challenge to identify with sufficient map precision field crop types and their boundaries. Thus, the 30 meter spatial resolution CDL fills this need.

Having highly accurate field-level crop type information ultimately allows one to isolate or mask the MODIS observations to only crop specific areas. This provides a clean signal of the vegetation profile throughout the growing season and improves yield model performance. Because the CDL is generated within season it can be leveraged as early as the August Crop Production report. More generally, the CDL can also be used to create year agnostic crop type or area masks (Johnson, 2012) by integrating the data over several seasons. This is useful for scenarios, such as needing to mask MODIS data that exists prior to the 2008 availability of national-level CDLs or when needing a crop predictive layer within season even before the CDL is available. This CDL-derived crop information is also useful for simplified yield modelling circumstances where managing year-specific crop maps is unwieldy.

Area Frame Stratification

A new automatic area frame stratification method (Boryan et al., 2014b; Boryan and Yang, 2017) was recently developed and implemented for NASS operations based on the Cultivated Layer. The NASS state-level area frames are stratified based on percent cultivated cropland within NASS Primary Sampling Units (Figure 3) and are used to select samples for NASS's annual JAS. For more than fifty years, the traditional area frame stratification method was conducted using visual interpretation of aerial photography or satellite data (Cotter et al., 2010). Research findings show that using the automated stratification method, based on the CDL, significantly improves area sampling frame stratification accuracies in intensively cropped areas (>75% cultivation) and overall stratification accuracies when compared to traditional stratification based on visual interpretation of aerial photography or satellite data, while reducing the cost of area frame construction (Boryan et al., 2014b).

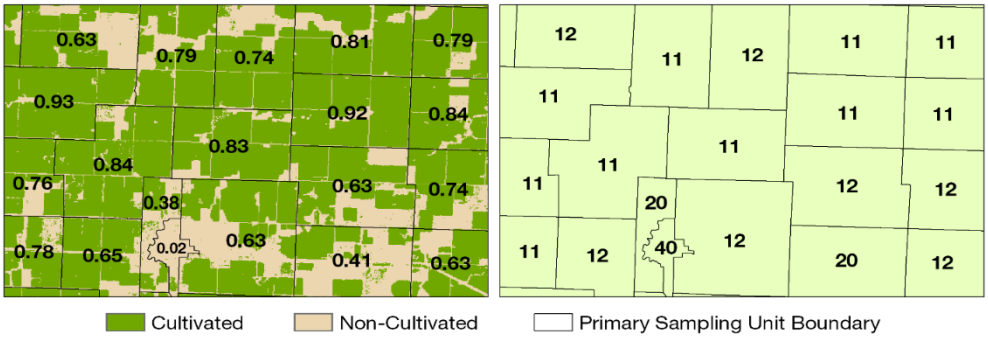


Figure 3: (left) CDL Cultivated Layer based Primary Sampling Unit with percent cultivation; (right) the same area sampling frame with CDL derived strata, where strata 11 represents greater than 75% cultivated, strata 12 represents 51-75% cultivated, strata 20 represents 15-50% cultivated, and 40 represents less than 15% cultivated. resents less than 15% cultivated.

Though the new fully-automated stratification method has improved stratification efficiency, objectivity, and accuracy in the intensively cropped areas, it achieves lower accuracies in low or non-agricultural areas. Consequently, a hybrid approach that integrates the automated stratification results with manual editing/review methods was implemented operationally. Since 2014, ten state-level area frames were built using the new integrated operational process. Boryan and Yang (2017) describe the area frame improvements, which include improved accuracy (30% improvement) and a reduction in labor costs, as well as other measurement criteria. As an example, the traditional Oklahoma area frame was constructed in 4,552 employee hours, while the hybrid frame required 1,980 employee hours.

June Area Survey Imputation

The JAS is the largest NASS annual survey in which approximately 9,000 square mile sample segments are visited by enumerators at the beginning of June to collect crop type and acreage information. Estimates of crop acreage and livestock inventories are based on these survey data (USDA/NASS, Understanding Statistics 2018). Although the JAS sampling frame provides complete coverage of all agriculture activity occurring on land pertaining to the target population, operator non-response and inaccessibility leads to the need to impute missing information. JAS imputation was traditionally based on historical records, which are not always available for the selected samples.

The CDL data provide a reliable alternative data source for imputation. Now NASS statisticians import the digital JAS segment boundary files directly into CropScape for analysis and use CDL data to 1) identify field-level planting history and crop rotation cycles, 2) confirm the area of irregularly shaped fields, and 3) review and resolve conflicts in reported data on segment-specific

portions of the JAS questionnaire. Additionally, the statisticians can view other information layers in CropScape, such as major road networks and the Crop Frequency Layers for major commodities to guide the manual imputation process.

Number of Farms

There are regions within the U.S. where pre-screening of small farming operations is time-consuming, expensive, and subject to misclassification. A procedure was developed to use the CDL to obtain land cover statistics within the JAS segments to more accurately identify potential land use, prior to survey pre-screening activities. The specific goal was to improve the official estimate for number of farms at the state and national-level for both the JAS and quinquennial US Census of Agriculture. The most recent five years of the CDL are used to categorize the land in each JAS segment and estimate percentages of several predetermined categories, such as percent cultivation, impervious surface, corn, soybean and pasture calculated at the JAS segment level. These percentages are used as covariates when modelling the probability that a record represents a farm and ultimately adjusting the weights to obtain the estimate of the number of farms for the JAS and Census of Agriculture.

Natural Disasters

NASS utilizes the CDL and Cultivated Layer to monitor and assess affected cropland and livestock in the U.S. caused by hurricanes, regional flooding, and fire events. This capability is now possible due to a refined methodology utilizing freely available geospatial data products, which include the newly launched and freely available Synthetic Aperture Radar (SAR) Sentinel-1 data, optical satellite imagery, and supplemental geospatial hurricane and fire location data. During disaster response, the confidential in-season CDL data are used to provide timely crop acreage estimates of impacted land to NASS stakeholders. The non-confidential (previous year) CDL data are used to provide crop acreage estimates that can be released within USDA and to the public on the NASS Disaster Analysis website (USDA/NASS Disaster Analysis 2018).

Specifically for hurricanes and heavy-precipitation events, NASS utilizes freely available Copernicus Sentinel-1 SAR data to conduct operational flood mapping of agricultural land in near real-time (Copernicus, 2018). NASS produces multiple binary inundation raster products that are then overlaid with agricultural information from the CDL and Cultivated Layer. From this analysis, NASS is able to estimate the extent of cropland, pasture/hay, and specific crop types that are inundated. This operational flood monitoring process provides accurate results based on independent manually-derived

ground reference data (above 95% producer's accuracy) (Boryan et al., 2018), and has been operationally used for Hurricanes Harvey and Irma in 2017 and Hurricanes Florence and Michael in 2018. During fire events, daily optical imagery, such as MODIS data, and active fire location geospatial data, such as Cal Fire (Cal Fire, 2019) and the USDA Forest Service Remote Sensing Applications Center (USDA/Forest Service Remote Sensing Applications Center, 2019), are combined with the CDL and Cultivated Layer to identify agricultural areas within an active fire perimeter. This methodology was implemented for agricultural areas impacted by the 2017 northern California wildfires 2017 and a 2018 Oregon substation fire.

3. Current and Future Efforts

The CDL and its derivatives have been used to inform NASS processes as in the yield modelling and the JAS imputation. NASS is now moving to integrate its survey and administrative data with the remotely sensed information. Two examples will be discussed here. First, as discussed above, official estimates of the acreages planted to various crops are published each year. Currently, initial estimates are obtained from survey data, FSA data, and the CDL. These estimates are combined using expert opinion during the Agricultural Statistics Board process. However, the information from all of these sources can potentially be geospatially integrated to provide an improved estimate. Efforts are now underway to evaluate this approach for the state of Iowa with the goal of providing estimates for all states.

Currently, the responses to the Census of Agriculture questionnaire provide the foundation for all official Census estimates. For the 2022 and subsequent censuses, NASS is moving rapidly to make full use of the remotely sensed information and administrative data. In addition to using current administrative data to reduce respondent burden, remotely sensed data will be used to inform estimates. Using the CDL with the FSA administrative data as the foundation with the questionnaire being used to fill in gaps in information is being explored for possible implementation for the 2027 Census.

4. Discussion and Conclusion

The numbers of uses and derivative products of the CDL have increased over its 11-year history. Most recently NASS has begun developing maps displaying the extent of the impact of natural disasters, such as floods and fires, as well as quantifying the extent of the potential damage by crop. The resulting products have been used, within NASS, within the US Department of Agriculture, and more broadly to inform relief efforts. Yet, other opportunities for using the CDL certainly exist, especially as NASS moves to integrate the CDL information with NASS survey and administrative data. By combining

these different sources, the potential exists to both reduce respondent burden and to produce more precise official estimates. A longer-range goal is to explore the potential of having remotely sensed data as the primary foundation for analysis of Census of Agriculture data with the Census questionnaire and administrative data providing ground-truthing and supplemental data. NASS continues to look toward the future for remote sensing and geospatial technologies that can enhance agricultural statistics and be integrated into Agency operations.

References

1. Allen, J. D. and Hanuschak, G. (1988). The remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987. U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08. https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/GIS_Reports/The Remote Sensing Applications Program of the National Agricultural Statistics Service 1980-1987.pdf.
2. Boryan C. and Yang, Z. (2017). Integration of the Cropland Data Layer Based Automatic Stratification Method into the Traditional Area Frame Construction Process. Survey Research Methods, 11 (3), 289-306. https://ojs.ub.uni-konstanz.de/srm/article/view/6725.
3. Boryan C., Yang, Z., and Di. L. (2012). Deriving 2011 cultivated land cover data sets using USDA National Agricultural Statistics Service historic Cropland Data Layers, Proc. of IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany. https://ieeexplore.ieee.org/document/6352699.
4. Boryan, C., Yang, Z., Di, L., and Hunt, K. (2014b). A New Automatic Stratification Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9 (11), 4317-4327. https://ieeexplore.ieee.org/document/6837419.
5. Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer Program. Geocarto International, 26 (5), 341-358. https://www.tandfonline.com/doi/abs/10.1080/10106049.2011.562309.
6. Boryan, C., Yang, Z., Sandborn, A., Willis, P., and Haack, B. (2018). Operational Agricultural Flood Monitoring with Sentinel-1 Synthetic Aperture Radar. Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), 2018 IEEE International, Valencia, Spain, July 22 – 27, 2018. https://ieeexplore.ieee.org/document/8519458.
7. Boryan, C., Yang, Z., and Willis, P. (2014a). US geospatial crop frequency data layers. Proceedings of the Third International Conference on Agro-

geoinformatics, August 11-14 2014, Beijing, China. https://ieeexplore.ieee.org/document/6910657.
8. Cal Fire, (2019). http://www.fire.ca.gov/general/firemaps. Ca.gov.
9. Copernicus, (2018). Copernicus Services Data Hub. https://cophub.copernicus.eu/. European Space Agency.
10. Cotter, J. Davies, C., Nealon, J., and Roberts, R. (2010). Area Frame Design for Agricultural Surveys in Agricultural Survey Methods (eds R. Benedetti, M. Bee, G. Espa and F. Piersimoni), John Wiley & Sons, TD, Chichester, UK. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470665480.ch11.
11. Han, W., Yang, Z., Di, L., and Mueller, R. (2012). CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Computers and Electronics in Agriculture, 84, 111–123. https://doi.org/10.1016/j.compag.2012.03.005.
12. Johnson, D.M. (2012). A 2010 map estimate of annually tilled cropland within the conterminous United States, Agricultural Systems, 114, 95-105. https://doi.org/10.1016/j.agsy.2012.08.004.
13. Johnson, D.M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment, 99, 341–356. https://doi.org/10.1016/j.rse.2013.10.027.
14. Johnson, D.M. (2016). A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products, International Journal of Applied Earth Observation and Geoinformation, 52, 65–81. https://doi.org/10.1016/j.jag.2016.05.010.
15. Lark, T.J., Mueller, R.M., Johnson, D.M., and Gibbs, H.K. (2017). Measuring land-use and land-cover change using the U.S. Department of Agriculture's Cropland Data Layer: Cautions and recommendations, International Journal of Applied Earth Observation and Geoformation, 62, 224-235. https://doi.org/10.1016/j.jag.2017.06.007
16. Mueller, R and Harris, J. (2013). Reported Uses of CropScape and the National Cropland Data Layer Program. ICAS VI, Rio de Janeiro, Brazil. https://repository.uantwerpen.be/docman/irua/78a9f0/130375.pdf.
17. USDA/Forest Service Remote Sensing Applications Center, (2019). https://fsapps.nwcg.gov/afm/gisdata.php. USDA-Forest Service, Salt Lake City, UT.
18. USDA NASS Disaster Analysis, (2018). https://www.nass.usda.gov/Research_and_Science/Disaster-Analysis/index.php. USDA-NASS, Washington, DC.

19. USDA NASS Cropland Data Layer, (2018). https://nassgeodata.gmu.edu/CropScape/. USDA-NASS, Washington, DC.
20. USDA NASS, Understanding Statistics, (2018). https://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Foundation_of_Estimates/Area_Frame_Samples/. USDA-NASS, Washington, DC

Modelling long memory stochastic volatility of crude palm oil price

Arifah Bahar[1,2,*], Shaymaa Mustafa[2], Kho Chia Chen, Haliza Abd Rahman[1], Nur Arina BazilahAziz[1,2], Zaitul Marlizawati Zainuddin[1,2], Zainal Abdul Aziz[1,2]
[1]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
[2] UTM Centre for Industrial and Applied Mathematics (UTM-CIAM), Ibnu Sina Institute for Scientific and Industrial Research (ISI-SIR) Universiti Teknologi Malaysia

Abstract
Crude palm oil (CPO) is one of the largest commodity for export in Malaysia. Forecasting its future price plays an important role in planning various investment and business activities for optimal resource allocation. However, this task is not a trivial one as it possesses long memory stochastic volatility. This study will handle this issue by using fractional Ornstein-Uhlenbeck (fOU) process to describe the time series of the CPO price so that the degree of its persistency can be estimated. Model will be constructed with long memory stochastic volatility (LMSV) based on 12 years daily CPO prices. The least square estimator (LSE) and quadratic generalised variations (QGV) method will be used to estimate the drift and diffusion coefficient of the volatility process respectively. The long memory parameter is estimated by the detrended fluctuation analysis (DFA) method. Small values of root mean square errors (RMSE) for the model and mean absolute percentage errors (MAPE) indicate a good forecast for future CPO price.

Keywords
Crude palm oil; fractional Ornstein-Uhlenbeck; long memory stochastic volatility; least square estimator; quadratic generalised variations.

1. Introduction

An accurate forecasting on crude palm oil (CPO) prices is one of the most important economic indicators in the world. All the participants in this industry including the producers, marketers, Policy-makers, consumers and financial participants monitor the CPO price behavior (Charles and Darné, 2014). There are more than 150 countries around the world consuming it (Rahim et al., 2018). The raising of the demand of CPO in last decades, especially in developing countries led to a continuous increase and volatile in the CPO prices (Rahim et al., 2018). This volatility and instability in trajectories and behavior of the prices is considered critical particularly in dealing with uncertainties and risks for the oil palm business (Charles and Darné, 2014).

Malaysia is one of the biggest exporters of CPO in the world and it is the leader of the production of the world palm oil (Arshad and Zainalabidin, 1994).

This industry is important and vital in Malaysia and consequently any change in crude oil prices will affect the country's earning indirectly and will form a risk for all the workers in this industry. Therefore, forecasting the CPO price is important to make ease decision in the immense economic instability event and to plan for various investment and business activities for optimal resource allocation. The high degree of accuracy of the predictions of CPO price is of paramount importance since any decisions will be taken based on this predictions will affect the performance of the true market of this commodity. This task is not an easy one where it includes long memory stochastic volatility (LMSV). For improving the accuracy of CPO price forecasting, it is necessary to monitor the CPO price and to record its growth data for long time period.

There are many stochastic volatility (SV) models have been developed in literature to forecast the CPO price. Arshad and Zainalabidin (1994) examined the ability of CPO future market to predict the forward prices efficiently. They found that the futures price method outperforms the other techniques such as moving average, Box Jenkins, exponential smoothing and econometric in forecasting the forward CPO price. One of the most popular model is the Box and Jenkins model that produces the results for linear time series data (Karia et al., 2013b). Arshad and Ghaffar (1986) forecasted the CPO price by using an univariate autoregressive-integrated moving average (ARIMA) model which is developed by Box Jenkins approach . Ahmad et al. (2014) applied ARIMA model to find the suitable time series that can simulate monthly CPO price in Malaysia. However, their residuals results were not normal and orthogonal. Moreover, ARIMA models can give inaccurate estimations with large sample sizes. Khin et al. (2013) compared between three statistical models (Vector Error Correction Method (VECM), Multivariate Autoregressive Moving Average (MARMA) and ARIMA model) that have been used to forecast the spot palm oil price in Malaysia. Their results demonstrated that MARMA model is the superior in comparison with VECM and ARIMA models. However, the results of the Root Mean Square Percentage Error (RMSPE) demonstrated that the model give high percentage of errors. Omar and Majid (2004) used the historical variances returns of spot and futures price to investigate the relationship between the spot and futures prices of CPO contracts that are traded in the Malaysian Derivatives Exchange.

All the previous studies predict the price of CPO for short memory volatility. The discovery of long memory behavior in the volatility of some financial data was started from the early 1990s. Ding et al. (1993) investigated that there is strong correlation between absolute returns of the daily standard and poor 500 index prices and they were among of the first people who discover this relation. Due to the instability and volatility in CPO price, the stochastics models such as autoregressive conditional heteroskedastic (ARCH), generalized ARCH (GARCH), exponential GARCH (EGARCH) or

standard (short-memory) stochastic volatility models cannot be used to forecast the CPO Price. Breidt et al. (1998) suggested a long memory stochastic volatility (LMSV) model in discrete time to overcome the limitations of the previous models. In LMSV model, the log-volatility is simulated as an autoregressive fractional integrated moving average (ARFIMA) process. The well-defined of LMSV in the mean square sense is one of main advantages that facilitates the establishing the stochastic features of LMSV model. Moreover, the LSMV model has counterparts in models for level series. These models gives their statistical properties to the LSMV model.

Karia et al. (2013a) applied ARFIMA model to solve the nonstationary persistency of the prices of CPO in the long-run data. They conducted a comparison between the ARFIMA over the existing ARIMA model and the results indicates that the ARFIMA model outperformed the existing ARIMA model. Karia et al. (2013b) forecasted the CPO price in Malaysia by using both the artificial neural network (ANN) and adaptive neuro fuzzy inference system (ANFIS). The predictability accuracy of ANN and ANFIS approaches was illustrated in regard with the statistical forecasting approach such as ARFIMA model. Their findings showed that the ANN model gives better results compared to the ANFIS and ARFIMA models. However, both models have a complicated time series characteristics and had relatively more parameters and consequently they need a bigger amounts of data. Karia et al. (2015) selected five different edible oils prices that have long memory behavior to investigate the effect of the over difference on the prices of these oils. They conducted a comparison by using the time series data that recorded with the over difference and long memory behavior between ARIMA and ARFIMA models. Their findings show mixed results for that the forecasting of oil prices for the two models and the existing of over difference seems not to have a significant effect neither ARIMA nor ARFIMA models. They also found that ARFIMA model does not give poor out-sample forecasting. Rahim et al. (2018) used weighted subsethood-based algorithm to generate fuzzy rules of predictions that are embedded in fuzzy time series data. This method is considered as a new approach to forecast the CPO price in order to enhance the accuracy of future prediction. They compared their model with previous models and with numerical results and the outcomes shows an increase in accuracies from the proposed method in predicting CPO price.

In Fact, volatility estimation is considered as a one of the complicated process in econometrics since the volatility can not be observed directly. There are no ideal method neither to simulate volatility nor to collect volatility data. To select the method we should consider many aspects such as financial support, data, expertise and manpower. Chen et al. (2017) evaluated the degree of persistence property of the data by constructing LMSV model. The model is developed by using fractional Ornstein-Uhlenbeck (fOU) process in

financial time series and it is applied for FTSE Bursa Malaysia KLCI over a period of 20 years. In this study, we will develop the LSMV model to forecast the CPO price by using fOU process that have the ability to capture the characteristic observed in the time series of the CPO. This model is useful to estimate the degree of its persistency of CPO time series and to simulate the relationship between the returns and the series volatility. The data will be collected for 12 years daily CPO prices. The drift and diffusion coefficient of the volatility process are estimated by using the least square estimator (LSE) and quadratic generalized variations (QGV) methods respectively. The long memory parameter is estimated by the detrended fluctuation analysis (DFA) method. This study aims to investigate the CPO prices behavior via the LSMV models. The model introduce a probabilistic approach in allowing different volatility states in CPO time series to overcome the excessive persistence problem in the composite linear and nonlinear models.

2. Methodology

The method of parameters estimation of long memory stochastic volatility (LMSV) is discussed in this section. At First, the specification of LMSV model is produced. Then, both the testing methods for the existence of long memory and the estimation methods of parameters on the drift and diffusion coefficient of the volatility process are discussed. Finally we assessed the model performance and the methods of parameters estimation.

2.1 LMSV Model specification

Suppose that there is a complete probability space. Then the LSMV model can be written in the state space form as follows:

$$X_t = m\sigma_t\varepsilon_t \qquad \varepsilon_t \sim N(0,\sigma_n^{\,2}) \qquad \sigma_t = \exp(\frac{Y_1}{2}) \tag{1}$$

$$dY_1 = -\lambda Y_1 dt + \beta d B_t^{\,H} \tag{2}$$

Where $\{X_t, t \geq 0\}$ is the returns series at time t and m is constant coefficient. The ε_t is mutually independent Gaussian white noise process. $\sigma_n^{\,2}$ is the variance of ε_t, $\{B_t^H, t \geq 0\}$ is the fractional Ornstein-Unlenbeck process (fOU) which is assumed to be followed by the volatility process $\{Y_t, t \geq 0\}$ in the model, λ is the drift, β is the volatility of the volatility, the fractional Brownian motion $\{B_t^H\}$ is with Hurst index $H \in (0,1)$ and has stationary increments which yields

$$Var[B_t^H - B_s^H] = |t-s|^{2H}. \quad (3)$$

The relation in Equation (3) defines its covariance structure as follows:

$$Cov(B_t^H, B_s^H) = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}), \quad (4)$$

Thus, for $H \neq \frac{1}{2}$, the fOU process is Gaussian and ergodic but it is neither Semimartingal nor Markovian. For $H > \frac{1}{2}$, the fOU process shows a long memory property. From Equation(2) we get:

$$Y_t = y_0 e^{-\lambda t} - \beta \int_0^t e^{-\lambda(t-s)} dB_s^H, \quad t > 0, \quad X_0 = x_0. \quad (5)$$

Let $y_0 \in L^0(\Omega)$, $-\infty \leq a < \infty$ and $\lambda, \beta > 0$. Then for all $s \in \Omega$ the integration $\int_a^t e^{\lambda s} dB_s^H$, $t > a$, exist as a Riemann-Stieltjes pathwise integral which is continuous in t, and the unique almost surely continuous solution of Equation (5) is

$$Z_t^{H,y_0} = e^{-\lambda t}(y_0 + \beta \int_0^t e^{\lambda s} dB_s^H), \quad t \geq 0, \quad (6)$$

Particularly, the restriction to $t \geq 0$ of the following almost surely continuous process

$$Z_t^H = \beta \int_0^t e^{-\lambda(t-s)} dB_s^H, \quad t \in \mathbf{R}, \quad (7)$$

which can solve Equation (5) with initial condition $y_0 = Z_0^H$. The Gaussian process $(Z_t^H)_{t \in \mathbf{R}}$ is stationary since it follows the stationarity of the increments of $\{B_t^H\}$. Additionally, for each initial condition $y_0 \in L^0(\Omega)$,

$$Z_t^H - Z_t^{H,x_0} = e^{-\lambda t}(Z_0^H - y_0) \to 0, \text{ as } t \to \infty. \quad (8)$$

This implies that every stationary solution of Equation (5) has the same distribution as $(Z_t^H)_{t \geq 0}$. The $(Z_t^{H,x_0})_{t \geq 0}$ is defined as a FOU process with initial condition y_0 and $(Z_t^H)_{t \in \mathbf{R}}$ is the stationary FOU process that is ergodic and for $H > \frac{1}{2}$ it exhibits as long range dependence. Since the values of H and β can be estimated without having the λ value, then the estimation for the unknown parameter $\theta = (\lambda, \beta, H)$ will be carried out in several steps. The covariance theorem of the fOU is shown in Chen et al. (2017).

2.2 Long memory estimation by using Detrended Fluctuation Analysis

The detrended fluctuation analysis (DFA) is developed by Peng et al. (1994) to investigate the correlation of long range power law of DNA nucleotides. The DFA analysis can estimates the scaling H of the series in nonstationary case and it can eliminate the spurious detection of long-range dependence. Additionaly, DFA implements the dispersion measurements that take the squared fluctuations around the time series trend. Conducting the DFA on the sub period avoids the effect of nonstationaries . However, DFA is still can be used to examine the long and short range correlation in both stationary and non-stationary series.

The first step of DFA is to integrate the time series $y(k)$ with N samples in purpose of analysation and this series will be divided into n non-overlapping segments. Next, for each segment, the local trend $y_n(k)$ will be calculated by using the least-square regression. The $y(k)$ series is detrended by subtracting the $y_n(k)$ in each segment. Finally, the root-mean-square fluctuation can be calculated as follows:

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_{n(k)}]^2} \tag{9}$$

The pervious process will be repeated at a range of different window sizes n for the whole signal. To test the self similarity or fractal properties, the log-log graph of n against $F(n)$ is created. If the plot is linear, then the power law scaling is exist. Then, the slop α of the plot line will be employed to characterize the fluctuations of the series as follows:

$$F(n) = Cn^{\alpha} \tag{10}$$

$$\log F(n) = \alpha \log n + \log C \tag{11}$$

where C is constant and α is the correlation value and represents the Hurst exponent H estimation. If $0 < \alpha < 1$ then it will have the properties of fractional Brownian motion. The α values that can be used to explain the series of self-correlations are summarised by Chen et al. (2017).

Bardet and Kammoun (2008) showed the DFA asymptotic properties for the fractional Gaussian noise. These properties can be used for long range dependent processes in stationary case. The asymptotic behaviour of the DFA for the FGN can be written as:

$$F(n) = c(\sigma, H).n^{H}, \tag{12}$$

where c is a positive function. Bardet and Kammoun (2008) also investigated the convergence of long range dependence parameter estimator. They found it has a reasonable convergence rate in the semi-

parametric frame of long memory stationary process while in many cases of trended long range dependent process, the estimator is not converge. Løvsletten (2017) explored the detrended fluctuation analysis consistency where $F(n)\ n^H$ for the stationary and non-stationary stochastic process with $0 < H < 1$ and $1 < H < 2$ respectively.

2.3 Volatility process drift estimation using least square estimator

The drift parameter is estimated by using the least square estimator. We assume that Equation (5) is derived by fractional Brownian motion $\{B_t^H\}$ with $H \geq \frac{1}{2}$. Then the solution of Equation(5) is given as follows

$$Y_t = \beta \int_0^t e^{-\lambda(t-s)} dB_s^H, \tag{13}$$

where $y_0 = 0$ and $\lambda > 0$. To estimate the value of λ, the following least square estimator will be used (Hu and Nualart, 2010):

$$\lambda_T = \lambda - \beta \frac{\int_0^T Y_t dB_t^H}{\int_0^T Y_t^2 dt}, \tag{14}$$

where $\int_0^T Y_t dB_t^H$ is a divergence-type integral (Biagini et al., 2008, Duncan et al., 2000). Another expression for estimating λ_T is :

$$\lambda_T = \frac{Y_T^2}{2\int_0^T Y_t^2 dt} + \beta^2 \frac{\int_0^T \int_0^t \xi^{2H-2} e^{-\lambda\xi} d\xi dt}{\int_0^T Y_t^2 dt}, \tag{15}$$

Consequently, if

$$\lim_{t\to\infty} \frac{1}{t} \int_0^T Y_t^2 dt = \lim_{t\to\infty} Var(Y_t) = \frac{\beta^2 \Gamma(2H+1)}{2\lambda^{2H}} \tag{16}$$

Then λ can be found as follows:

$$\hat{\lambda}_N = \left(\frac{2\hat{\varphi}}{\beta_N^2 \Gamma(2H_N + 1)} \right)^{-\frac{1}{2H}} \quad (17)$$

Where φ is the empirical moment of order 2 that can be given as :

$$\hat{\varphi} = \frac{1}{N} \sum_{n=1}^{N} X_n^2 \quad (18)$$

The asymptotic distribution of Least square estimator is given by (Chen et al., 2017)

2.4 The estimation of Diffusion Coefficient on the Volatility Process by using Quadratic Generalized Variations

The quadratic generalized variations (QGV) can be employed to estimate the Diffusion coefficient β in the discretely fOU process. The Hurst exponent H and the Diffusion coefficient β can be estimated simultaneously. Let $a = (a_0, ..., a_k)$ be a discrete filter of K+1, $K \in \mathbf{N}$, with order $L \geq 1$, $K \geq L$. Then

$$\sum_{k=0}^{K} a_k k^{\ell} = 0 \quad \text{for} \quad 0 \leq \ell \leq L-1 \quad \text{and} \quad \sum_{k=0}^{K} a_k k^{L} \neq 0 \quad (19)$$

which can be normalised as $\sum_{k=0}^{K} (-1)^{1-k} a_k = 1$. The filter a is expanded to a^2 by using the following relation:

$$a_k^2 = \begin{cases} a_k & if \quad k = 2k \\ 0 & otherwise \end{cases} \quad \text{for} \quad 0 \leq k \leq 2K \quad (20)$$

Since $\sum_{k=0}^{2K} {a_k}^2 k^r = 2^r \sum_{k=0}^{K} k^r a_k$, then a^2 and a produce same filtration. The QGV associated to the filter a is presented as (Istas and Lang, 1997):

$$V_{N,a} = \sum_{i=0}^{N-K} \left(\sum_{k=0}^{K} a_k X_{i+k} \right)^2. \quad (21)$$

If we denote

$$\rho_H^{a^m, a^n}(i) = \frac{\sum_{k=0}^{mK} \sum_{\ell=0}^{nK} a_k^m a_\ell^n \left| mk - n\ell + i \right|^{2H}}{(mn)^H \sum_{k,\ell} a_k a_\ell \left| k - \ell \right|^{2H}}, \quad (22)$$

then, the values of H and β can be estimated as follows:

$$H_N = \frac{1}{2}\log_2 \frac{V_{N,a^2}}{V_{N,a}}, \tag{23}$$

and

$$\beta_N = \left(-2 \frac{V_{N,a}}{\sum_{k,\ell} a_k a_\ell \left|k-\ell\right|^{2H_N} \Delta_N^{2H_N}} \right). \tag{24}$$

The asymptotic distribution of QGV is given by (Chen et al., 2017).

2.5 LMSV model simulation

To simulate the LMSV process, , the Monte Carlo simulation method was implemented. The parameter $\theta = (\lambda, \beta, H)$ is now can be estimated and it is based on the real data . As produced by Euler Maruyama discretization in Equations (1) and (2), The LMSV process is defined as:

$$\Delta X_i = X_{i+1} - X_i = k e^{Y_i/2} \varepsilon_i \tag{25}$$

$$\Delta Y_i = Y_{i+1} - Y_i = -\lambda Y_i \Delta t + \beta (B^H_{t_{i+1}} - B^H_{t_i}) \tag{26}$$

To illustrate the numerical process of LSMV model on Crude palm oil data, we should follow the following algorithm

Step 1: Obtain the fractional Brownian motion by Generating the stationary fractional Gaussian noise via fast Fourier transform. The fractional Brownian motion is defined as partial sum of fractional Gaussian noise.

Step 2: Use Euler- Maruyama approach to simulate the process $Y(.)$ as presented by Equation (26) for different values of λ, β and H. the simulation will be conducted for length $\Delta t = 1$ of samples particles , $nT = 2\text{^}9.$

Step 3: Perform the simulation for a sample path of p=100 and take the average of each point of the path.

Step 4: Generate Gaussian white noise $Y(.)$ and then $X(.)$ processes are simulated by using $Y(.)$ result of for different values of k with assumption that $k \le 1$.

Step 5: Calculate the root mean square error (RMSE) between the estimated returns $X(.)$ and the empirical returns (log returns).

3. Results

3.1 Historical Prices of Crude Palm Oil (CPO)

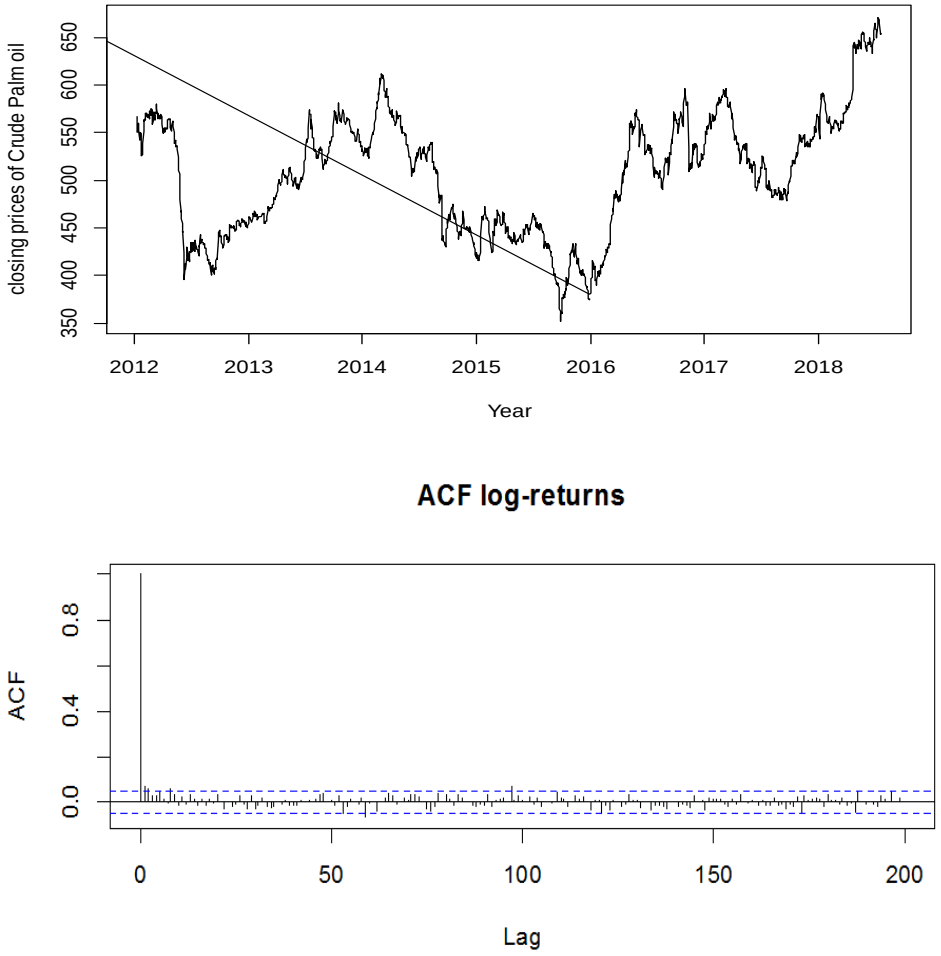


Figure 1 CPO Prices (1 June 2012 until 31 May 2018) and ACF

Table 1: Descriptive statistics of CPO closing prices

Mean	506.20
Median	514.6
Standard deviation	62.84
Skewness	0.0394
Kurtosis	2.418
Coeff of Variation	0.124

3.2 Data Transformation

The returns is defined as :

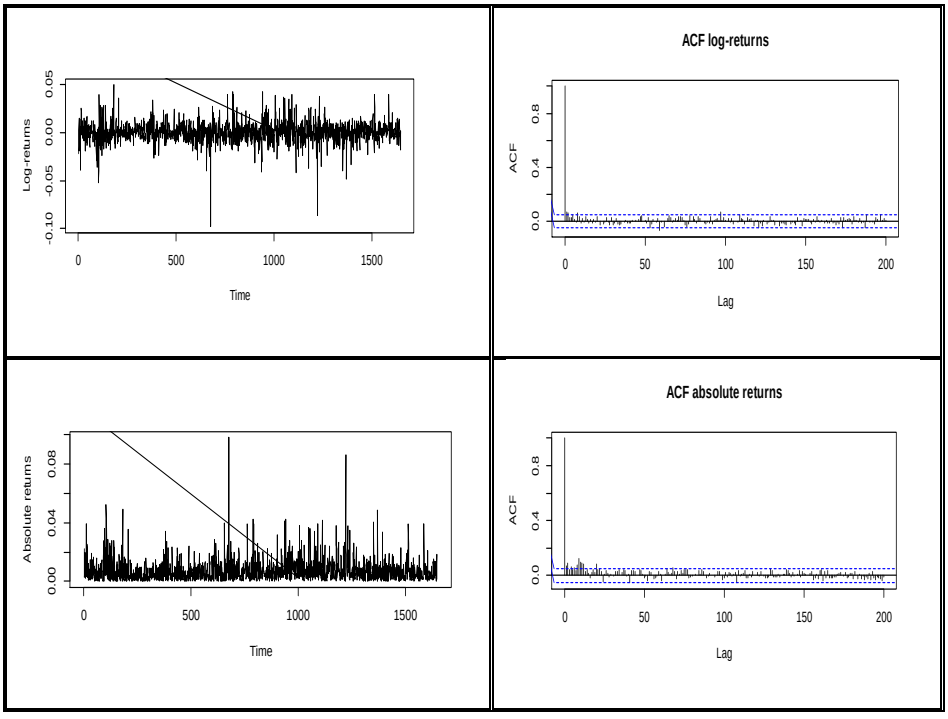
$$X_t = \log(S_t) - \log(S_{t-1}) = \log(S_t / S_{t-1}) \quad (27)$$

Where $\{X_t, t > 0\}$ is the returns and $\{S_t, t > 0\}$ is the closing CPO prices. Relatively, the proxies of volatility are represented by the absolute returns $|X_t|$ and squared of returns X_t^2,

Table 2: Descriptive statistics for the series of X_t, $|X_t|$ and X_t^2

Data	Mean	Median	Standard deviation	Skewness	Kurtosis	CV
Closing Prices of Crude PalmOil	506.1958	514.6	62.8421	0.03935	2.41425	0.12414
Log Return	8.625×10^{-5}	-0.00018	0.01082	-0.62033	11.85372	125.4492
Absolute Return	0.00737	0.00518	0.00792	3.12662	23.29788	1.07462
Squared Return	0.00012	2.683e-5	0.00039	14.62716	311.1487	3.25

3.3 Returns and Volatilities



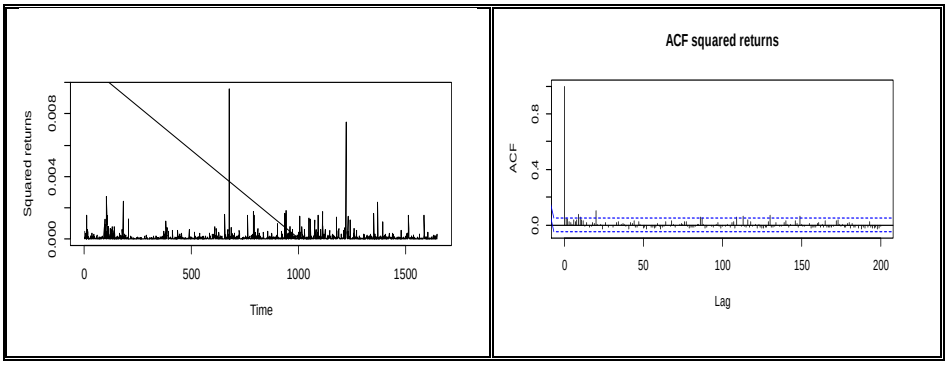
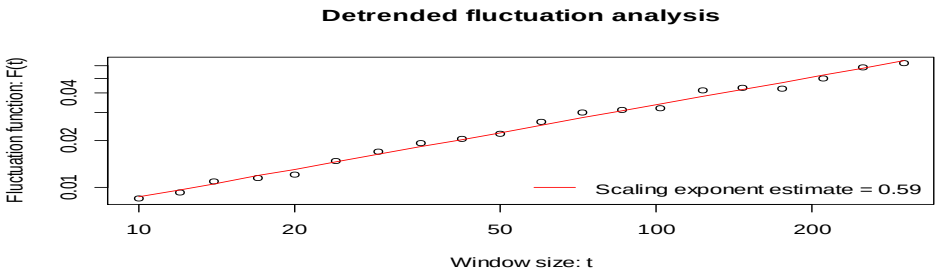


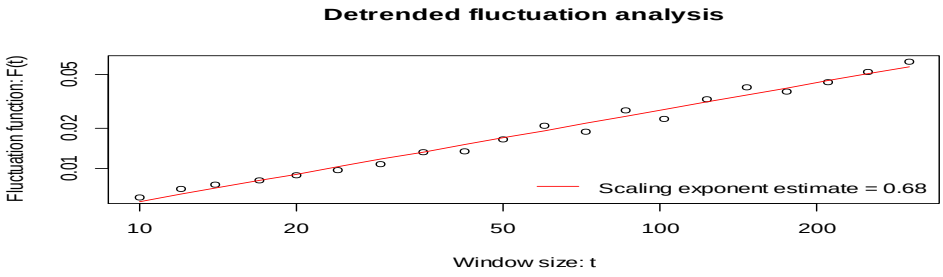
Figure 2: Volatilities and ACF for X_t, $|X_t|$ and X_t^2.

3.4 Parameters Estimation of the LMSV

X_t,



$|X_t|$



X_t^2

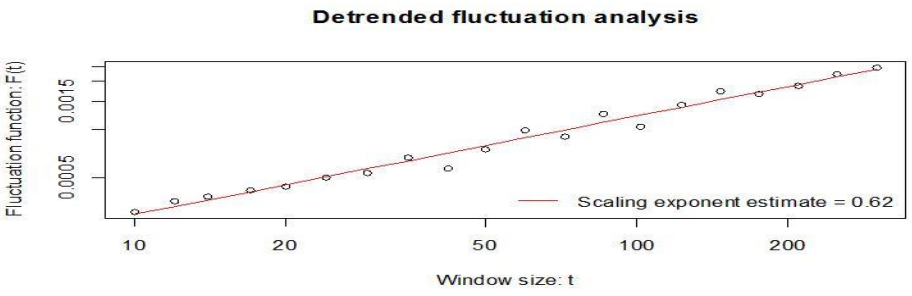


Figure 3: DFA analysis for Long Memory Detection

Table 3: Parameters estimation of λ and β with known H and RMSE between simulated fOU process and data

Data	(LSE)	(QGV)	(DFA)	RMSE
Absolute Return	0.49845	0.01435	0.68367	0.001
Squared Return	0.87036	0.000664	0.6176	0.000092

The value of the drift parameter indicates the volatility process was ergodic where $\lambda > 0$. The diffusion coefficient of the volatilities had a quite small value, implying that the fluctuation of CPO prices was not very significance along the period. The RMSE of between the X_t^2 and its simulated fOU process was smaller than the $|X_t|$.This indicates that squared returns could be more suitable to be chosen as the proxy of volatility for the estimation of the LMSV model.

3.5 Numerical Illustrations of LMSV model

Table 4: Descriptive statistics of estimated returns, and RMSE between estimated returns and empirical returns from squared return for with p=50

θ	$\lambda = 0.87036, \quad \beta = 0.000664, \quad H = 0.6176$
σ_{X_t}	0.01082
p	50
Mean	-2.1540×10^{-4}
Median	-2.8476×10^{-4}
Standard Deviation	0.0110
Skewness	-0.01084
Kurtosis	3.0254
RMSE	0.0155

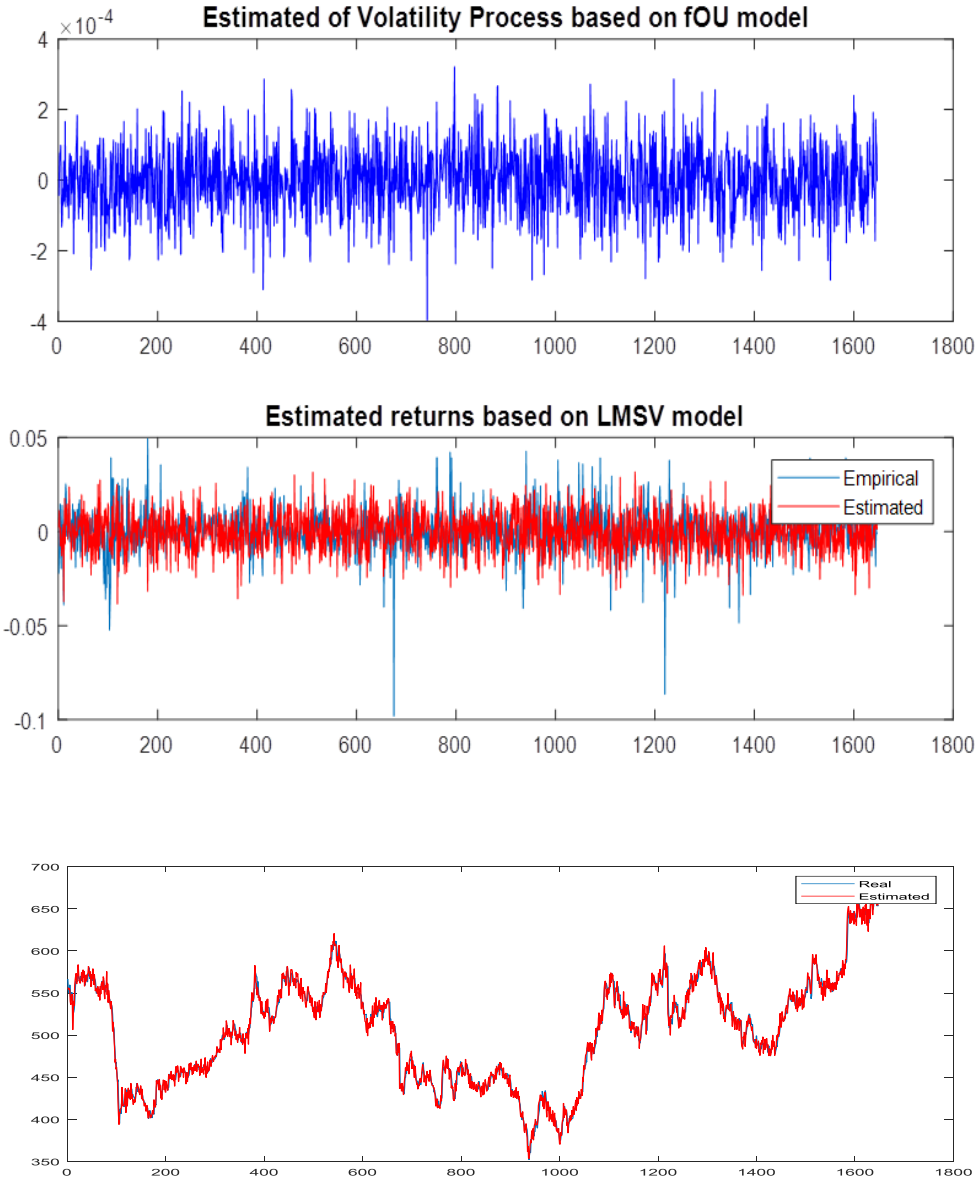


Figure 5: FOU volatility process based on X_t^2, and the comparison between the empirical returns and estimated returns (a) p = 50, k = 0.001

3.6 CPO Forecast Price

14 Days Forecast with MAPE = 1.05

Date	Forecast
6/1/2018	652.9451
6/4/2018	652.766
6/5/2018	656.5104
6/6/2018	650.6179
6/7/2018	651.445
6/8/2018	638.5913
6/11/2018	630.5371
6/12/2018	632.9981
13/6/2018	632.7982
14/6/2018	641.9716
15/6/2018	647.309
18/6/2018	633.0494
19/6/2018	634.6982
20/6/2018	649.5477

4. Conclusion

The framework for handling LMSV had been applied for detection of long memory process of the CPO time series. This study presents the results of the estimated volatility process based on the proxies of volatilities, where the parameters of the LMSV model had been correspondingly estimated. Procedures have been established to construct the LMSV model and estimation methods suitable to explain the CPO market tendency in Malaysia with small errors.

References

1. Ahmad, M. H., Ping, P. Y. & Mahamed, n. 2014. Volatility modelling and forecasting of Malaysian crude palm oil prices. Applied Mathematical Sciences, 8, 6159-6169.
2. Arshad, F. & Zainalabidin, M. Price discovery through crude palm oil futures market: An economic evaluation. Proceedings of the 3rd Annual World Business Congress on Capitalising the Potentials of Globalisation-Strategies and Dynamics of Business, 1994. 73-92.
3. Arshad, F. M. & Ghaffar, R. A. 1986. Crude Palm Oil Price Forecasting Box-Jenkins Approach, Universiti Pertanian Malaysia.
4. Bardet, J. & Kammoun, I. 2008. Asymptotic Properties of the Detrended Fluctuation Analysis of Long-Range-Dependent Processes. IEEE Transactions on Information Theory, 54, 2041-2052.

5. Biagini, F., Hu, Y., Øksendal, B. & Zhang, T. 2008. Stochastic calculus for fractional Brownian motion and applications, Springer Science & Business Media.
6. Breidt, F. J., Crato, N. & De Lima, P. 1998. The detection and estimation of long memory in stochastic volatility. Journal of econometrics, 83, 325-348.
7. Charles, A. & Darné, O. 2014. Volatility persistence in crude oil markets. Energy policy, 65, 729-742.
8. Chen, K. C., Bahar, A., Ting, C.-M. & Rahman, H. A. 2017. Modeling and estimation on long memory stochastic volatility for index prices of FTSE Bursa Malaysia KLCI. Malaysian Journal of Fundamental and Applied Sciences, 13, 315-324.
9. Ding, Z., Granger, C. W. J. & Engle, R. F. 1993. A long memory property of stock market returns and a new model. Journal of Empirical Finance, 1, 83-106.
10. Duncan, T. E., Hu, Y. & Pasik-Duncan, B. 2000. Stochastic calculus for fractional Brownian motion I. Theory. SIAM Journal on Control and Optimization, 38, 582-612.
11. HU, Y. & Nualart, D. 2010. Parameter estimation for fractional Ornstein–Uhlenbeck processes. Statistics & Probability Letters, 80, 1030-1038.
12. Istas, J. & Lang, G. 1997. Quadratic variations and estimation of the local Hölder index of a Gaussian process. Annales de l'Institut Henri Poincare (B) Probability and Statistics, 33, 407-436.
13. Karia, A. A., Bujang, I. & Ahmad, I. 2013a. Fractionally integrated ARMA for crude palm oil prices prediction: case of potentially overdifference. Journal of Applied Statistics, 40, 2735-2748.
14. Karia, A. A., Bujang, I. & Ismail, A. 2013b. Forecasting on crude palm oil prices using artificial intelligence approaches. American Journal of Operations Research, 3, 259.
15. Karia, A. A., Hakim, T. A. & Bujang, I. 2015. World edible oil prices prediction: evidence from mix effect of overdifference on Box-Jenkins approach. The Business & Management Review, 7, 279.
16. Khin, A. A., Mohamed, Z., Malarvizhi , C. A. N. & Thambiah, S. 2013. Price forecasting methodology of the Malaysian palm oil market. The International Journal of Applied Economics and Finance, 7, 23-36.
17. Løvsletten, O. 2017. Consistency of detrended fluctuation analysis. Physical Review E, 96, 012141.
18. Omar, A. & Majid, S. 2004. Improving the price forecast of crude palm oil futures using historical return variances. Oil Palm Industry Economic Journal, 4, 23-28.

19. Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. & Goldberger, A. L. 1994. Mosaic organization of DNA nucleotides. Physical review e, 49, 1685.
20. Rahim, N. F., Othman, M., Sokkalingam, R. & Kadir, E. A. 2018. Forecasting Crude Palm Oil Prices Using Fuzzy Rule-Based Time Series Method. IEEE Access, 6, 32216-32224.

On monitoring stock price movements using Markov Switching Model

NUR Iriawan[1], Wiwik Prihartanti[2]

[1] Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia.

[2] Department of Business Administration, Faculty of Social and Political Science, Universitas WR Supratman, 60111 Surabaya, Indonesia.

Abstract

Stock price movements and interest rate fluctuations have always been the attention of investors in the capital market. Lower interest rates will reduce interest in investment instruments in banks. That way, investment instruments in the capital market will be more attractive and much in demand. Optimal modeling that can help monitor stock price movements, therefore, is needed. One of them is the proposed Markov Switching model, which can be coupled with the first step analysis to detect the price change patterns. Then, in turn, the average run length that the price will stay at the same price can be calculated. A combination of the probability switching pattern and the average run length can be used as a basis for determining the predicted price of the stock later. The combination of these two methods, furthermore, will further sharpen the ability of investors to place their investments in a more profitable position.

Keywords

Markov Switching Model, EM Algorithm, Regime model, probability switching pattern, Average Run Length

1. Introduction

The main measure of transaction liquidity is the value of transactions in the regular market. This measurement is always developed to sharpen the criteria for stock liquidity, including measuring the number of trading days and the frequency of transactions into this measure of liquidity. The movement of stock prices is often also used as a way of supporting in seeing transaction liquidity. LQ45 is one index made by the Indonesia Stock Exchange (IDX). This index comes from the calculation of 45 issuers which meet the criteria for assessing liquidity and taking into account market capitalization. The list of shares is periodically updated every six months, namely the period August-January and February-July. Stocks that have a good performance in meeting the LQ45 liquidity criteria will be selected to enter the list of shares in LQ45 replacing stocks whose performance has declined and does not meet the criteria for liquidity. This is due to, among others, the influence of changes in

market volatility on the price of a stock which can suddenly affect the serial changes in stock prices at any time. These changes can occur in three categories, namely changes in mean, variance, or mean and variance simultaneously. Therefore, some good stocks can maintain their position in the LQ45 for a long time, but there are also stocks that have only been in the LQ45 for a few periods.

In this paper, the Markov Switching model (MSwM) coupled with the Expectation Maximization (EM), which is attached by the method of calculating run length, is proposed to be applied to two different long-standing stocks into LQ45, namely Astra Agro Lestari, Tbk (AALI) and Sawit Sumbermas Sarana, Tbk (SSMS). Both shares come from the plantation sub-sector and have been registered in sharia stocks, since December 9, 1997, and December 12, 2013, respectively. Since July 2003, AALI is registered in LQ45 and has been removed in the period February - July 2018. Meanwhile, SSMS is only registered as a member of LQ45 in the period February - July 2015 (Britama, 2019).

These two stocks with a significantly different length of stay in LQ45 were used to show the ability of the MSwM coupled with EM method in capturing patterns of stock price fluctuations with structural break phenomena. Integrated with that, this combined method will also demonstrate their ability to monitor the run length of each structural regime.

2. Methodology

2.1 Markov switching model

The model that contains structural break due to changes in mean and variance simultaneously can be represented in equation (1) ((Hamilton,1996) and (Kim & Nelson, 1999)).

$$z_t = \mu_{s_t} + \varepsilon_{s_t}. \quad (1)$$

This is a normality based model with μ_{s_t} is the mean model of regimes s_t , $s_t \in \{1,2,\dots,K\}$ and $\varepsilon_{s_t} \sim N(0, \sigma_{s_t}^2)$ is the related residual. Compatibility with Markov chains, a regime can be set as a state. Thus, the movement of stock prices from regime *i* to regime *j* is represented in the Markov probability transition matrix as as equation (2).

$$p\{s_t = j \,|s_{t-1} = i, s_{t-2} = k, \dots\} = p\{s_t = j | s_{t-1} = i\} = p_{ij}, \quad (2)$$

Where $p_{ij} \geq 0$ for $i, j = 1,2,\dots,K$ and $\sum_{j=0}^{K} p_{ij} = 1$. This is an MSwM with a certain number of regimes as a Markov process with a finite *K*-state.

When each of the *K* regimes has μ_{s_t} in equation (1) as Autoregressive pattern on order *r*, it is called as the Markov Switching Autoregressive

model written as MSw(K)-AR(r). For r=1, the MSw(K)-AR(1) of equation (1) can be written as

$$z_t = \left(\emptyset_{0s_t} + \emptyset_{1s_t} z_{t-1}\right) + \varepsilon_{s_t}, s_t = 1,2,\dots,K \quad (3)$$

The parameters estimation will estimate an unknown and optimal number of regimes including their parameters simultaneously. However, determining the optimal number of states representing any of the time jumps is a hidden step in the MSwM estimation process. To do this, the ordinary likelihood ratio tests may not be fulfilled, although it has a regular condition for asymptotic x^2 distribution. Persio and Frigo (2016) have tried to use an MLE to estimate the MSwM in a serial financial time modeling by involving the external factors from its own data series for identifying the hidden number of state.

2.2. Expectation Maximization and Average Run Length

The EM algorithm is a method for estimating a model having latent parameters that are not given directly by the data. It works through two stages, namely the Expectation stage and the Maximization stage, iteratively until convergence to estimate parameters in the non-close form likelihood function ((Dempster et. al, 1977) and (Susanto, 2018)).

Model MSw(K)-AR(1) in equation (3) has likelihood function as in equation (4).

$$L(\varepsilon) = \prod_{t=1}^{n} \sum_{s_t=1}^{K} \left(\pi_{s_t} \left(\frac{1}{\sigma_{s_t}^2 \sqrt{2\pi}} \right) exp\left(-\frac{1}{2} \left(\frac{\varepsilon_{ts_t}}{\sigma_{s_t}} \right)^2 \right) \right) \quad (4)$$

where n is the number of data, s_t is the regime number, π_{s_t} is the contribution of regime s_t in the model, $\pi_{s_t} \geq 0$, and $\sum_{s_t=1}^{K} \pi_{s_t} = 1$. EM algorithms will involve regimes identifier as the unobservable latent variables by utilizing a dummy vector $\tau_t = (\tau_{t1}, \tau_{t2}, \dots, \tau_{tK}), t = 1,2,\dots,n$, in estimating MSw(K)-AR(r). In the expectation stage, for the certain t-th data, the value τ_{ts_t} is estimated as ${\pi_{s_t}}/{\sigma_{s_t}} \exp\left(-{1}/{2}\left({\varepsilon_{ts_t}}/{\sigma_{s_t}}\right)^2\right)$ divided by the sum of this value for all K regimes repeatedly. The biggest τ_{ts_t} will represent that the t-th data is more likely to belong to the regime s_t, then set τ_{ts_t} =1 and $\tau_{ts_t} = 0$ for the other s_t. For all n number data, $\tau_{s_t} = \sum_{t=1}^{n} \tau_{s_t}$, and in favor of certain t-th data for all K regimes, $\sum_{s_t-1}^{K} \tau_{ts_t} = 1$. Finally, the estimated proportion π_{s_t} in equation (4) is estimated as $\pi_{s_t} = \frac{\tau_{s_t}}{n}$. In the second step, the maximization stage is carried out, namely for estimating $\hat{\sigma}_{s_t}^2$ by using the estimated τ_{is_t} and τ_{s_t} as $\hat{\sigma}_{s_t}^2 = \frac{1}{\tau_{s_t}} \sum_{t=1}^{n} \tau_{ts_t} \varepsilon_{ts_t}^2$.

This process will be repeated for some different predetermined K regimes implemented to MSw(K). Finally, the most representative models with an optimal number of regimes for the data will be selected using the smallest AIC ((Frühwirth-Schnatter, 2006) and (Chuffart, 2015)).

When the EM is running, the first step analysis in a Markov process can always happen in regime $s_t = k$ at the following (t+1)-th. It can happen when the process for the first time leaving from the regime $s_t = l$, i.e. when τ_{tk}=0 and τ_{tl}=1 is assigned to be $\tau_{(t+1)k} = 1$ and $\tau_{(t+1)l}$=0 (Huang, et. al.,2013). Calculating a run length of regime $s_t = l$, therefore, can be done by counting the length of serial time t, when τ_{tl}=1 before it is set to $\tau_{(t+m)l} = 0$ for $m = 1,2, \dots, n$. When the EM process has met the convergence, each latent variable τ_{ts_t}, $s_t \epsilon \{1,2, \dots, K\}$ will have the series value, 0 and 1 for the length of n. The length of 1's series for regime $s_t = l$ represents its run length during the associated period. Furthermore, the distribution of run length, especially ARL, for each regime can be empirically determined by using its run-length distribution.

3. Results

The increasing price of both shares does not seem linear from day to day transaction. Both trends tend to decline at the end of the recording period which the Dicky Fuller test is said not to reject the null hypothesis. This non-linear price change shows the AALI shareholders scooped up share prices doubled almost 75 times, while for SSMS shareholders only doubled almost 3.5 times.

Simplification for easier analysis, differencing on lag 1 to both serial data shares has to be done. Dicky Fuller's test rejects the null hypothesis to the new both data. The serial plot data coupled with their marginal plot are shown in Figure 1. Their leptokurtic and fat tail would be impossible to be represented as a uni-modal normal distribution. Changes in variance during the transaction make the preliminary testing using the Wolfram Mathworld chi-square in Mathematica software report that there is still a small difference in mean and even variance. It is meant that there is a multi-modal pattern and showing that there are structural changes in the model (Weisstein, 2019).

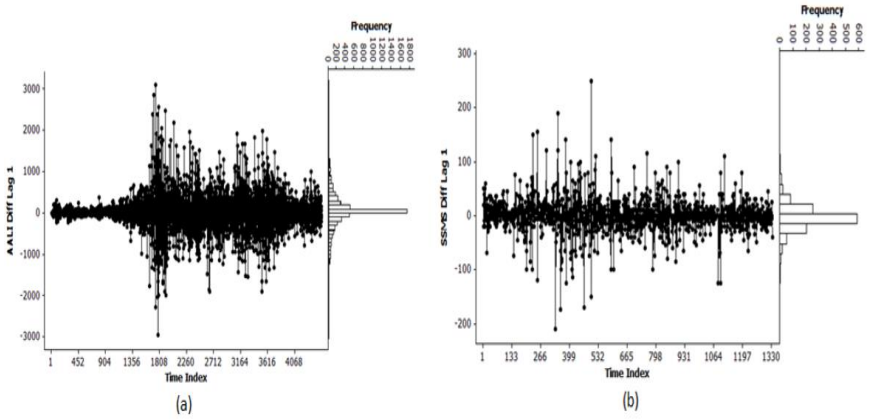


Figure 1 Time series plot and histogram of shares on differencing lag 1: (a) AALI and (b) SSMS

Based on the analysis using MSwM coupled with EM provided by R software employing the overfitting modeling of autoregressive on the predetermining the fixed number of switching regime, the smallest AIC belongs to MSw(2)-AR(3) for AALI and MSw(3)-AR(2) for SSMS. Two regimes in AALI shows two difference behaviors, firstly through the transition probability representing that Regime 1 has a bigger probability to be time-sequentially recurrent, i.e. 0.965119 than Regime 2, that has only 0.939683. Secondly, through the striking difference standardized residual, Regime 1 has almost 8 times bigger standardized residual, 496.3468, than the second regime having 56.37536. The first regime represents a platen model pattern to capture the fat-tail-ness data as shown in the marginal plot Figure 1. (a)., while the second captures the leptokurtic one. The significant difference between these two regimes gives proof that there are multi-modalities as said by the Wolfram Mathworld chi-square which had been done in the preliminary analysis.

The transition probability for SSMS stock price movements, on the other hand, has a lower probability to have time-sequentially recurrent than AALI, those are 0.838113, 0.886688, and 0.878966 for each regime. Changes in the regime during daily transactions in the capital market are more likely to spread in all periods of the transaction (see its Regime 1 plot in Figure 2. (b)), rather than the more observable AALI clustered in a series of ordered and adjacent daily transactions (see its Regime 2 plot in Figure 2. (a)). Regime 2 shows the significant difference in its standardized residual, twice from Regime 3 and 6 times bigger than Regime 1 has. Regime 1, therefore, will be surely able to explain its leptokurtic pattern, Regime 2 explains the platykurtic distribution with fat-tail pattern by employing the greatest standardized residual, and Regime 3 will mostly capture the mesokurtic pattern of data. For investors, investing in SSMS share requires more caution and precision in predicting the changes in this stock prices through the patterns of changes in such regimes than AALI due to its more frequently changing regimes.

The estimation process that produces a good model is also able to provide identification of the first step analysis for each regime change in each daily transaction. Giving values on latent variables during the EM iteration process to converge can be obtained an empirical record of the membership of each daily transaction data as a member of which regime. At the same time, the distribution of the run length of each regime would be known.

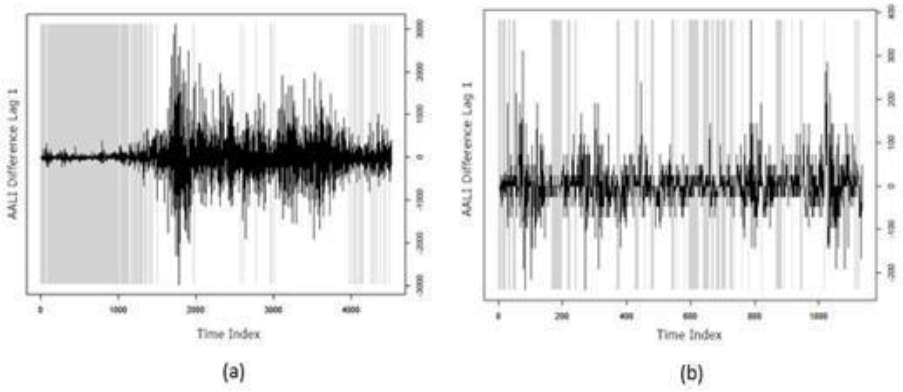


Figure 2 Plot of the Dominated Regime 2 of AALI and Regime 1 of SSMS during the Trading. (a) Regime 2 of AALI and (b) Regime 1 of SSMS.

There is almost the same number of hitting time and ARL of all regime in AALI, those are 227 and 228 times with ARL 12.79 and 12.04, respectively. Figure 3 demonstrates the run length distribution for the second regime of AALI and the first regime of SSMS. The longest time to stay in each regime as a maximum run length (MRL) is 342 run length daily transaction for Regime 2 (see Figure 3. (a)) and 223 for Regime 1. SSMS trading time which is only one-third AALI, on the other hand, has been divided into three structural change regimes. Its first regime is the most frequently visited, which is 110 times with ARL 6.236 and MRL 27 days of transactions (see Figure 3. (b)). Next is followed by Regime 2 with 61 times visited having ARL 4.279 and MRL 31 days of transactions. Finally, the third regime with 52 times visited with ARL 3.577 and MRL 17 days of transactions.

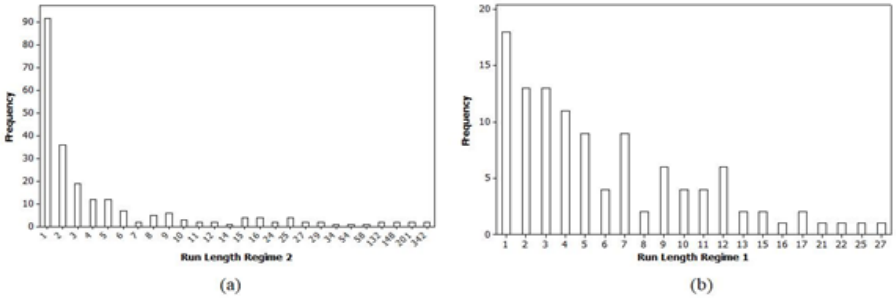


Figure 3 Run Length distribution Regime 1 of AALI and SSMS.
(a) Regime 2 of AALI, and (b) Regime 1 of SSMS
(b)

4. Discussion and Conclusion

The existence of EM as a numerical estimation method in the estimation of the parameters of the MSwM (.) - AR (.) model can be shown to be able to help us in estimating the first step visit of each regime. This first visit can be calculated as a run length for the raised regime. Therefore, ARL can finally be obtained as a result of a side process of EM during the parameter estimation of MSw () - AR (). The results of the run length and ARL recording process can be used as a material for monitoring stock price movements for day-to-day transactions. At the end of the series which AALI had been removed from LQ45 during the last run length, AALI backed to the first regime as the period before it listed in LQ45 at 2003. Investments in stocks such as AALI will tend to be easier to monitor the movements of the share than on stocks such as SSMS, which changes the regime quickly.

The estimation method with EM is always constrained by the number of regimes that must be predetermined first. This is a new challenge if there is a sudden process that will change the number of regimes when this method is applied to monitoring in real observation.

References

1. Britama (2019) Daftar Saham dalam Indeks LQ-45 Mulai dari Agustus 2003 hingga Sekarang, http://britama.com/index.php/2016/03/daftar-saham-dalam-indeks-lq-45-mulai-dari-agustus-2003-hingga-sekarang/, downloaded on April 10, 2019
2. Chuffart, T. (2015) Selection Criteria in Regime Switching Conditional Volatility Models, Econometrics, 2015, 3, 289-316. DOI:10.3390/econometrics3020289.
3. Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B, 39, 1-38. DOI: 10.1.1.133.4884.
4. Frühwirth-Schnatter, S. (2006) Finite Mixture and Markov Switching Models, Springer, New York.
5. Hamilton, J.D. (1996) Specification Testing in Markov-switching Time Series Models. Journal of Econometrics, 70: 127-157.
6. Huang, X., Xu, N., & Bisgaard, S. (2013) A Class of Markov Chain Models for Average Run Length Computations for Autocorrelated Processes, Communications in Statistics - Simulation and Computation, 42:7, 1495-1513, DOI: 10.1080/03610918.2012.667474
7. Kim, C.J & Nelson C.R. (1999) State Space Models with Regime Switching, Classical and Gibbs Sampling Approaches with Applications. Cambridge, MA: MIT Press.

8. Persio, L.D. & Frigo, M. (2016) Gibbs sampling approach to regime switching analysis of financial time series, Journal of Computational and Applied Mathematics, 300, 43–55, DOI: 10.1016/j.cam.2015.12.010.
9. Susanto, I., Iriawan, N., Kuswanto, H., Suhartono, Fithriasari, K., Ulama, B.S.S., Suryaningtyas, W.,& Pravitasari, A.A. (2018) On the Markov Chain Monte Carlo Convergence Diagnostic of Bayesian Finite Mixture Model for Income Distribution, Journal of Physics: Conference Series, 1090(1), 012014, DOI: 10.1088/1742-6596/1090/1/012014.
10. Weisstein, E.W. (2019) "Normal Distribution." From MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/NormalDistribution.html, downloaded on March 20, 2019.

Two-stage stochastic programming approach for oil refinery production planning

Zaitul Marlizawati Zainuddin, Arifah Bahar, Norshela Mohd Noh
Universiti Teknologi Malaysia

Abstract

Two-stage stochastic models are the common model in oil refinery stochastic optimization problem. We propose this model as a framework to maximize the expected profit of production planning for oil refinery industry. Geometric Brownian motion (GBM) is used to describe the uncertainty for price and demand of petroleum products. This model generates the future realization of the price and demand in scenario tree that provides input to the stochastic programming. The prices and demand data was obtained from Malaysia Energy Information Hub (MEIH), Suruhanjaya Tenaga Statistics and was tested for the oil refinery production planning. The result indicates that stochastic model provided a better prediction of oil refinery profit margin.

Keywords

Two-stage stochastic programming; geometric Brownian motion; scenario tree; oil refinery optimization

1. Introduction

Oil refinery is one of the most complex industries due to many different processes with different chemical reactions involved to produce multi finished products. In order to optimize oil refinery profitability, important decisions such as to determine the right amount of crude oil to purchase, amount of products to produce, and amount of raw materials and finished products inventory with the best use of the existing resources are needed. This is more crucial now especially recently in facing uncertainties such as fluctuations of crude oil prices, unpredictable demand of finish products and unstable oil production that cause difficulty to know the company's direction in next year ahead.

Nowadays stochastic programming has become one of the main modeling technique in dealing with refinery planning optimization problem under uncertainty because the deterministic model has a limited capability in handling uncertainties. Stochastic programming can be divided into programming with recourse and probabilistic programming as described in Sahinidis [1] and Khor et al [2]. Two-stage stochastic programming with recourse is the most popular model in the review study of oil refineries planning under uncertainty [3]. The scenario construction for stochastic

parameters and dimensional problem are the main challenges in solving two-stage stochastic programming. Therefore we propose this model as a framework to maximize the expected profit of production planning for oil refinery industry. Leiras et al.[3] conducted a review of oil refineries planning under uncertainties for journal articles from the 90s to 2010 and none of the articles have emphasised on an approach to apply the behaviour of uncertainties and the method of representation for stochastic parameters in oil refinery optimization. Moreover, their study are limited to articles before 2011 thus, we focus on papers on oil refinery from 2011 onwards with the representation of stochastic parameters. Ribas et al. [4] constructed the stochastic parameter scenario tree based on the Brazilian refineries expertise of employees to evaluate between stochastic and robust approach in considering uncertainty in oil refinery activity in 2012 and stochastic model gain highest profit expected value compared to robust approach .Geometric Brownian motion is used to model the end products price uncertainties in developing stochastic linear programming to maximizing profit of biofuel supply chain in North Dakota as studied by Awudu and Zhang [5] in the year 2013. However, scenarios generated by the solution of GBM increased the size of the problem and they applied Benders decomposition to solve large scale mathematical programming. In 2015, Ruoran Chen et al. [6] proposed an approach to apply the behaviour of crude oil prices follow GBM and employed approximate dynamic programming to solve multiperiod multiproduct oil refinery optimization problem at Shandong, China and they are also facing difficulty in a high dimensional problem. Meanwhile, Relvas et al [7] built a scenarios to describe the future realization of uncertainties, oil price and demand with ARIMA and SARIMA model that provides the input to the two-stage stochastic programming in maximizing the expected profit of Portugues oil network. In this study, we improve the formulation of two-stage stochastic programming by Khor et al [2] with constructing the scenario tree based on GBM to generates all possible future realization of the price and demand instead of only considering element in the set of event sequences is highest and lowest value. The prices and demand data from 1990 to 2015 was obtained from Malaysia Energy Information Hub (MEIH), Suruhanjaya Tenaga Statistics and was tested for the oil refinery production planning to maximize oil refinery profitability.

2. Methodology

2.1 Deterministic model

In the deterministic model, crude oil price, finish products sales price and operating cost are constant with mean values from historical data are used and each cost or sales prices are in dollar/barrel. The objective function is to

maximize midterm production planning of oil refinery considered as revenue from products sales minus the raw material cost and operating cost. In this study, the simplified oil refinery operation based on Khor et al. [2] study is used to describe a formulation of the deterministic linear program which the production flowrate variables are in barrel/year.

2.2 Stochastic Model

Two-stage stochastic models are the common model in oil refinery stochastic optimization problem and general formulation for this stochastic approach was defined by Dantzig [8]

$$max C^T x + E\left[Q(x,\xi)\right] s.t\ Ax \le b,\ x \ge 0 \tag{1}$$

Where $Q\left[(x,\xi)\right]$ is the optimal value for the second stage problem,

$$max q^t y\ s.t\ Tx + Wy \le h, y \ge 0 \tag{2}$$

The decision variables are divided to the two stages in the two-stage stochastic model. The first stage variables are decided before the realization of uncertain parameter denoted by x. Matrix A, vector b and vector C are known with certainty. Meanwhile the second stage variables are decided after the realization of uncertain parameter denoted by y and also interpreted as corrective measures against any infeasibility arising due to realization of uncertainty. In the second stage problem, elements $q, T, W\ and\ h$ are viewed as random.

In this study, the framework of two-stage stochastic programming with recourse for discretely distributed random vector ξ is considered, equation (1) and (2) takes on the form

$$max C^T x + \sum_{sc \in SC} p_{sc} q_{sc}^T y_{sc} \text{ subject to } Ax \le b \tag{3}$$

$$Wy_{sc} \le h_{sc} - Tx,\ x \ge 0,\ y_{sc} \ge 0,\ sc \in SC \tag{4}$$

The probability of scenario sc will occur, $P_{sc} (P_{sc} \ge 0, \sum_{sc=1}^{SC} P_{sc} = 1, sc \in SC)$.

2.2.1 Mathematical programming model

Let us define the variables of the two-stage stochastic model in this study. The amount of crude oil type i, P_i and production capacity of process j, x_j are the first stage decision variables. After the prices and demand for finished products uncertainty are revealed, y_i^s production flowrate of product i per

realization of scenario s and the recourse costs are imposed based on corrective action. These corrective actions are associated with $z_{i,s}^{+}$ and $z_{i,s}^{-}$, amount of unsatisfied demand and excess finished products i per realization of scenario S. The recourse costs are denoted as penalty cost incurred due to shortfall and surpluses in oil refinery production planning due to the uncertainty of finished products demand. In oil refinery field, stochastic programming seeks to maximize profit of the first stage decision plus expected profit of the second stage recourse decision.

A new reformulated objective function is the form of

$$Max\, Z_1 = Z + E[Z] - E_{s,demand} \tag{5}$$

consists of summation of maximization of the expected profit (considered as revenue from products sales minus raw material cost and operating cost) and minimization of summation of expected recourse penalty due to shortfall and surpluses in production.

$$Z = -\sum_{i \in I} \lambda_i P_i - \sum_{j \in J} c_j x_j \ , i \in I \ , j \in J \tag{6}$$

denotes the raw material and operating costs, while the expectation objective function with random price coefficient is given by

$$E[Z] = \sum_{i \in I} \sum_{s \in S} p_s c_{i,s} y_{i,s}, i \in I_{price}^{random} \subseteq I, s \in S \tag{7}$$

where p_s is the probability of scenario s for prices uncertainty, $c_{i,s}$ is the coefficient for prices uncertainty that obtained from scenario tree and $y_{i,s}$ is a production flow rate of products type i corresponding to scenario S. The expected recourse penalty due to uncertainties in demand for the second stage cost is given by

$$E_{s,demand} = \sum_{i \in I} \sum_{s \in S} p_s \left(c_i^{+} z_{i,s}^{+} + c_i^{-} z_{i,s}^{-} \right), i \in I_{demand}^{random} \subseteq I, s \in S \tag{8}$$

where p_s is the probability of scenario S for demand uncertainty, $z_{i,s}^{+}$ and $z_{i,s}^{-}$ are an amount of underproduction and overproduction of product type i per realization of scenario S and c_i^{+} and c_i^{-} are penalty costs for shortfall and surplus in production of product type i.The objective function (5) is subject to constraints

$$x_{j,t} \le x_{j,t-1} \quad \forall j \in J \tag{9}$$

$$P_i + \sum_{j \in J} b_{i,j} x_j - y_i^s = 0 \; i \in I, \; j \in J \tag{10}$$

Equation (9) refers to the limitation plant capacity for crude distillation unit (CDU) and catalytic cracker unit. Meanwhile equation (10) refers to mass balances constraints classified as fixed production yields, fixed blends and unrestricted balances.

$$x_i + z_{i,s}^{+} - z_{i,s}^{-} = d_{i,s}, i \in I_{demand}^{random} \subseteq I, s \in S \tag{11}$$

The demand deterministic constraints are replaced by the new constraints to model the number of generated scenarios in the stochastic model as in equation (11) where $d_{i,s}$ is demand of finished products i corresponding to demand scenario s. However, to construct reasonable scenarios with the appropriate probabilities from the historical data is one of the challenges in two-stage stochastic programming. Thus effective scenario construction for stochastic parameters is needed.

2.2.2 Scenario tree

GBM is a continuous time stochastic process where logarithm of the randomly varying quantity follows a random movement where the explicit solution is $S_t = S_0 exp\left[\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right]$ and has property that the log ratio follows normal distribution, $\ln\left(\frac{S_t}{S_{t-1}}\right) \sim N\left(\left[\mu - \frac{\sigma^2}{2}\right]t, \sigma^2 t\right)$. In general to solve mathematical program with uncertain parameters described as continuous distributions is complicated. Thus we discretized the continuous stochastic process in discrete binomial model as proposed by Jarrow and Rudd (9)

$$S_{t+1} = \begin{cases} S_t u \text{ with probability } p \\ S_t d \text{ with probability } 1\text{-}p \end{cases} \tag{12}$$

The first and second moment of binomial steps are matching with the GBM.

$$\begin{aligned} p \ln u + (1-p) \ln d &= \left(\mu - \frac{\sigma^2}{2}\right)\Delta t \\ p(1-p)\left[\ln u - \ln d\right]^2 &= \sigma^2 \Delta t \end{aligned} \tag{13}$$

with the probability $p = \frac{e^{\frac{\sigma^2}{2}\Delta t} - e^{-\sigma\sqrt{\Delta t}}}{e^{\sigma\sqrt{\Delta t}} - e^{-\sigma\sqrt{\Delta t}}}$. If we calculate the limit we obtain $\lim_{\Delta t \to 0} \mathrm{p} = \lim_{\Delta t \to 0} \frac{e^{\frac{\sigma^2}{2}\Delta t} - e^{-\sigma\sqrt{\Delta t}}}{e^{\sigma\sqrt{\Delta t}} - e^{-\sigma\sqrt{\Delta t}}} = \frac{1}{2}$. Thus the probability for next prices and demands upward and downward value is 0.5. Insert $p = 0.5$ to equation (13) and we get $u = e^{\left(\mu - \frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t}}$ $d = e^{\left(\mu - \frac{\sigma^2}{2}\right)\Delta t - \sigma\sqrt{\Delta t}}$ represent as high (H) and low (L) value in the scenario tree. Uncertainties are discretely represented by scenario realization modeled as a scenario tree.

3. Results

In this study, we consider ten uncertain parameters for prices and finished products demand uncertainty. Each parameter takes on two values, high value which is denoted as H and low value which is denoted as L. The probability of each occurring high and low value is 0.5. We obtain 32 scenarios (2^5) with the probability of occurrence of each scenario is 0.03125 by multiplying the probabilities of uncertain parameter in each scenario as presented in the scenario tree in Figure 1. For example, in scenario 1, $\{H, H, H, H, H\}$ is a set of event sequences denotes as high prices for Gasoline, Naphta, Kerosene, Diesel, Fuel oil as well as demand uncertainty scenario tree shown in Figure 2.

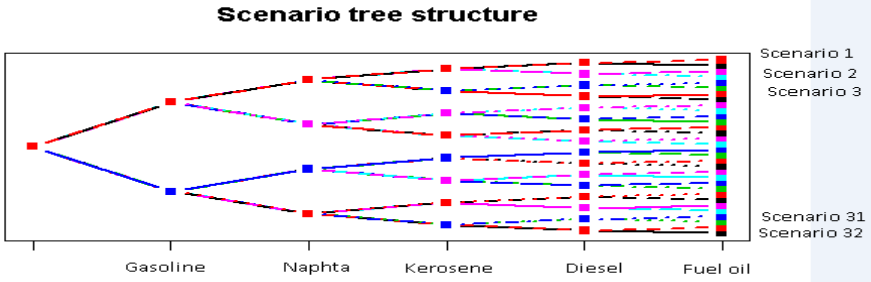


Figure 1. Scenario tree for prices of finish products uncertainty

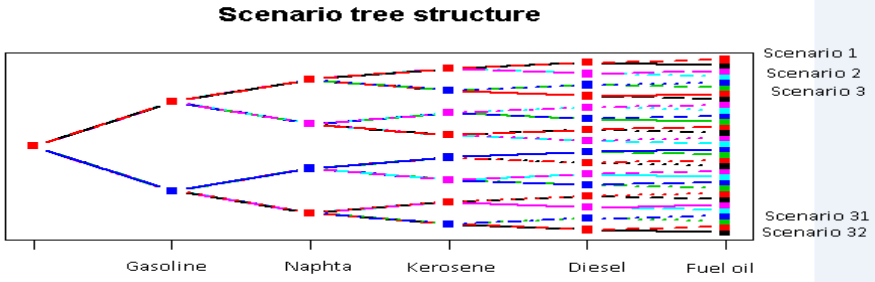


Figure 2. Scenario tree for demand of finish products uncertainty

First and second stage programming model can be summed up to the large linear programming model as shown in equation (3) by taking all possible scenarios for prices and demand uncertainty. The objective function equation is extended to large scale linear programming model and there are 160 new constraints for demand uncertainty. R was used to generate scenarios and construct the scenario tree. Meanwhile, the stochastic programming model was implemented using GAMS.

Table 1. Comparison of oil refinery profit (dollar/year)

Model	Profit/Expected profit
LP	117,096,783
Two stage stochastic linear programming	170,639,512

Value of stochastic solution (VSS) is used to evaluate the uncertainty parameters by calculating the expected profit from two-stage stochastic model over the deterministic model.

$$VSS = 170,639,512 - 117,096,783 = 53,542,729 \tag{14}$$

The VSS result, 53.5 million dollar which the stochastic model gain 45.73% more than deterministic model provided a good solution by including the uncertainty into the model. Thus we conclude that the stochastic model gives a better prediction of oil refinery profit margin as shown in Table1.

4. Discussion and Conclusion

This paper presents a study of stochastic optimization model to find the optimal operation mode of units and stream flows that maximize the oil refinery profit while observing all possible constraints by including the prices and demand uncertainty. This model is called two-stage stochastic programming with recourse and GBM is used to describe the uncertainty by generates the future realization of scenarios with probabilities as input to the stochastic programming. The expected profit from the stochastic model gain 45.73% more than deterministic model provided a good solution by including the uncertainty into the optimization model. Thus we conclude that the stochastic model gives a better prediction of oil refinery profit margin.

Appendix

Set and Indices

I set of material or products i

J set of processes j

S set of scenarios s

P_t amount of crude oil purchase in period t
$b_{i,j}$ stoichiometric coefficient for material i in process j
λ_i unit purchase price of raw material i
c_j operating cost of process j
$x_{j,t}$ production capacity of process j during period t

References

1. N. V Sahinidis, "Optimization under uncertainty : state-of-the-art and opportunities," *Comput. Chem. Eng.*, vol. 28, pp. 971–983, 2004.
2. C. S. Khor, A. Elkamel, K. Ponnambalam, and P. L. Douglas, "Two-stage stochastic programming with fixed recourse via scenario planning with economic and operational risk management for petroleum refinery planning under uncertainty," Chem. Eng. Process. Process Intensif., vol. 47, no. 9–10, pp. 1744–1764, 2008.
3. A. Leiras, G. Ribas, S. Hamacher, and A. Elkamel, "Literature review of oil refineries planning under uncertainty," Int. J. Oil, Gas Coal Technol., vol. 4, no. 2, p. 156, 2011.
4. G. P. Ribas, A. Leiras, and S. Hamacher, "Operational planning of oil refineries under uncertainty Special issue : Applied Stochastic Optimization," IMA J. Manag. Math., pp. 1–16, 2012.
5. I. Awudu and J. Zhang, "Stochastic production planning for a biofuel supply chain under demand and price uncertainties," Appl. Energy, 2013.
6. R. Chen, T. Deng, S. Huang, and R. Qin, "Optimal crude oil procurement under fluctuating price in an oil refinery," Eur. J. Oper. Res., vol. 245, no. 2, pp. 438–445, 2015.
7. C. Lima, S. Relvas, and A. Barbosa-Póvoa, "Stochastic programming approach for the optimal tactical planning of the downstream oil supply chain," Comput. Chem. Eng., 2018.
8. George B. Dantzig, "Linear programming under uncertainty," Manage. Sci., vol. 1, pp. 197–206, 1955.
9. R. Jarrow and A. Rudd, "Approximate option valuation for arbitrary stochastic processes," J. financ. econ., vol. 10, no. 3, pp. 347–369, 1982.

Study on star formation history of nearby galaxies

Tanuka Chattopadhyay
Department of Applied Mathematics, University of Calcutta.

Abstract

Star formation scenario in galaxies of various morphological types is significant in a sense that it characterizes the structure formation in the Universe. Star formation Rate (SFR) is an important index to study the above phenomenon. But direct measurement of SFR_{true} (i.e., *True values of SFR*) is not at all possible as one has to count stars formed per year in a galaxy accurately. In this paper the star formation is investigated by Gaussian Mixture Model Based Clustering technique (GMMBC) to form the clusters of the galaxies (using a large data set of galaxies in the Local Volume (LV)) and the groups are discussed.

Keywords

SFR; Bayesian Analysis; Bayesian LASSO; GMMBC

1. Introduction

During the past few decades, cosmology has progressed a lot, though galaxy formation is far from completeness. It is the great unsolved problem in modern astrophysics. Galaxy formation can be properly delineated if the corresponding star formation history can be explored in greater detail. From current observations it is speculated that galaxy formation and evolution is strongly influenced by environment e.g., star formation activity is reduced in high density environment (?; ?; ?; ?). In addition gas density plays an important role. In fact a quantitative measurement of the relation between SFR and gas density evolution. It gives the clue that how efficiently gas is turned into stars and hence an essential input to simulations models of galaxy formation (?; ?; ?). Another important parameter driving galaxy evolution is the stellar mass of galaxies. There is correlation between SFR and stellar mass in local (?; ?) as well as distant Universe (?; ?; ?). ? have analyzed specific star formation rate (sSFR) for SDSS galaxies of different masses and redshift and they differ in massive and dwarf galaxies. ? have studied the star formation history of dwarf galaxies and traced out SFH as a function of galactocentric radius. He found contrast in SFH for dSph from dwarf transition types (dTr).

Now SFR is the number of stars formed per year. This is not possible to measure SFR directly by counting stars in a galaxy (i.e., SFR_{true}). Hence another aspect is to properly quantify SFR of a representative galaxy. It is important to

note that SFR_{true} values can never be directly deter-mined, i.e., these are hypothetical values. One can use physical models to find model based SFR values as representative of SFR_{true}, denoted by SF R (say). Several authors starting from Kennicut (1998) have used fluxes of electromagnetic waves of various wave lengths related to star formation e.g. Balmer lines (viz. H_α, Kennicut 1998). In this work the author has estimated the integral $\widetilde{SFR}$ in a galaxy using the relation $SFR\left(M_\odot yr^{-1}\right) = 0.945 \times 10^9 F_c(H_\alpha) D^2$, where D is the distance and $F_c(H_\alpha)$is the integral flux in H_α line corrected for Galactic extinction. The SFR estimate, in the above relation, used H_α that corresponds to a time scale of ∼10 Myr. This is the characteristic time scale for the glow of most massive stars in a galaxy. Another estimate has been suggested by ? where the $\widetilde{SFR}$ is found from the integral fluxes of far ultraviolet (FUV) lines that correspond to a time scale of ∼ 100 Myr. This estimate not only considers massive stars but also includes intermediate solar type stars and hence is a more trust worthy estimate. ? have estimated SFR based on the strength of [*OII*] emission line. They found that the SFR is reduced in a cluster environment compared to field galaxies for the similar concentration indices. Tekola et al. (2011) have shown that tidal forces in groups and clusters of galaxies has a strong correlation with SFR at a fixed stellar mass. Wuyts et al. (2011) have suggested another estimate of $\widetilde{SFR}$ on the basis of multiwavelength photometry starting from FUV to infrared and the estimated $SFR_{FUV\ +IR} > 100 M_\odot yr^{-1}$ and out to z ∼3. These spectroscopically (i.e., through flux of energies) estimated SFR (i.e., $\widetilde{SFR}$) are later computed for huge datasets of galaxies by expressing $\widetilde{SFR}$ in terms of magnitudes e.g., m_c^{FUV} (?) or H_α (?), using a mathematical relation, which are photometric parameters hence easily measurable quantities for a large number of galaxies. Finally these physical model based values have been used to compare with our statistically estimated SFR (SFR_{est}) values. But the above wavelengths or magnitudes are not the only indicators of SFR. There are many more parameters e.g neutral hydrogen mass, environment (as mentioned above), which along with the previous ones simultaneously affect the SFR. Hence an ideal model of predicting the SFR of a galaxy is to include as many such parameters as possible through a multivariate set up. The present work aims such challenge.

Moreover previous studies especially used the scatter diagrams of any two parameters at a time while ignoring the effect of others and taking whole data set at a time. These are not suitable in a multivariate set up where all the above mentioned parameters have the simultaneous effect on star formation history of galaxies. Also, there are various types of dwarf galaxies depending on color and surface brightness (?). Therefore while tracing the star formation history of any particular data set of galaxies we have to be aware of the fact that we can depict a convincing theory only if we concentrate on homogeneous galaxy groups, in some way identified. From the above point of view we have first

classified the galaxies into homogeneous groups and then we have studied the star formation properties of those groups separately.

In the present work we have classified the data set of nearby galaxies using GMMBC on the basis of star formation related variables and discussed the groups in terms of their SFr properties.

The rest of this paper is organized into four further sections. Section 2 delineates the data set used and its features and importance for inclusion in our study. Overview and relevant discussions on the statistical methodology used are discussed in Section 3. Section 4 is based on the analyses and interpretations. Finally, the paper concludes with some discussions and provisions for further research pertaining to this area.

2. Overview of The Dataset

We have prepared a suitable dataset of nearby galaxies from an original dataset compiled by ?. Primarily we concentrated on the absolute magnitudes and the parameters associated with the star formation rates e.g., mass of the neutral hydrogen (MHI) etc. In the dataset we have computed $\widetilde{SFR}$ of the galaxies following ? which is more robust (?) given by,

$$\log\left(\widetilde{SFR}\left(M_{\odot}yr^{-1}\right)\right) = 2.78 - 0.4m_{FUV}^{c} + 2logD \quad (1)$$

and

$$m_{FUV}^{c} = m_{FUV} - 1.93(A_B^G + A_B^I) \quad (2)$$

where, D is the distance of a galaxy, m_{FUV}^{c} is the Far-Ultra-Violet apparent magnitude corrected for extinction and A_B^G, A_B^i are galactic extinction in B-band (?) and internal extinction in galaxies in B-band (?) respectively.

We have considered the variables associated with the star formation history, e. g.,

(i) Axial ratio (b/a) measured at the Holmberg isophote,
(ii) Absolute magnitudes in K, B bands (M_K, M_B) and in HI line (M_{21}),
(iii) Logarithm of Mass of the neutral hydrogen (log(*MHI*)),
(iv) Holmberg radius (A_{26}),
(v) Tidal index (**Θ**1),
(vi) Surface brightness (*SBB*) in B-band within Holm-berg radius, and
(vii) Neutral hydrogen mass to K band luminosity ratio (MHI/L_K),

for primary statistical analyses along with $log(SFR)$ (mentioned above and also others) for subsequent comparison.

After several boxplots of the above parameters together with the computation of standard errors we have finally selected the complete data of 596 samples. The remaining parameters which have a considerable amount of missing values and morphological indices (estimated in a subjective way) have not been considered for statistical analyses are: Radial velocity of the galaxy

relative to Local Group Centroid with apex parameters (V_{LG}), Morphological Type (T), Heliocentric Radial Velocity (V_h), Amplitude of rotational velocity (V_m) adjusted to inclination (i), HI line width (W_{50}). They are used to study the properties of the coherent groups of galaxies, once identified. Also Star Formation Efficiency parameter ($SFE = LK/MHI$) has been computed for studying the property of the groups.

Since we have used a reduced data set from an original data set of ?, it requires a checking for completeness. For this purpose we have performed V/V_{max} test. The test was first used by ? for studying space distribution of a complete sample of radio quasars from $3eR$ catalogue. According to this test let F_m be the limiting flux of a survey data. Two colours $V(r) = \frac{4\pi r^3}{3}$ and $V_{max} = \frac{4\pi r_m^3}{3}$ are defined where r is the radial distance and r_m is the maximum distance observed. If the distribution of object is uniform then V =Vmax is uniformly distributed over [0,1], then $< V/V_{max} > = 0.5$. Accordingly we have computed $< V/V_{max} >$ of several significant parameters (e.g. M_{21}, distance D, L_B etc.) and they are all close to ~ 0.3 - 0.6 i.e. the present data set used is complete up to an accuracy of 60% - 80%.

3. Overview of The Statistical Methods

Cluster Analysis (CA) is the art of finding homogeneous (in terms of some parameters) groups that are already present in the data. It is to be noted that according to physical notation we have denoted variables by parameters. We start this section with a coherent review of K-*means* Cluster Analysis; discussing its merits, demerits and why we choose not to use it in our present work. Then we proceed to the extensive discussion on Model-Based Clustering methods (MBC); the method that has been employed in the present study.

3.1 The K-*Means* Cluster Analysis

The K-*Means* Algorithm (?) is one of the simplest unsupervised learning algorithms which tries to partition a given set of points/observations into *K* clusters, such that the points within each of the clusters tend to be near each other in term of some distance measure. Most commonly this distance is taken to be the "Euclidean Distance" with either standardized or non-standardized observations. K-*Means* has been used in many disciplines. It is very much popular for Astronomical datasets as well.

The algorithm aims at minimizing an objective function, known as *Squared-Error Function* given as follows:

$$\mathcal{O}_K = \sum_{k=1}^{K} \sum_{i=1}^{n} \| x_i - c_k \|^2 \quad (3)$$

where, $\| x \| = \sqrt{x^T x}$; the norm of x, $(x_1, \ldots, x_n)$ are the data points and $(c_i, \ldots, c_k)$ are the cluster centers for the k clusters: $\mathcal{G}_1, \ldots, \mathcal{G}_k$. $\sum_{k=1}^{K} \sum_{i=1}^{n} \| x_i - c_k \|^2$ is the Euclidean Distance between a data point x_i and cluster center c_k of the k^{th} cluster $\mathcal{G}_k$. This is basically an indicator of the distance of the n data points from their respective cluster centers. Finally, the algorithm has the following steps:

3.2 The Mixture-Model Based Clustering Technique

K − *Means* clustering is an iterative relocation method which minimizes the *intra-cluster* variance. Model Based Clustering (MBC) is also an iterative method but unlike K −*Means*, it has the provision for variability and structure
of the data. In finite mixture model based clustering, each of the component probability distribution corresponds to a cluster. The usual questions in applied cluster analysis, i.e., choice of appropriate clustering method and determination of number of clusters, can be reformulated as a *Statistical Model Selection* problem where models that differ in number of components and/or in component distribution can be compared. Outliers as well can conveniently modeled by adding one or more component(s) representing a different distribution for the outlying data (?).

As already noted, K − *Means* assumes homogeneous and spherical groups/clusters. This can be viewed as a procedure which approximately maximizes the multivariate normal classification likelihood when the covariance matrix is equal for each of the mixing component probability distributions and is proportional to the identity matrix. On the other hand, MBC can tackle the problem of overlapping and non-spherical clusters having different covariance structures. (?).

Suppose we have the data: $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where $\boldsymbol{x}_i$ is a *d*-dimensional vector. Now, for a given number of components of length *G,* assume the points are generated in an iid (independently and identically distributed) manner from the finite-mixture model:

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{G}} \boldsymbol{\tau_k} \boldsymbol{f_k}(\boldsymbol{x}|\boldsymbol{\theta_k}) \tag{4}$$

where, $\boldsymbol{f_k}(\boldsymbol{x}|\boldsymbol{\theta_k})$ represents the density of the k^{th} group/cluster parameterized by $\boldsymbol{\theta_k}$. $\boldsymbol{\tau_k}$: $= \Pr(x_i \in \mathcal{G}_k$ is called the mixing proportion/weight where $\sum_{k=1}^{G} \tau_k = 1$. The complete set of parameters for a mixture-model with *G* components is:

$$\Theta = \{\boldsymbol{\tau_1}, \ldots, \boldsymbol{\tau_G}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_G}\}$$

Most often and throughout the rest of this work, $\boldsymbol{f_k}$ is taken to be Multivariate Normal (Gaussian) distribution $\phi(\boldsymbol{x}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})$, parameterized by

its mean vector $\boldsymbol{\mu_k}$ and covariance matrix $\sum_{\boldsymbol{k}}$ having the following explicit form:

$$f_k(\boldsymbol{x}|\boldsymbol{\theta_k}) = \phi(\boldsymbol{x}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}) \equiv \frac{\exp\{-1/2(x-\boldsymbol{\mu_k})^T \sum_k^{-1}(x-\boldsymbol{\mu_k})\}}{\det(2\pi\sum_{\boldsymbol{k}})^{-1/2}} \quad (5)$$

This set-up of finite-mixture models are called **Gaussian-Mixture-Model based Clustering** (GMMBC). Estimation of the parameters := $\{\tau_k, \boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}\}$ under such setup is done via the **Expectation-Maximization Algorithm** (EM) (? & ?).

3.2.1 The EM-Algorithm for the parameter estimates under GMMBC setup

Under the above discussed setup, the EM algorithm has the following steps:

(a) Obtain some initial values (randomly) of the parameters and the mixing proportions for the Gaussian mixtures:

$$\{\tau_k^*, \boldsymbol{\mu_k^*}, \Sigma_k^* : k = 1,2, \dots, G\}$$

(b) **E-Step**: Given the initial values, the E-step calculates the conditional probability that the i^{th} observation comes from the k^{th}group as,

$$\tau_{ik}^* = \frac{\tau_k^* \phi(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})}{\sum_{j=1}^G \tau_j^* \phi(\boldsymbol{x_i}|\boldsymbol{\mu_j}, \sum_{\boldsymbol{j}})} \quad (6)$$

this follows from direct application of the *Bayes' Rule*.

(c) **M-Step**: Now, use the estimate of the mixing proportions obtained from E-Step, calculate new parameter values. Let, $n_k = \sum_{i=1}^n \tau_{ik}^*$ i.e., the sum of the mixing proportions for the k^{th} component :– this is the effective number of data points assigned to the component *k*. M-Step gives the following estimates:

$$\tau_k^{new} = \frac{\sum_{i=1}^n \tau_{ik}^*}{\sum_{k=1}^G \sum_{i=1}^n \tau_{ik}^*} = \frac{n_k}{n} \quad (7)$$

$$\mu_k^{new} = \frac{\sum_{i=1}^n \tau_{ik}^* x_i}{n_k} \quad (8)$$

and

$$\Sigma_k^{new} = \frac{\sum_{i=1}^n \tau_{ik}^* (x_i - \mu_k)(x_i - \mu_k)^T}{n_k} \quad (9)$$

(d) Continue iterating (each pair of E & M steps is considered one iteration) between E-Step and M-Step until convergence is attained after which an observation can be assigned to the group for which the corresponding posterior probability is the highest.

Under iid assumption, the MLE (Maximum-Likelihood Estimate) method is generally used to check for the convergence of the above EM. This involves computation of the *log-likelihood* after each iteration and stopping the process when there appears to be no significant change from one iteration to the next. The log-likelihood is defined as follows:

$$\log l\,(\Theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{G} \tau_k \phi_k(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})\right) \tag{10}$$

where, $\phi(.)$ is the Multivariate Gaussian Density.

The results of the above discussed EM are highly dependent on the initial values. Model-based Hierarchical Clustering can often be a good source of initial values for datasets that are moderate in size (?; ?; ?).

The EM solution driven by MLE can fail to converge; instead it can diverge to a point of infinite likelihood. For many mixture models, the likelihood is unbounded and there are paths in parameter space along which the likelihood tends to infinity (?). For examples of such an instance, we refer to the paper by ?. To avoid this instance the failure of convergence can be tackled by replacing the ML-Estimate by the *maximum a posteriori* (MAP) estimate from a Bayesian analysis. This can be achieved by assuming a prior distribution on the parameters that eliminates failure due to singularity; while having little effect on the stable results obtainable without any prior assumption. Under such setup, the Bayesian predictive density for the data is assumed to be of the form:

$$\mathcal{L}(\boldsymbol{Y}|\tau_k, \boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})\, \boldsymbol{\mathcal{P}}(\tau_k, \boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}|\boldsymbol{\theta}) \tag{11}$$

where $\mathcal{L}$ is the mixture likelihood and is given by:

$$\mathcal{L}(\boldsymbol{Y}|\tau_k, \boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{G}} \tau_k \phi_k\,(\boldsymbol{x_i}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})$$

$$\equiv \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{G}} \frac{\exp\{-1/2(\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \sum_k^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu_k})\}}{\det(2\pi \sum_{\boldsymbol{k}})^{-1/2}} \tag{12}$$

and $\mathcal{P}$ is a prior distribution on the parameters $\tau_k, \boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}$, which includes other parameters denoted by θ. Under this setup, we try to find a posterior mode or MAP (maximum a posteriori) estimate rather than a ML-estimate for the mixture parameters. For further discussions on regularization, we refer to ?. This method has been implemented in the software **mclust** (https://cran.r-project.org/web/packages/mclust/index.html) built under the **R** environment by the same authors of the paper. We have used **mclust** for our work and did not face issue with convergence of the EM algorithm (instead, we found excellent computation speed with fast convergence) and thus have not used any prior $\mathcal{P}$.

3.2.2 Model selection under GMMBC setup

As already discussed, the problems in applied cluster analysis: *selection of clustering method and that of number of clusters* can be postulated into one single problem of *Statistical Model Selection* under the MBC setup (?). The approach taken to the problem is based on *Bayesian Model Selection* via the use of Bayes' Factors (?) and posterior model probabilities. The idea goes like this: Let several models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\}$ are considered with prior probabilities of getting selected as $p(\mathcal{M}_k)$, $k = 1, \ldots, K$ often taken to be equal. Now, by applying Bayes' Theorem, the posterior probability of $\mathcal{M}_k$ getting selected given the data *(D)* is:

$$p(D|\mathcal{M}_k) \propto p(D|\mathcal{M}_k)p(\mathcal{M}_k)$$

where,

$$p(D|\mathcal{M}_k) = \int p(D|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k$$

where, $p(\theta_k|\mathcal{M}_k)$ is the prior distribution of θ_k: the parameter vector for the model $\mathcal{M}_k$. $p(D|\mathcal{M}_k)$ is known as Integrated Likelihood of the model $\mathcal{M}_k$. This likelihood will help us in deciding the best model. We will choose that model which is maximum likely a posteriori. Assuming the priors $p(\mathcal{M}_k)$s are equal, then we select the model with highest integrated likelihood, i.e., if we are comparing $\mathcal{M}_i$ and $\mathcal{M}_j$, then we calculate:

$$\mathcal{B}_{ij} = \frac{p(D|\mathcal{M}_i)}{p(D|\mathcal{M}_j)}$$

With the comparison favouring $\mathcal{M}_j$, if:

$$\mathcal{B}_{ij} > 1 \Leftrightarrow p(D|\mathcal{M}_i) > p(D|\mathcal{M}_j)$$

For regular models the *integrated likelihood* can be approximated simply by the **BIC** (Bayesian Information Criterion). Using BIC will add a penalty to the log-likelihood based on the number of parameters, and has shown good performance in a number of applications. (?; ?; ?). BIC can be calculated as follows:

$$2\log p(D|\mathcal{M}_i) \approx 2\log p(D|\theta_i^*, \mathcal{M}_i) - \nu_i \log n \equiv BIC_i \quad (13)$$

where, $\log p(D|\theta_i^*, \mathcal{M}_i)$ is the maximized likelihood for the model and data and ν_k is the number of independent parameters to be estimated from model $\mathcal{M}_i$ (?).

Finally, we can adopt the following strategy to combine all of the methods discussed so far to select the optimal model:

- Select a maximum number of components to consider for our mixture model. Let us call it Gmax.
- Estimate the parameters via the EM and MAP estimate method for each parameterization and each number of components up to Gmax.
- Compute BIC for the mixture likelihood taking the parameter estimates from the EM for up to Gmax clusters.
- Select the model (parameterization/number of mixture components) having the maximum BIC.

3.3 Dimension reduction for visualization

After performing cluster analysis to a group of data it is usually desired to check the distinctness of the clusters created. Popular measure for cluster validity e.g., *Silhouette Width* (?) utilizes Euclidean Distances or any standardized metric for checking the validity of $K-Means$ clustering. But the distance to be used in case of clusters arising from **GMMBC** and other clustering methods are not clearly delineated. ? proposed a methodology to reduce the dimensionality of data so that it can be projected to a subspace of 2 or 3 dimensions and thus we will have a convenient visual representation of the clusters created from a finite mixture of Gaussian densities. Information on the dimension reduced subspace is taken from the various group-specific measures such as, group means and depending on the estimated mixture model: variation on group covariances. The proposed method aims to reduce the dimensionality by identifying a set of linear combinations - called *Directions* - ordered by importance as quantified by the associated eigenvalues of the original features which capture most of the cluster structure contained in the data. After performing all these, observations may then be projected on a dimension reduced subspace. This will facilitate various summary plots which will help us to visualize the clustering structure. The method uses the Gaussianity of

the mixture distributions along with *Bayesian Feature Selection* to extract the directions. For an extensive discussion with mathematical details and implementations to both real and simulated datasets please refer to ?.

The **mclust** software has a function MclustDR() which implements the method discussed in the paper mentioned above. The function, MclustDR() also has an argument lambda which is basically a tuning parameter in the range [0, 1] as described in ?. This argument can be tuned to recover the directions that mostly separate the estimated clusters. The package is maintained by the author himself and thus provides the best possible way to implement the discussed method.

We use this method in our present work just to see how distinct our clusters are and thus helping us to validate the obtained results.

4. Analyses and Interpretations of the Data

The dataset that we have prepared has many desirable characteristics of our nearby galaxies. We proceed to the analysis part with the following steps:

(i) We first perform Gaussian-Mixture-Model-Based Clustering (GMMBC) on the selected number of variables. This step selects the number of optimal clusters with the memberships in each of the clusters/groups.

(ii) Next, we perform the Bayesian LASSO within the selected clusters with all of the variables that we have used for prediction purpose. This step will create a variable selection within the clusters created from the previous step in order to find SFR_{est} values.

(iii) Next, we perform full Bayesian Regression on these clusters/groups with the selected variables from the previous step.

(iv) After all of the above, we come to the interpretation part.

Figure 1 shows the plot of Bayesian Information Criterion (BIC) against the number of mixture components required, i.e., the number of groups. We will choose that number of mixture components (the number of clusters) to be optimal for which the BIC is maximum. The BIC criterion (Figure 1) for selecting the optimal number of clusters was giving rise to 6 clusters. After performing Bayesian LASSO, we observed that Group 4 & 5 had the same set of variables selected and after merging these two clusters gave the same set of selected variables from the LASSO. Also, GMMBC with 6 clusters and GMMBC with 5 clusters only had a difference of 94.1 (1.2%) in BIC. Thus, noting these points and for better interpretability of the Galaxy Clusters we continue our work with 5 clusters (viz. G1 - G5) each of them having unique set of variables explaining their characteristics.

The methods described in Section 3.4 are performed in our data accordingly. The two different directions describing the maximum amount of separation or uniqueness between the clusters/groups are extracted from the data set using the methods described in the paper ?.

We now use these directions for various visualizations helping us in distinguishing the clusters/groups and evaluating their uniqueness. Figure 2 shows the density plots and boxplots of the *Direction 1* of the dimension reduced subspace for 5 different clusters. Figure 3 shows the scatter-plot of the two directions : *Direction 1* & *Direction 2* together explaining about 97% of variation present in the dataset in terms of Eigenvalues and Eigenvectors. Scatters for 5 different clusters/groups are shown in 5 different colors. This plot helps us in visualizing the separation of the clusters between them and also the homogeneity within them.

Table 1 gives the Means and Standard-Errors (given in the brackets below the means) of various parameters of the Galaxies under study. The table gives the values separated by the groups/clusters created.

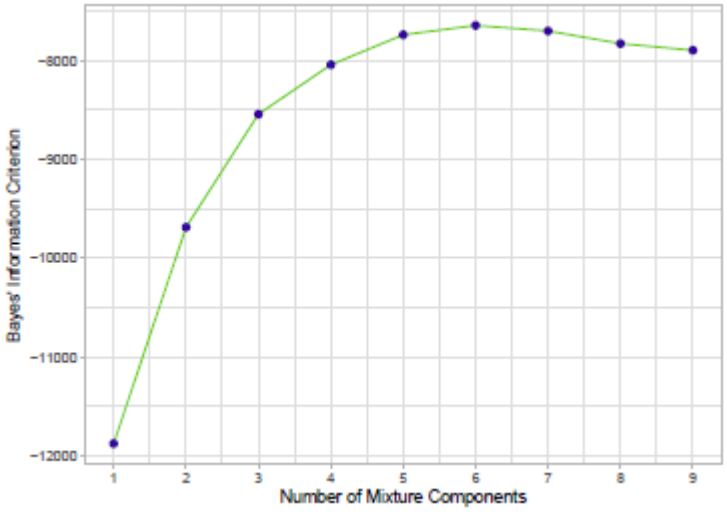


Figure 1. Bayesian Information Criterion for choosing the number of Clusters/Groups. The number of clusters/groups are along the horizontal axis and the corresponding BICs are in the vertical axis of the plot.

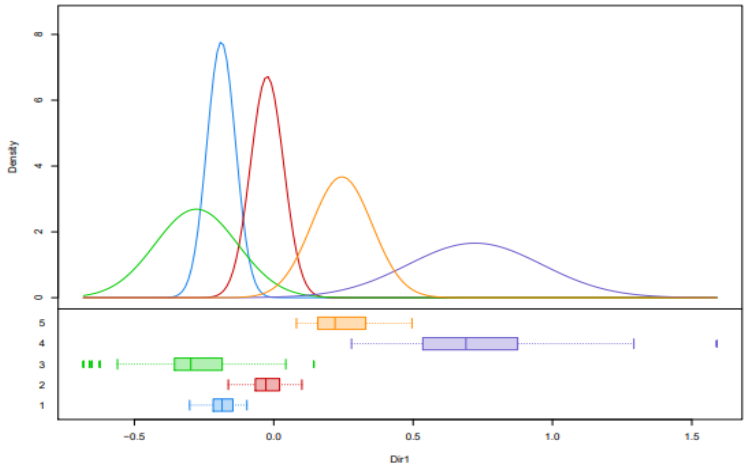


Figure 2. Density plots and Box-Plots for the 5 different clusters/groups (viz. G1 - G5 from left to right) of the Direction 1 of the dimension reduced subspace extracted from the data. The plots of the 5 different clusters are displayed in 5 different colors.

Eigenvalues and Eigenvectors. Scatters for 5 different clusters/groups are shown in 5 different colors. This plot helps us in visualizing the separation of the clusters between them and also the homogeneity within them.

Table 1 gives the Means and Standard-Errors (given in the brackets below the means) of various parameters of the Galaxies under study. The table gives the values separated by the groups/clusters created.

4.1 Properties of the groups and their star formation mechanisms

It is clear from Table 1 and Figures 1 to 4 that G4 and G5 are strongly rotating (viz V_m in Table 1) disc dominated (viz. A_{26} in Table 1) late type spiral galaxies (viz. mostly Sa - Sd type as evident from 'T' indices) with largest size, HI mass and follow moderately close trends toward Tully - Fisher relation in B band, $logL_B \propto 1.55\, logV_m$ in G4 and $logL_B \propto 1.69\, logV_m$ in G5 (viz . $L_B \sim V_m^3$, ?). Their SFRs are independent of the environmental influence (viz. $\Theta_1 \propto .298$ in G4 and -0.115 in G5 and Figures 11 and 12) (probably due to strong rotation) and depends only on gas mass and surface brightness.

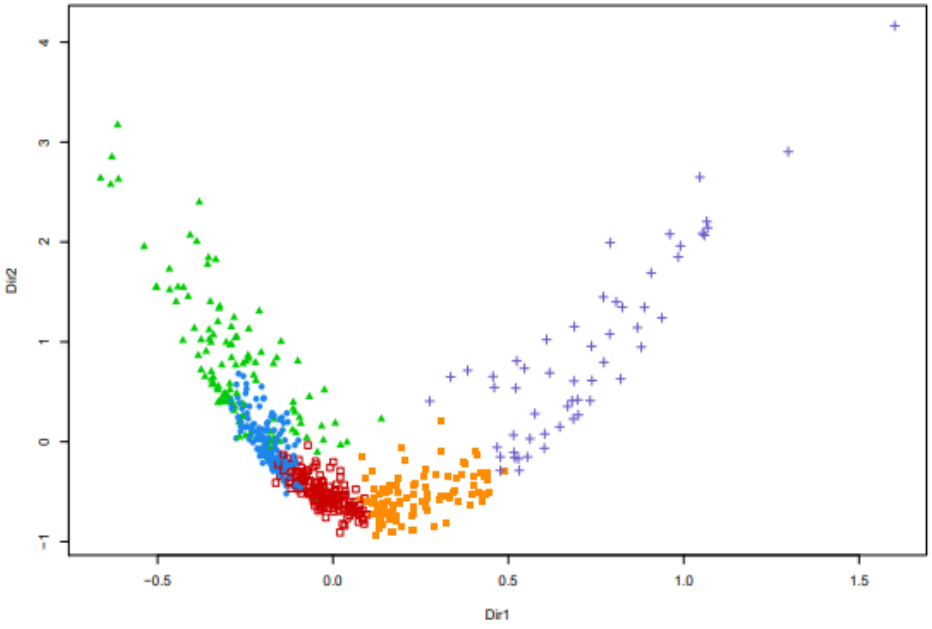


Figure 3. Scatter plot of the two directions of the dimension reduced subspace showing the separation of the 5 clusters (G1 - G5 from left to right). The five data points for different clusters are displayed in different colors

G3 galaxies are faintest , smallest in size having lowest neutral hydrogen mass and their SFRs are dominated by environment. This is also evident from the list of significant parameters found in LASSO for Bayesian Regression and highest Θ_1-value of Table 1). This group is predominated by lenticular (viz. almost 50 %) and Irregular types. G3 has a large percentage of bulge dominated galaxies (viz T $\sim -3/-1$ for 51 objects). Though the SFR of G4 is highest and that for G3 is smallest, SFEs (viz L_K/MHI) are comparable for G4 and G3 (viz. $\sim$ 13.08 and 14.06. Table 1). The reason might be due to rapid consumption of neutral gas for dwarf galaxies than for larger disc galaxies (?).

G2 and G1 have moderate rotations and they do not follow Tully-Fisher relation (viz. for G2, $logL_B \propto logV0.814_m$, for G1, $logL_B \propto 0.705\ logV_m$). This might be due to the fact that turbulent gas motions for galaxies in these groups play significant role in mass distribution. G2 contains most of the BCDs (viz 50) than G3 (viz 25). In G1, SFR is weekly controlled by environment (viz. Table 1 and Figures ?? and ??). In G2 and G3, SFE are very low. It might be due to the fact that though in G1 SFR is moderate, due to moderate influence of the environment ($\Theta_5 \sim 0.498$) and a higher value of gas depletion time ($\sim$ 15.30 Gyr, Table 1) the gas is depleted in longer time and due to low availability of HI mass (viz. Table 1) SFE becomes low. On the other hand in G2, massive stars form in a shorter time (mean M_{FUV} of G2 is greater than mean M_{FUV} of G1, viz. Table 1) which depletes neutral gas (not in enough amount compared to G4 and G5) by galactic wind decreasing the SFE to a minimum value.

Also the SBB profiles are different in the groups indicating different star formation histories of the galaxies in the groups. It is clear from the A_{26} values from Table 1 that G4 and G5 have maximum surface brightness followed by G2 and G1 and finally G5 group of galaxies have minimum ones. This indicates that G4 and G5 are largest in size and G3 is the smallest group of early type bulge dominated lenticular galaxies. The same scenario has been reflected from their several SFR distributions . The SFRs are maximum in groups G4, G5 and G2 and lowest in G3.

Table 1. Means and Standard Errors of various Galaxy-Parameters for G1 (Galaxy Cluster/Group 1) to G5 (Galaxy Cluster/Group 5). Standard Errors are given in the brackets below the means of the variables.

Parameters	G1	G2	G3	G4	G5
Membership of Groups	159	162	115	56	104
b/a	0.6857 (0.0129)	0.5819 (0.0138)	0.6959 (0.0147)	0.5357 (0.0347)	0.4902 (0.0254)
M_{FUV}	-9.525 (0.185)	-12.482 (0.112)	-5.760 (0.448)	-16.030 (0.192)	-14.632 (0.122)

M_k	-15.164 (0.0966)	-16.876 (0.0786)	-13.957 (0.267)	-22.471 (0.194)	-19.374 (0.120)
M_{21}	-11.509 (0.109)	-13.869 (0.0831)	-9.224 (0.384)	-17.128 (0.187)	-15.750 (0.0914)
A_{26}	1.4654 (0.0337)	3.6219 (0.0817)	1.410 (0.127)	27.92 (1.35)	10.198 (0.332)
$M_{\boldsymbol{B}}$	-12.591 (0.0931)	-14.536 (0.0766)	-10.718 (0.285)	-19.705 (0.156)	-17.025 (0.113)
SBB	24.460 (0.0707)	24.496 (0.0648)	25.481 (0.161)	23.721 (0.107)	24.223 (0.0970)
$log(\widetilde{SFR})$ (calculated using m_{FUV})	-3.2303 (0.0701)	-2.0788 (0.0428)	-4.727 (0.180)	-0.3166 (0.0802)	-1.0791 (0.0494)
$log(MHI)$	6.9369 (0.0435)	7.8815 (0.0334)	6.023 (0.154)	9.1855 (0.0749)	8.6342 (0.0368)
$MHI = L_k$	0.5208 (0.0325)	0.9836 (0.0660)	1.508 (0.261)	0.2056 (0.0336)	0.5752 (0.0524)
Θ_1	0.248 (0.119)	-0.1870 (0.0937)	1.466 (0.180)	0.298 (0.166)	-0.115 (0.150)
Θ_5	0.498 (0.109)	0.0772 (0.0840)	1.627 (0.168)	0.525 (0.154)	0.138 (0.136)
$log\ L_k$	7.3864 (0.0387)	8.0680 (0.0314)	6.902 (0.107)	10.323 (0.0781)	9.0766 (0.0486)
$\log M_{26}$	7.0483 (0.0736)	8.2970 (0.0510)	7.223 (0.175)	10.691 (0.0689)	9.5389 (0.0435)
MH_α	-7.001 (0.297)	-10.330 (0.157)	-4.141 (0.485)	-14.957 (0.243)	-13.131 (0.138)
W_{50}	31.02 (1.23)	54.61 (1.93)	38.89 (3.96)	247.5 (17.8)	114.47 (5.07)
V_h	466.8 (30.6)	610.3 (27.6)	217.0 (42.1)	540.5 (39.7)	503.9 (28.4)
D	6.487 (0.248)	8.102 (0.266)	4.620 (0.441)	8.183 (0.521)	9.144 (0.517)
i	53.28 (1.44)	62.72 (1.39)	54.14 (1.79)	59.11 (2.91)	62.27 (2.06)
V_m	12.02 (1.01)	26.87 (1.27)	18.80 (3.14)	141.31 (9.48)	60.54 (2.74)
AB_i	0.000440 (0.000242)	0.00846 (0.00238)	0.00217 (0.00124)	0.4088 (0.0445)	0.1588 (0.0184)
V_{LG}	433.5 (24.6)	569.7 (24.7)	238.7 (30.9)	519.1 (29.4)	500.1 (22.8)
$\tau(Gyr)$	15.30 (1.18)	10.934 (0.519)	49.96 (3.43)	3.782 (0.367)	6.075 (0.419)
$SFE\ = L_K/MHI$	4.676 (0.462)	2.716 (0.303)	13.08* (2.26)	14.7* (2.96)	5.311 (0.825)

* After removing few outliers (at 5 % level through box plots)

Clustering and classification of Astronomical objects- A new paradigm in Statistics

Asis Kumar Chattopadhyay*
Department of Statistics, University of Calcutta, India
Visiting Scholar, Concordia University, Canada

Abstract

This collection of works involves the application of statistics to astronomy and the development of statistical methods to solve the problems related to the universe, leading us to discoveries of new astrophysical phenomena. Data collection missions like Galaxy Evolution Explorer, Kepler Space Telescope, Hubble Space Telescope and virtual archives like Sloan Digital Sky Survey, Multi-mission Archive at STSCI, NASA Extragalactic Data base preserve petabytes of data, which can be used for big data analyses. Usually collection of data on celestial bodies is obscured by bad weather conditions, obstruction by another celestial object or instrumental restrictions and it cannot be repeated. Hence we often get data contaminated with noise, affected by outliers or sparsely distributed. In all such situations, the usual statistical methods fail and we need to use their adaption or to introduce new methods as per requirements. To overcome such problems, there are various transformations and denoising techniques available in the literature (e.g. kernel principal component analysis (KPCA)). Sometimes there are rare objects, unevenly spaced data of unequal lengths where classical statistical methods are only applicable when the data is interpolated to get into a form suitable for the methods. We have suggested some possible solutions under the above scenario.

Keywords

Astrostatistics, Clustering, Classification, Kernel Principal Component Analysis, Missing Value.

1. Introduction

Astronomy, perhaps the oldest observational science, has got its spectacular emergence with the advent of theoretical astrophysics. During the last two decades galaxy formation theory and their related star formation histories have drawn interests among the astrophysicists to a great extent to uncover these mysteries using the reach treasure of virtual archives. Scientists working in the theoretical areas like cosmology and relativity are gradually becoming interested in database analysis because of technological advancements enabling us to have data related to such physical phenomenon.

While digging the pathway, a new branch (Chattopadhyay et al.(2014)) Astrostatistics (or Statistical Astronomy) has emerged since 1980s. It is a blending of statistical analysis of astronomical data along with the development of new statistical techniques useful to analyze astrophysical phenomenon. The target is not only to explore the formation and evolutionary history of galaxies but also to uncover the unknown facts related to star formation, gamma ray bursts, supernova and other intrinsic variable stars.

Gamma-ray bursts (GRBs), the brightest explosion in the universe, since the Big Bang, show huge variation in their duration. This duration may vary from ten milliseconds to several hours, indicating the variation in formation of them. To explore the possible sources, clustering of GRBs is performed in different ways (Chattopadhyay et al. (2007) and references therein). Among the controversy that the number of natural groups in GRBs is 2 or 3, Modak et.al. (2018) use kernel principal component analysis(KPCA) (Scholkopf & Smola (2002), chapter 14) to GRB data set to perform clustering as well as dimension and noise reduction. Previous work of kernel principal component analysis on astronomical data includes supernovae (Ishida et al. (2013), (2012)), image denoising, etc. Kernel principal component analysis is a nonlinear transformation on raw data, where non-linear features are extracted from data in terms of kernel principal components. It is a generalization of linear transformation performed in standard principal component analysis, where linear features are extracted from data in terms of principal components.

Statistical analysis with missing data is an important problem as the problem of missing observation is very common in many situations. In astrostatistics one should look at missing value problem from a different angle (Chattopadhyay (2017)) since the causes of missing observation are sometimes inherent in the process. The imputation method may not be applicable to some astronomical data sets as the missing value may arise from physical process and imputing missing values is likely to be misleading and can skew subsequent analysis of data. For example, the Lyman break technique (Giavalisco, M. (2002)) can identify high-redshift galaxies based on the absence of detectable emissions in bands corresponding to the FUV rest frame of the objects. Such high-redshift galaxies were previously unobservable. De et al. (2016) has tackled the problem by including the knowledge of missing proportion in a classification rule.

In the present work we have discussed about some of the above mentioned applications of statistical methods.

2. Clustering Gamma Ray Bursts Data-methodology

Modak et al. (2018) retrieved a dataset from the fourth BATSE Gamma-Ray Burst Catalog (revised) (Paciesas et al. (1999)), consists of information on 1972 GRBs for the following 9 variables. F1, F2, F3, F4 are time-integrated fluence in

20−50, 50−100, 100−300 and > 300 keV spectral channel, respectively; P64, P256, P1024 are peak flux measured in 64,256 and 1024 ms bin, respectively; T50, T90 are time within which 50% and 90% of the flux arrive. Unit of fluence is given in ergs per square centimeter (ergs cm−2), unit of peak flux is count per square centimeter per second (cm−2 s−1) and unit of time is second (s). First, observations on each variable are standardized because the ranges of the variables vary largely. Then, for a particular choice of kernel, KPCA is performed on them. They extracted nonlinear features using the significant kernel principal components. Then using them as study variables, k-means clustering method (Hartigan–Wong clustering algorithm (Hartigan et al. 1979)) is performed on them in which the number of clusters is determined with the help of gap statistic (Tibshirani et al. (2001)).

Kernel Principal Components are supposed to carry less information and more noise with increasing order and after a certain order they fail to give any relevant information regarding the data under study. So, they started choosing first few KPCs and number of chosen KPCs is increased as long as their performance gets better in terms of an accuracy measure. In this context, they used one accuracy measure as the Dunn index (Dunn (1974)), which indicates the internal validation of a performed clustering, taking values between 0 and ∞ with greater value indicating better clustering.

3. Results and interpretation

First, k-means clustering method is applied to the standardized variables of GRB data set in which gap statistic gives no clustering in GRBs, i.e., raw GRB data set fails to reveal the inherent clustering nature in GRBs. Then the same method is applied to the principal components, extracted from the GRB data through principal component analysis. Linear features (first two PCs), explaining more than 80% variation in data, results in one group of GRBs. Thus linear information on data can't expose the natural groups present in GRBs. A new choice of Kernel successfully reveals the inherent clustering nature in GRBs, by extracting the relevant nonlinear information from raw data in terms of kernel principal components. We see the first two KPCs, extracted through kernel (10) with $p < 1$ and for every choice of s considered, are enough to describe the data. In Chattopadhyay et al. (2007), k-means clustering approach is directly applied to differently chosen study variables and 1594 GRBs are clustered in three groups of sizes 622, 423, and 549, respectively with 4.08 % 1-NN classification error rate. While their clustering of 1972 GRBs based on the first two kernel principal components, extracted by proposed kernel with $p = 1/2$ and $s = \sigma 1$, group those 1594 GRBs in three clusters of sizes 827, 438, and 329, respectively with 0.2% 1-NN classification error rate.

Here they not only reduced the burden of the data, but also extracted the inherent information from the data, on which simple clustering method reveals

the natural groups in GRBs. We propose a new possible way, kernel principal component analysis, to analyze GRB data set as well as a new kernel, which makes the clustering results better in comparison with the other existing kernels and gives three physically interpretable groups in GRBs. However explanation of the sources of these three groups will be more prominent in the future with more data collection.

4. Classification under bivariate gamma set up with incomplete observation

The analysis of De, Bhattacharya and Chattopadhyay(2016) was based on the sample of Globular clusters (GCs) of the early-type central giant elliptical galaxy in the Centaurus group, NGC 5128, whose structural parameters have been derived by Mclaughlin et al. The distance is that adopted by McLaughlin et al. (2008), namely 3.8 MPc. The sample consists of 130 GCs whose available structural and photometric parameters are tidal (Rtid,in pc), Core radius (R_c, in pc), half light radius(r_h,in pc), central volume density ($log\rho_0$) in $M_\odot\, pc^{-3}$), $\sigma_{p.0}$ (predicted line of sight velocity dispersion at the cluster center (in kms^{-1})), twobody relaxation time at the model projected half mass radius(trn,in years), galactocentric radius (Rgc, in kpc), Concentration (c $\sim log(R_{tid}/R_c)$), dimensionless central potential of the best fitting model (W_0), extinction-corrected central surface brightness in the F606W bandpass (μ_0 in $magarcsec^{-2}$) , v surface brightness averaged over r_h (< μ_v > h) in $magarcsec^{-2}$, integrated model mass($logM_{tot}$ in MJ), washington T_1 magnitude, extinction corrected color $(C-T_1)_0$. and metallically ([Fe/H])index determined from the color $(C-T_1)_0$. However, for their purpose, they have considered only $logR_h$ and $logM_{tot}$. It was found that 127 among the 130 data points (i.e. 97.69%) satisfy the restriction $logR_h < logM_{tot}$. Thus there was a clear indication of the order restriction among the realized values of $logR_h$ and $logM_{tot}$. These two variables were found to be jointly distributed as bivariate Gamma. They have considered three forms of bivariate Gamma distributions proposed by Mckay, Nadarajah & Gupta and a transformed form proposed by the authors.

Hence the problem was to form the discriminant function for observations coming out from some bivariate gamma distribution. In particular, considering two groups, say Group 1 and Group 2 and a random observation *(X,Y)* such that *(X,Y)* $\sim f_i(x,y)$ under Group i=1,2, where $f_i(x,y)$ is the density of a bivariate gamma distribution the discrimination function was developed. Assuming that 'loss' is the cost associated with misclassifying an observation, they have denoted the loss for classifying an observation to group 1 when it originally belonged to group 2 by c(1|2) = l_1 and the loss for classifying an observation to group 2 when it originally belongs to group 1 by c(2|1) = l_2. Then one should classify (x,y) to Group 1 if

$$\log_e \frac{f_1(x,y)}{f_2(x,y)} > \log_e \frac{\pi_2 c(1|2)}{\pi_1 c(1|2)} \text{.....Rule1}$$

else, classify to group 2, where π_i is the prior probability for the i th group, i = 1,2. If we assume $\pi_1 = \pi_2 = 1/2$, then the above rule R1 classifies (x,y) to Group 1 if

$$\log_e \frac{f_1(x,y)}{f_2(x,y)} > \log_e(l_1) - \log_e(l_2)$$

For the estimation of unknown parameters involved in the classification rule, they have used the maximum likelihood estimates.

The total cost of misclassification (TCM) according to Rule 1 corresponding to three forms of bivariate Gamma distribution and two sets of choices of classification cost are shown in the following table-1. The analysis was carried out on the basis of above mentioned data set on the galaxy NGC5128 and starting with two groups obtained by k-means clustering with k=2.

Table 1: TCM for NGC 5128 data set

Gamma distribution	TCM ($l_1 = 0.9, l_2 = 0.3$)	TCM ($l_1 = 0.2, l_2 = 0.8$)
First form	0.3409	0.2657
Second form	0.1365	0.0182
Third form	0.1501	0.0265

The authors have proposed under the above set up, a classification rule by including the knowledge of missing proportion in the construction of classification rule, as described below.

Assume few observations are missing in a data set containing n observations. If *l* is the cost per observation, the total cost is *nl* when no observation is missing. But for *m* missing observations in the data, total cost reduces to $(n-m)l$. Now $(n-m)l = n(1-\frac{m}{n})l$. If p is the proportion of missing observations in the data, then $p = \frac{m}{n}$. Hence, $(n-m)l = n(1-p)l \propto (1-p)l$, which indicates that it will be reasonable to take the loss as $l(1-p), 0 < p < 1$, in such a situation.

Consequently, if we have two groups to classify, we redefine the loss due to misclassification as $c(1|2) = l_1(1-p)$ and $c(2|1) = l_2(1-p_2)$, where l_i are known constants, p_i is the proportion of missing observations in the i th group ,i = 1,2. It can be noted that when there is no missing value in the data, i.e., $p_1 = p_2 = 0$, then $c(1|2) = l_1$ and $c(2|1) = l_2$. So, the proposed loss due to misclassification becomes equal to that without missing observations. Hence, one can apply earlier adopted Rule 1 after discarding the missing observations (i.e. marginalization) in the data or after substituting the missing observations (i.e. imputation). But under marginalization there would be loss of information and when the proportion of missing observation is quite significant in the data application of imputation techniques may result in loss

of accuracy and the severity of error increases with the increased proportion of missing observations. To tackle this situation, they proposed this alternative classification rule by including the proportion of missing observation in the construction of classification rule.

For the redefined loss due to misclassification, the classification rule with prior probabilities π_1 and π_2 would be to classify an observation (x, y) to group 1 if

$$\log_e \frac{f_1(x,y)}{f_2(x,y)} > \log_e \frac{\pi_2 c(1|2)}{\pi_1 c(1|2)} = \log_e \frac{\pi_2 l_1 (1-p_1)}{\pi_1 l_2 (1-p_2)} \text{.....Rule 2}$$

else classify to group 2.

For Rule 2 the total cost of misclassification is less than that for Rule 1.

In order to compare the performances of the rules on the basis of above mentioned astronomical data, the following exercise had been carried out.

As before, for the NGC 5128 data set, they have first performed a k-means clustering with k=2 and formed two clusters(55 observations in cluster 1 and 72 observations in cluster 2).Here the only difference was at each step they have created few missing observations in the data.

Then they performed the classification in two different ways:

1. Rule 1a: The missing observations were discarded and Rule 1 was used treating the resulting data set as a whole
2. Rule 2a: Rule 2 was used considering the missing proportions.

As earlier, they considered the three candidate distributions and found the concerned TCM considering different choices of (l_1, l_2) and (p_1, p_2). Rule 1a and Rule 2a are essentially the modified versions of Rule 1, adjusted for the presence of missing observations. The resulting analysis is shown in Table-2 given below.

Table 2: TCM for NGC 5128 data with missing values

Rules	Gamma distribution	TCM (l_1 =0.9, l_2 =0.3) (p_1 = 0.06, p_2 = 0.15)	TCM (l_1 =0.2, l_2 =0.8) (p_1 = 0.06, p_2 = 0.15)
Rule 1a	First form Second form Third form	0.170 0.050 0.029	0.045 0.019 0.004
Rule 2a	First form Second form Third form	0.097 0.039 0.013	0.104 0.022 0.009

The important findings from the above tables are, the total cost of misclassification for second and third form is less than the total cost of misclassification for the first form. Also, when one includes the missing

proportions in the classification rule, the misclassification proportion reduces even for all the three forms.

5. Discussion and Conclusion

From the above two case studies, it is clear that for astronomical data analysis standard statistical methods may fail and it is necessary to develop proper methods.

References

1. Chattopadhyay, T., Misra, R., Chattopadhyay, A. K., & Naskar, M.(2007). Statistical evidence for three classes of gamma-ray bursts, Astrophysical Journal, 667, 1017-1023.
2. Chattopadhyay , A.K and Chattopadhyay,T. (2014), Statistical Methods for Astronomical Data Analysis, Springer Series in Astrostatistics, New York.
3. Chattopadhyay, A.K. (2017) Incomplete data in Astrostatistics, Wiley StatsRef: Statistics Reference Online, 1-12, John Wiley & Sons.

Unsupervised classification of galaxy spectra and interpretability

Didier Fraix-Burnet
Univ. Grenoble Alpes / CNRS / IPAG, Grenoble, France

Abstract

Dealing with large amount of data is a new problematic task in astrophysics. One may distinguish the management of these data (astroinformatics) and their scientific use (astrostatistics) even if the border is rather fuzzy. Dimensionality reduction in both the number of observations and the number of variables (observables) is necessary for an easier physical understanding. This is the purpose of classification which has been traditionally eye-based and essentially still is, but this becomes not possible anymore. In this talk, I present an unsupervised classification of 700 000 spectra of galaxies of 1500 wavelengths each, with a model-based subspace clustering algorithm (Fisher-EM). I also show some preliminary results on the interpretation of the classes using data bases of modelled spectra.

Keywords

Unsupervised classification; machine learning; spectra; astrophysics; galaxies

1. Introduction

Astrophysics has now entered the era of Big Data and the new telescopes and instruments that will come into operation in the next few years (EUCLID, VLT/MOONS, LSST, SKA...) face technological challenges for the management and the analysis of the data. Spectra are particularly spectacular since they contain several thousands of wavelengths making matrices of about a million observations described by thousands of parameters.

These spectra contain all the astrophysical information that an astronomer can dream of, apart from the morphological structure: the composition of the stellar populations, the history of the stellar formation events, the content in gas and its physical conditions, the presence of hot regions such star forming regions, hot nebulae, active galactic nuclei hosting black holes, and the global kinematics of the galaxy. Basically the spectrum of a galaxy is made of a continuum due to the thermal emission from the stars, plus some absorption features due to the cold gas, and emission lines due to hot gas. An atlas of typical galaxy spectra is provided in Kennicutt (1992) or Dobos et al. (2012).

Classification in astrophysics traditionally uses an eye-based approach which also serves as the bases for supervised learning studies. In contrast, unsupervised learning is not common (Fraix-Burnet et al. 2015). In the case of

spectra of galaxies, the first study using k-means is recent (Sánchez Almeida et al. 2010) and has been disputed in De et al. (2016). Indeed, the data set is so large that this simple technique is not able to detect structures.

This prompted us to use more sophisticated techniques to automatically and objectively build a statistical classification of galaxy spectra. Firstly, we want to use unsupervised clustering since we are convinced that using supervised learning is currently weird since the training set is devised by human subjectivity. Data mining is a better approach to know what kind of structures in the data set algorithms are able to detect. Secondly, the large amount of parameters (wavelengths) is very probably redundant. Sánchez Almeida, et al. (2010) as selected a priori physical interesting regions of the spectra, but this introduces biases for the discriminative power of the classification. Using dimensionality reduction techniques such as Principal Component Analysis has only been used up to now to separate spectra of galaxies from those of stars. However, principal components are known to be unadequate to perform a clustering (Chang 1983). As a consequence, we have chosen a discriminative latent subspace clustering approach designed both to reduce the dimensionality and to perform an unsupervised clustering, as described in the Methodology section.

2. Methodology

The data consists in 702248 spectra of galaxies and quasars with resdshift smaller than 0.25 that were retrieved from the Sloan Digital Sky Survey (SDSS) database, release 7 (http://www.sdss.org/dr7/). These data and their preparation are described in De et al. (2016) except that we here do not select wavelength bands to reduce the number of wavelengths initially of more than 3000 to around 1500. Rather, we applied a wavelet filtering of the noise followed by a binning by a factor of 2.

The unsupervised clustering was performed with the algorithm FisherEM available in R (Bouveyron & Brunet 2012). The Fisher-EM algorithm is a discriminative latent mixture model that estimates both the discriminative subspace and the parameters of the mixture model. It is based on the Expectation-Maximization (EM) algorithm from which an additional step, named F-step, is introduced, between the E- and the M-step. This F-step uses the Fisher's criterion under orthonormality constraints and conditionally to the posterior probabilities to optimize the clustering.

Due to computational constraints and an algorithm currently written in R, we analyzed several sets of 100 000 spectra, as well as one with 300 000 spectra. The latter took two weeks of computation with the current non-parallelized R code.

3. Results

We here present the result on one set of 100 000 spectra (Fig. 1) as well as the one for 300 000 spectra (Fig. 2) for comparison. In all cases, the optimal number of groups according to the integrated completed likelihood (ICL) criterion is found to be 20.

For the first set, four of the groups have only one or two spectra which are very odd, while for the second set three groups are in this case. These odd spectra are either too noisy, or caused by an technical problem at the time of observation, or coming from another kind of source (star?) and misclassified in the automatic procedure. They are not considered in this paper and not shown in the figures, but the fact that they were isolated among the large data set is in itself a nice outcome since real outliers could thus be found.

Another noticeable result is that both classifications agree very well, the same kinds of spectra can be easily identified. Indeed, this is true for the results on all the sets of 100 000 spectra, showing the robustness of the clustering algorithm. A more quantitative comparison is in progress.

The dispersion of spectra within groups is always low, there is little overlap implying that the groups are highly specific. This is remarkable since both the number of observations and the number of variables are very large. In addition, galaxies are characterized by quantitative properties that vary continuousy over large ranges of values, so that the data space made by their spectra could be expected to be more or less homogeneously populated, without clearcut structures. Our result tends to show that this is not the case however.

For astronomers, the most remarkable outcome is that the groups map remarkably well the diversity of galaxy spectra that has been observed or modelled. This is certainly the first time in astrophysics that without any physical a priori, a classification is obtained and provides immediate and easily identifiable physical properties.

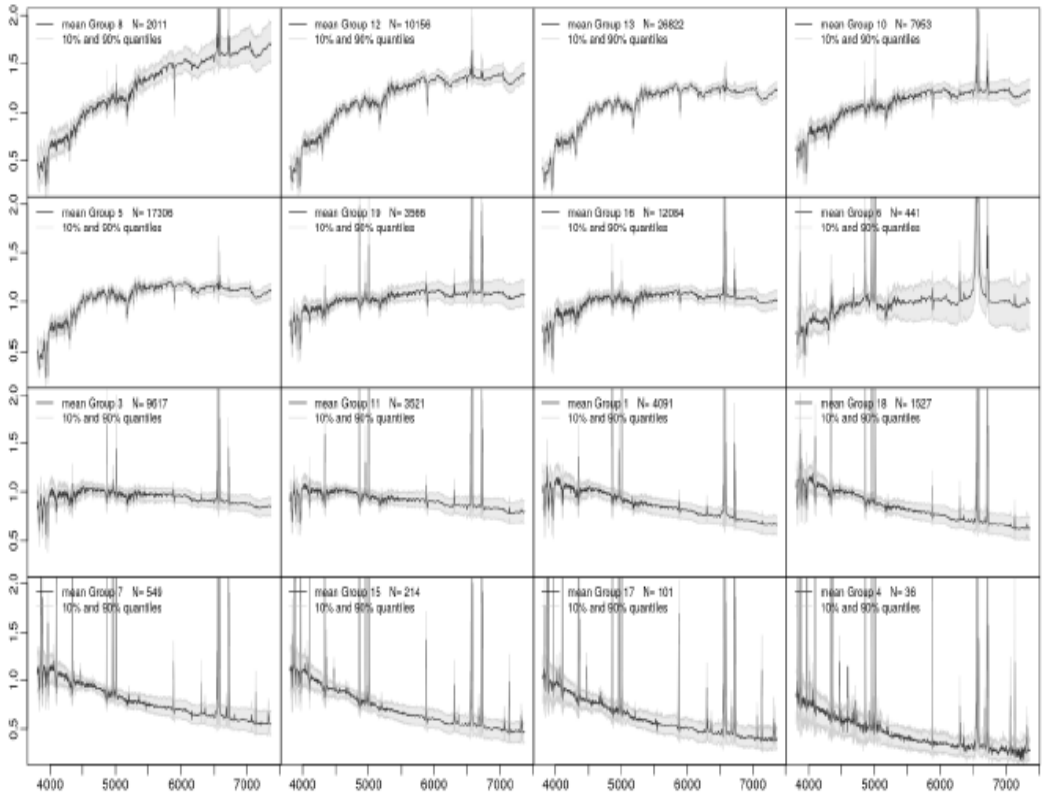


Figure 1. Median spectra (black lines) for sixteen groups obtained with a set of 100 000 spectra. The abscissae are the wavelengths in Ångström (Å), and the ordinates is the flux. All spectra before the analysis have been normalized using the integrated flux between 4300 and 5000 Å. The shaded areas show the region between the 10% and 90% quantiles in each group. The legends give the number N of spectra in each group.

On the two figures, the spectra are ordered according to their global slopes (continuum). This continuum emission is produced mainly by the populations of stars: the first groups on the upper left of the two figures are clearly populated by old (red) stars, while the groups toward the bottom have young (bluer) stars. The latter also form stars since there are many emission lines spiking out. These emission lines can also be due to an active galactic nuclei, a central region where very probably a lot of matter is attracted and heated by the massive black hole that we now know is present in most of the galaxies, if not all.

It must be kept in mind that galaxies are made of billions of stars and many gaseous regions. They are thus mixtures of different populations of stars, of different ages and chemical compositions. This explains why several types of spectra are identified among red and blue galaxies. As a consequence, there seems to be several nearly identical spectra (for instance in the two first lines in the figures), varying by subtle differences of slopes together with more significant ones for the emission lines. Another ingredient of galaxies is the

dust (in reality cold gas and very small grains) that can absorb the blue part of the spectrum and make it steeper than in reality. This phenomenon is very probably present in the spectra of the first group, at least.

Absorption lines give some indication on the chemical composition of the stars. It is easy to see that their intensity and their ratios generally differ from one group to the other. Sometimes a more detailed examination of the median spectra is necessary.

The emission lines are characteristics of several atomic elements (hydrogen, carbon, nitrogen, oxygen...) and their ratios give invaluable information on the physical conditions of the hot gas. This gas can be heated by young and massive stars, by shocks in the interstellar medium or by the presence of a black hole in its vicinity, most generally near the center of the galaxies. The differences between the last eight groups in the two figures are spectacular and reveal really different kinds of objects.

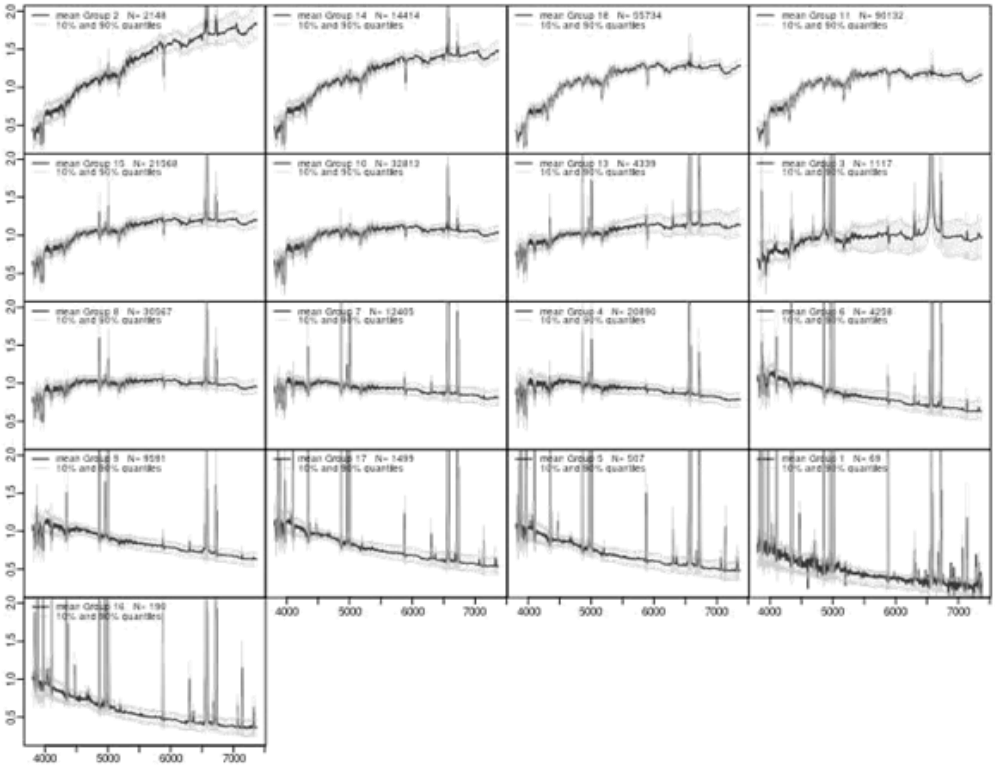


Figure 2. Idem Fig. 1 for the set of 300 000 spectra. Seventeen groups are shown

4. Discussion and Conclusion

The unsupervised classification results we have obtained on a very large sample of astrophyscial spectra appear remarkable in that it can automatically identify recognizable typical galaxy properties. We have shown that the

classification is statistically robust through analyses of several subsets of 100 000 observations. This classification appear quite relevant through an eye-based examination of the median spectra. We have given some hints on the physical meaning of the spectra and the relevance of the groups and their specificities. However, a more detailed analysis is required to estimate the astrophysical properties of the galaxies of the different groups. This work is in progress.

We want to insist on the fact that the number of groups is chosen thanks to an objective likelyhood indicator. In addition, this number of groups does not vary with the different subsets of 100 000 distinct spectra, nor with the 300 000 subset. This probably shows that the true number of typical spectra that can be distinguished in the data is around 16 or 17 is we exclude the few outliers that came out in all analyses. This number of typical spectra is given by statistics, and may or may not correspond to the true number of typical galaxies. To check this point, careful fits of each spectra will have to be performed (Moultaka et al. 2004 ; Noll et al. 2009).

Even if it is possible to fit millions of spectra with models and derive physical properties of galaxies (Comparat et al. 2017), classification remains invaluable since it allows to perform more precise fits on much fewer typical spectra. In any case, classification would still be required to simplify and understand the properties of millions of galaxies and we find highly preferable to classify the data themselves rather than the derived values prone to their own uncertainties, errors, degeneracies and model inadequacies.

The present results open the possibility of an automatic and objective classification procedure for the big databases already available and yet to come. An extension of the algorithm, called sparse-FEM, allows to identify the most discriminant parameters (here wavelengths) that explain the classification. In other words, it is then possible to perform a supervised classification of new spectra by using these parameters which are much less (say 100), alllowing for a nearly real time classification.

Finally, an important outcome of our work is the possibility to easily detect outliers or interesting and rare objects (Baron & Poznanski 2017). The real time supervised classification could also be used to detect quickly any modification in the spectrum of a known galaxy. This would be quite useful for the transient alerts sent to other telescopes.

References

1. Baron, D. & Poznanski, D. (2017). The weirdest SDSS galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society,* **465,** 4530-4555.

2. Bouveyron, C. & Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing,* **22**, 301-324.
3. Chang, W.-C. (1983). On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions. *Applied Statistics,* **32,** 267-275.
4. Comparat, J., Maraston, C., Goddard, D., Gonzalez-Perez, V., Lian, J., Meneses-Goytia, S., Thomas, D., Brownstein, J. R., Tojeiro, R., Finoguenov, A., Merloni, A., Prada, F., Salvato, M., Zhu, G. B., Zou, H. & Brinkmann, J. (2017). Stellar population properties for 2 million galaxies from SDSS DR14 and DEEP2 DR4 from full spectral fitting. *Astronompy & Astrophysics*, submitted (arXiv:1711.06575).
5. De, T., Fraix-Burnet, D. & Chattopadhyay, A. K. (2016). Clustering large number of extragalactic spectra of galaxies and quasars through canopies. *Communication in Statistics - Theory and Methods,* **45,** 2638-2653.
6. Dobos, L., Csabai, I., Yip, C.-W., Budavári, T., Wild, V. & Szalay, A. S. (2012). A high-resolution atlas of composite Sloan Digital Sky Survey galaxy spectra. *Monthly Notices of the Royal Astronomical Society,* **420,** 1217-1238.
7. Fraix-Burnet, D., Thuillard, M. & Chattopadhyay, A. K. (2015). Multivariate Approaches to Classification in Extragalactic Astronomy. *Frontiers in Astronomy and Space Sciences,* **2**, 3.
8. Kennicutt Jr., R. C. (1992). A spectrophotometric atlas of galaxies. *The Astrophysical Journal Suppl.*, **79**, 255-284.
9. Moultaka, J., Boisson, C., Joly, M. & Pelat, D. (2004). Constraining the solutions of an inverse method of stellar population synthesis. *Astronomy & Astrophysics,* **420,** 459-466.
10. Noll, S., Burgarella, D., Giovannoli, E., Buat, V., Marcillac, D. & Muñoz-Mateos, J. C. (2009). Analysis of galaxy spectral energy distributions from far-UV to far-IR with CIGALE: studying a SINGS test sample. *Astronomy & Astrophysics,* **507**,1793-1813.
11. Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C. & de Vicente, A. (2010). Automatic Unsupervised Classification of All Sloan Digital Sky Survey Data Release 7 Galaxy Spectra. *The Astrophysical Journal*, **714,** 487-504.

A statistically robust approach to the detection of astrophysical transient and periodic phenomena

Guillaume Belanger
European Space Agency (ESAC), Madrid, Spain – gbelanger@sciops.esa.int

Abstract

Transient phenomena are interesting and potentially highly revealing of details about the processes under study that could otherwise go unnoticed. It is therefore important to maximise the sensitivity of the detection methods used. We present a general procedure based on the likelihood function for identifying transients that is ideally suited for real-time applications because it requires no grouping or pre-processing of the data. The method use all the information available in the data throughout the statistical decision making process, and is suitable for a wide range of applications. Here we consider those most common in astrophysics which involve searching for transient sources, events or features in images, time series, energy spectra, and power spectra, and demonstrate the use of the method in the case of a short-lived quasi-periodic oscillation in a power spectrum. We present two new periodogram statistics, the $\mathcal{R}_k^2$ and the $\mathbb{Z}^2$, and derive two fit statistics, the K and B statistics, relevant to model fitting in frequency space.

Keywords

Transient phenomena; periodic phenomena; likelihood function; astrophysics; X-rays.

1. Introduction

The way in which the measurements are distributed defines the appropriate statistical treatment. Each measurement considered individually, and the collection of measurements as a whole, carry statistical evidence that can be used to assess the agreement between a given hypothesis or model and the data. Treating data as evidence is a powerful means to detect changes, differences, deviations or variations. This is done using the likelihood function.

The detection of an event localized in time, involves identifying something that was not there before. Whether it rises, dwells, and decays over weeks and months like a supernova, or whether it just appears and disappears in a fraction of a second like a γ-ray burst; whether it manifests as a complete change of shape of the energy spectrum during a state transition in a black hole, or as the short-lived emission line from an accretion event; whether it comes as a sudden change of spectral index in the power spectrum or as the appearance of an ephemeral quasi-periodic oscillation (QPO); all of these

phenomena, independently of timescale, share in common that they appear as a sharp change.

We here address the task of identifying *transients*—any feature or change that can be identified in the data as *statistically distinct* from the underlying process. It should be understood that if a feature cannot be distinguished by statistical means, it cannot be detected and identified, whether this is because the transient is too weak or too long-lived. The limitations of a detection procedure can always be accurately established before applying it.

The power spectrum refers to the power spectral density distribution of a physical process, whereas the periodogram refers to an estimate of the power spectrum. The most common choice of a periodogram statistic is the Discrete Fourier Transform, and it is generally used in the form of a computationally fast algorithm called Fast Fourier Transform (FFT; see Press et al., 2002) applicable only to grouped data. More sensitive periodogram statistics include the Rayleigh or R^2-test (Leahy et al., 1983), the Z^2-test (Buccheri et al., 1983), and the H-test (de Jager et al., 1989). Two important features that make these tests more powerful than the standard FFT periodogram are: (1) they can be applied directly to event arrival times, and thus access all variability timescales present in the data, and (2) they impose no constraints on the frequencies that can be tested, and are thus said to oversample the periodogram by testing timescales other than those corresponding to independent frequencies.

But oversampling without taking into account that the variables computed to estimate the power are correlated within an independent Fourier spacing (IFS) leads to frequency-dependent artifacts that distort the periodogram and that can be interpreted as signatures of coherent periodic modulations. Each of the above R^2, Z^2 and H tests suffers from this.

Although it is powerful—the most powerful according to Leahy et al. (1983)—in detecting sinusoidal modulations in event data, the R^2 achieves this by estimating the power using the fundamental harmonic only. This advantage for sinusoidal signals is a limitation for non-sinusoidal pulses. The Z^2 statistic was devised for this purpose as a generalization of the R^2 that combines any number of harmonics. Accessing the power in higher harmonics confers the Z^2 an important advantage over the R^2, and explains why it is the statistic of choice for event data where pulses are peaked or irregular in shape. A powerful and reliable periodogram statistic must fulfil three conditions: it must (1) be able to use event arrival times in order to access all variability timescales, (2) allow for oversampling in order to explore frequency space without restrictions, and (3) take into account the oscillation in the mean, variance, and covariance of the Fourier components as a function of frequency. These are met by the *modified* Rayleigh statistic.

We present a general transient detection method based on the likelihood function applicable to a wide range of problems.[1] We present two new periodogram statistics: the generalized (kth order) modified Rayleigh statistic, labeled $\mathcal{R}_k^2$; and the modified Z^2 statistic, labeled $\mathcal{Z}^2$, that benefit from all the features of their predecessors but do not suffer from the artifacts caused by unaccounted for correlations in the trigonometric moments. Finally, we present two fit statistics, the K and B statistics, optimal for χ^2 and exponential variables respectively, and thus relevant to the model-testing in frequency space.

2. Methodology

A transient can only be identified as such in relation to the underlying background, which can either be constant or variable. In most applications where transient sources are searched for, the background is constant or nearly so. Intensity ratios and variability timescales must be worked with as parameters in order to establish optimal detection algorithms and thresholds. The first measurement gives the first estimate of the reference value: the value we expect to measure under usual conditions when there is no transient. The second measurement gives a second estimate of that reference value, but it can also be evaluated for its potential of being a transient. The third measurement gives a third estimate of the reference, the likelihood of measuring such a value is evaluated by the ratio of the single-measurement likelihood function centered on the maximum likelihood reference value given by the previously calculated joint likelihood.

With each subsequent measurement we: (1) compute the likelihood of the newly measured value based on the single-measurement function defined by the current reference value; (2) if the likelihood is less than the defined threshold, issue a transient event trigger. Do not update the estimate of the reference value; (3) if the likelihood is more than the defined threshold (within the likelihood interval), recalculate the joint likelihood function including the new measurement and update the reference. The thresholds must be optimized for the application.

The detection of transients in images, time series, energy spectra, and power spectra follows the same methodology, applying different analytical probability functions depending on the input data. The illustration uses the periodogram for which the $\mathcal{R}_k^2$ statistic is used for greatest sensitivity to weak signals. Unlike images, time series, and energy spectra, the values of power estimates in frequency channels are related to χ^2 and exponential distributions.

[1] The mathematical statistics of likelihood are from the work of Fisher (1912, 1922); the philosophical basis is primarily from Royall (1997); and other technical details of data analysis and statistics are mostly from Cowan (1997).

The reason is that each power estimate is calculated from a sum of squared standard normal variates.

The natural choice for a general optimal χ^2 fit statistic is twice the negative of the log-likelihood, $K = -2\ln L$, dropping terms that do not depend on the parameters k_i:

$$K = -2\sum_i \left[\left(\frac{k_i}{2} - 1\right) ln x_i - \frac{k_i}{2} ln 2 - ln\Gamma\left(\frac{k_i}{2}\right)\right] \quad (1)$$

The K statistic is optimal for x^2 data—for fitting a model to a set of measurements that are samples of random x^2 variables with potentially different dof ki in each channel i.

The ideal case of a globally flat power spectrum is the simplest manifestation of a red noise with a spectral index of zero. The power values in red noise are related to one another through the relation $Power \propto f^{-\alpha}$, where f is the frequency and α is the power spectral index. In this case, we are working with *scaled* χ_2^2 *variables*. This can be verified by dividing the power estimates by the best-fit power-law model and thereby recovering the χ_2^2 distribution. This would also be true for *any* power spectral shape *if* the process can be considered as one of simply scaling the basic χ_2^2 variable that results from summing two squared standard normal variables by the underlying model, whatever the shape. We assume this to be true, and thus work with the power estimates at a given frequency as though they were χ_2^2 variables scaled differently in each channel. This implies they are distributed according to the exponential density function.[2] We can therefore construct another fit statistic specifically for fitting periodograms (Duvall & Harvey (1986) also derive and use this statistic for this purpose):

$$B = 2\sum_i (\ln \tau_i + x_i/\tau_i) \quad (2)$$

where x_i is the measured and τ_i is the model-predicted power in frequency channel i.[3](see Belanger, 2013, for further details)

In searching for periodic modulations, as sensitive as the R^2 and Z^2 statistics may be to weak sinusoidal signals and non-sinusoidal pulse profiles, both suffer in exactly the same way from over-sampling artifacts caused by

[2] Having recognized that the powers in a frequency channel of any periodogram are exponentially distributed with a mean given by the expected power in that channel, the one-sided tail probability of finding a power value of 60 or greater, for instance, when the expectation is 30, is 0.135 or 13.5%, quite low in terms of statistical significance. However, using normal statistics (mean power of 30 and standard deviation of √30, say), finding a value of 60 or greater is a 5.47o result with a probability of about 10^{-8}.

[3]Sensitivity to detect changes in the overall shape of the spectrum increases quickly with the number of iterations. However, fluctuations in the power estimates in each channel always remains important due to the intrinsically high variance of exponential variables for which the standard deviation is equal to the decay constant (j. = r, o^2 = r^2 and thus o = r).

correlations within each IFS most noticeable at lower frequencies. For independent frequencies, the integral of the sine and cosine components is always zero. For all other

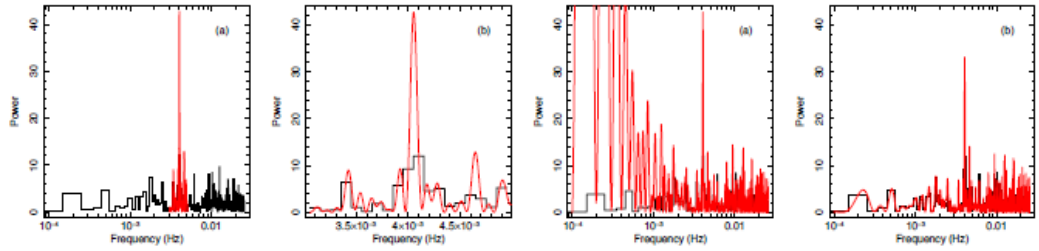


Figure 1: Comparing the FFT, R^2, and $\mathcal{R}_1^2$ periodograms on simulated white noise (duration T = 10 ks, mean rate μ = 0.5s^{-1}), with a 10% pulsed fraction for a sinusoid at 0.00405 Hz (≈247 s). The FFT periodogram in black is computed on 512 time bins (δt=19.531 s) and thus 256 real frequencies with $\nu_{min} = \delta f =$ 10^{-4} Hz and ν_{max} = 0.0256 Hz. The R^2 periodogram (in red) is computed on a restricted range around ±1 IFS of the peak (0.003–0.005 Hz) with sampling of 21 frequencies per IFS. Panel (a) shows the full range of the FFT, and panel (b) shows the range of the R^2 periodogram. A period that falls between two independent frequencies best illustrate the major difference in sensitivity that can be achieved. Artifacts in the R^2 periodogram are shown on a truncated vertical linear scale in panel (c). R^2 estimates between independent frequencies deviate noticeably from the FFT below ≈ 3×10^3Hz. Lastly, panel (d) shows the $\mathcal{R}_1^2$ statistic, free of artifacts, detecting the signal at the same frequency (0.004052 Hz) but at a lower power from the more accurate calculation (33.1 instead of 42.8: probability of 10^{-8} instead of 10^{-10}.) frequencies, it is not. Similarly, their variances—assumed to be equal to one-half, and covariance—assumed to be zero, also oscillate. Hence, FFT and R^2 powers are equal or nearly so at independent frequencies, but can vary wildly in between. Figure 1 panel (a) shows the R^2 periodogram, and panel (b) shows the modified Rayleigh statistic and demonstrates the advantage it has over the standard FFT periodogram for detecting weak signals peaking between independent frequencies without the severely limiting disadvantages of the classical Rayleigh statistic. $R^2{}_k$ is identically as sensitive as $R2$ for the fundamental harmonic (by mathematical definition), but it is, in addition, equally sensitive for any other harmonic. We define the generalization of the modified Rayleigh statistic for any harmonic:

$$\mathcal{R}_k^2 = \begin{pmatrix} C_k - \langle C_k \rangle \\ S_k - \langle S_k \rangle \end{pmatrix}^T \begin{pmatrix} \sigma^2 C_k & \sigma C_k S_k \\ \sigma C_k S_k & \sigma^2 S_k \end{pmatrix}^{-1} \begin{pmatrix} C_k - \langle C_k \rangle \\ S_k - \langle S_k \rangle \end{pmatrix} \quad (3)$$

The dependency on the harmonic is carried by the variable k in the argument of the sine and cosine functions to yield the following expressions for C_k and S_k:

$$C_k = \frac{1}{N}\sum_{i=1}^{N} \cos k\emptyset\, i \qquad \text{and} \qquad S_k = \frac{1}{N}\sum_{i=1}^{N} \sin k\emptyset_i. \tag{4}$$

The other terms are defined as follows:

$$\langle C_k \rangle = \frac{1}{k\omega T}\left[\sin k\omega t\right]_{t_1}^{t_2}, \tag{5}$$

$$\langle S_k \rangle = \frac{1}{k\omega T}\left[\cos k\omega t\right]_{t_1}^{t_2}, \tag{6}$$

$$\sigma_{C_k}^2 = \frac{1}{2N}\left(1 + \frac{1}{k\omega T}\left[\sin k\omega t\ \cos k\omega t\right]_{t_1}^{t_2}\right) - \langle C_k \rangle^2, \tag{7}$$

$$\sigma_{S_k}^2 = \frac{1}{2N}\left(1 - \frac{1}{k\omega T}\left[\sin k\omega t\ \cos k\omega t\right]_{t_1}^{t_2}\right) - \langle S_k \rangle^2, \tag{8}$$

$$\sigma C_k S_k = \frac{1}{2k\omega TN}\left[sin^2 k\omega t\right]_{t_1}^{t_2} - \langle C_k \rangle \langle S_k \rangle, \tag{9}$$

The terms $\langle C_k \rangle$ and $\langle S_k \rangle$ are the expectation values, $\sigma_{C_k}^2$ and $\sigma_{S_k}^2$ are the variances, and $\sigma C_k S_k$ is the covariance of C_k and S_k. The *modified* Z^2 statistic, $\mathbb{Z}^2$, is a sum of $\mathcal{R}_k^2$ components,

$$\mathbb{Z}^2 = \sum \mathcal{R}_k^2, \tag{10}$$

and is also ideal for harmonic decomposition of the pulse profile (see details in Belanger, 2016).

3. Results

As a demonstration, we consider a hypothetical observation in X-rays of a bright (500 s−1) accreting system whose variable emission comes mostly from two components: the accretion disk, and the hot and turbulent gas in the inner flow. In both, the emission processes are connected on all timescales, and thus each gives rise to a red noise component. The accretion disk is much larger in extent and has a sharp inner radius. It dominates at lower frequencies with a power-law index $\alpha = -1$, and has a high-frequency cutoff beyond which it does not contribute to the power spectrum. The turbulent inner flow is much smaller in extent because it is bounded by the inner edge of the disk. Its emission is more variable and dominates the high-frequency part of the spectrum with a power-law index $\alpha = -3$.

We are interested in monitoring the range of frequencies between 0.1 and 10 Hz for a weak, short-lived,transient QPO that we expect to appear at or near the break in the power spectrum at 1 Hz, which marks the boundary between the disk and the turbulent inner flow. For this, we make a periodogram every 10 s with the events accumulated during this time interval, and monitor the power. Because we are interested in a short-lived transient, we cannot rely on it persisting in more than one "measurement", and therefore must establish a

single detection threshold that is constraining enough for our application. This threshold is established using simulations.[4] The observation and the analysis are presented in Figure 2 where we see that the transient QPO is clearly detected in the likelihood monitoring at 1Hz, but because it is very short-lived, is not at all evident in the periodogram of the whole observations.

4. Conclusion

The transient detection method here presented is well suited to handle transients in a non-variable background without any further considerations. Naturally, the identification efficiency depends intimately on the strength of the signal. The method is perfectly suited for analyzing archival data. It is, however, also powerful for real-time applications. Handling the third class of transients characterized by a variable background requires additional care, a work that will be presented in a future publication.

Briefly, the crucial consideration is that of the timescales involved: that of the transient with respect to that of the underlying variability. More specifically, since the stationarity of the probability distribution can be considered as being a function of the timescale at which the process is viewed, in general it is possible to have a running estimation of that probability distribution which is stationary up to a given timescale, but evolves on longer timescales. In this way, the likelihood function and all the associated statistics are well defined at any point in time, and the method becomes a more general, time-dependent form of the procedure presented. The power of the method relies on simulations for an accurate estimation of the statistics of the process, and for defining the detection thresholds. The generality of the formalism is such that it can be applied to identifying transients in other parameter spaces, where the independent variable is not time.

[4]We have done this for the power at 1 Hz to first determine the average expected power (35), and then establish a threshold (log-likelihood of −10.1, and thus a likelihood of $4.1 \times 10-5$) that ensures a level of false detections of 5%.

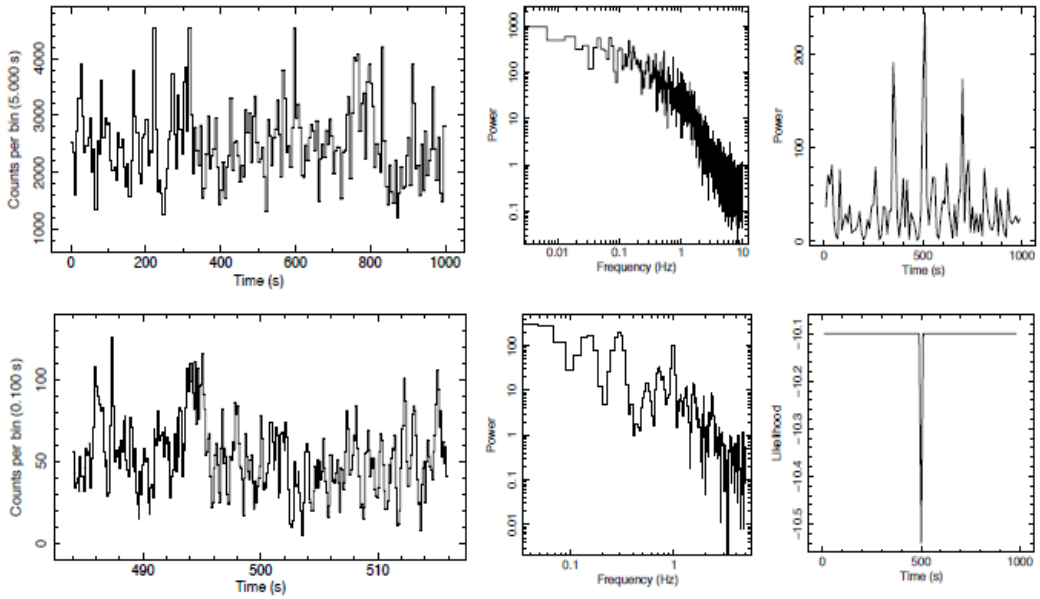


Figure 2: The top row shows the time series of the entire observation (binned to a resolution of 5 s for clarity of presentation); the periodogram made from the Kalman-filtered, 0.05 s resolution time series of the event arrival times; and the power at 1 Hz estimated at 10 s intervals from the Rayleigh periodogram of the event arrival times as a function of time. The bottom row shows a zoom on the time series during the transient QPO from its start at 485 s until its end at 515 s after the start of the observation; the periodogram of the Kalman-filtered 0.05 s resolution time series; and the log-likelihood as a function of time where only detections beyond the established threshold are shown. The QPO is characterized by 30 cycles of an almost periodic signal centered on 1 Hz with a standard deviation of 1/20 and a pulsed fraction of 27%.

When performing a search for a weak periodic signal—weak because it is seen over few cycles, because it has a very low pulsed fraction, or because it is short-lived—the sensitivity of the periodogram statistic we use is critical. In event data, the most sensitive are those presented in Equations (3) and (10), which we have labelled the $\mathcal{R}_k^2$ and $\mathbb{Z}^2$ statistics. Fitting models to data requires an appropriate fit statistic. An optimal fit statistic is one derived from the likelihood function, and must reflect the statistical properties of the random variable in question. For a collection of χ^2 or a collection of exponential variables, the optimal fit statistics are those presented in Equations (1) and (2), which we have labelled the K and B statistics.

References

1. Belanger, G. 2013, ApJ, 773, 66
2. Belanger, G. 2016, ApJ, 822, 14
3. Buccheri, R., Bennett, K., Bignami, G. F., et al. 1983, A&A, 128, 245
4. Cowan, G. 1997, Statistical Data Analysis (Oxford: Clarendon) de Jager, O. C., Raubenheimer, B., & Swanepoel, J. 1989, A&A, 221, 180
5. Duvall, T. L. J., & Harvey, J. W. 1986, in Proc. NATO Advanced Research Workshop, Seismology of the Sun and the Distant Stars, ed. D. O. Gough, (Greenbelt, MD: NASA, Goddard Space Flight Center), 105
6. Fisher, R. A. 1912, Messenger Math., 41, 155
7. Fisher, R. A. 1922, RSPTA, 222, 309
8. Leahy, D. A., Elsner, R. F., & Weisskopf, M. C. 1983, ApJ, 272, 256
9. Press, W. H., Teukolsky, S. A., Vetterling, & Flannery, B. P. 2002, Numerical Recipes in C++ : the Art of Scientific Computing (Cambridge: Cambridge Univ. Press)
10. Royall, R. M. 1997, Statistical Evidence, A Likelihood Paradigm (New York: Chapman & Hall/CRC)

Statistical definition of employment in the environmental sector and green jobs: Theory and practice

Valentina Stoevska
International Labour Office, Department of Statistics

Abstract

The 19th International Conference of Labour Statisticians (ICLS), Geneva, 2-11 October 2013, adopted guidelines concerning statistical definition of employment in the environmental sector and green jobs. The guidelines defined the environmental sector as consisting of all economic units producing, designing and manufacturing at least some goods and services for the purposes of environmental protection and resource management. The guidelines draws a distinction between employment in the production of environmental goods and services for consumption by other economic units (i.e. employment in production of environmental outputs) and for consumption by the economic unit in which the activity is performed (i.e. employment in environmental processes). Green jobs are specifically referred to as a subset of employment in the environmental sector, meeting the requirements of decent work.

Following the 19th ICLS, a number of countries participated in pilot programmes designed to test concepts and definitions presented in the guidelines and new data collection methodologies.

This paper first will describe the main statistical concepts and definition of employment in the environmental sector, present the results of the surveys in Mongolia and Albania, and outline some of the challenges in generating data through statistical surveys and monitoring the transition towards a green economy. It will also highlight its possible application for assessing the sustainability of the tourism industries.

Keywords

statistics framework; employment; environmental sector; green jobs; international guidelines

1. Introduction

The concept of the green economy has become a focus of policy debate in recent years and has been mainstreamed into the work of the United Nations and its specialized agencies. Much of the discussion has focused on the potential of the green economy to provide significant opportunities for investment, growth, and jobs while addressing the global environmental challenges, especially climate change. This has led to an increasing need for

the statistical community to deal with the difficult task of defining and measuring the concepts of "green jobs" in order to produce internationally harmonized statistics that would inform the ongoing policy debate on the economic and employment impact of "greening" the economy.

In order to provide clear statistical definition of the green jobs that would facilitate production of internationally comparable data, draft guidelines concerning statistical definition of employment in the environmental sector and green jobs were discussed and adopted by 19th ICLS, 2013

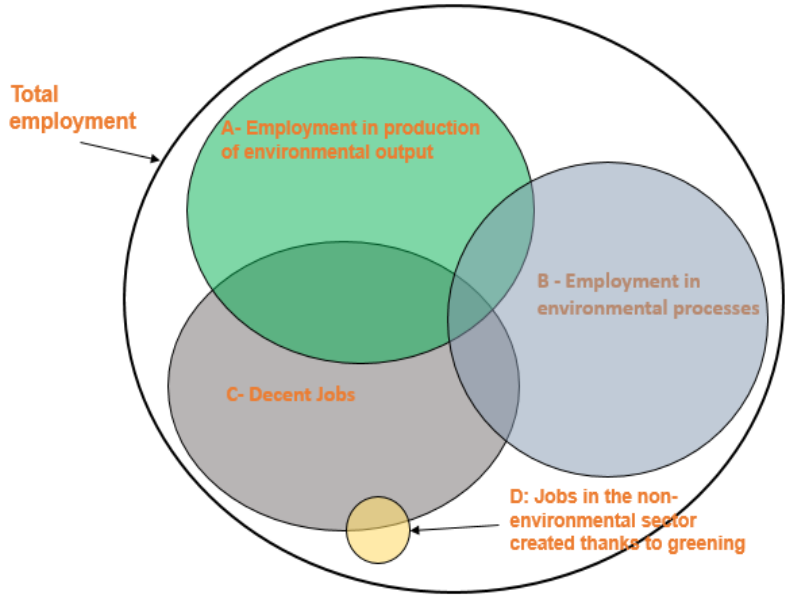
2. Concepts of employment in the environmental sector and green jobs

The 19th ICLS guidelines define the environmental sector as comprising all economic units that carry out environmental activities – those defined in the Central Framework of the United Nations' System of Environmental-Economic Accounting (SEEA) as economic activities whose primary purpose is to reduce or eliminate pressures on the environment or to make more efficient use of natural resources. They are grouped into two broad types of environmental activity:

- *Environmental protection activities* defined as activities whose primary purpose is the prevention, reduction and elimination of pollution and other forms of degradation of the environment.
- *Resource management activities* defines as activities whose primary purpose is the preservation and maintenance of the stock of natural resources and hence safeguarding against depletion.

Environmental activities can be carried out by all economic units, as main, secondary or ancillary activities. A distinction is made between (i) specialist producers, (ii) non-specialist producers, and (iii) own-account producers of environmental goods and services. These units produce, design and manufacture at least some goods and services for purposes of environmental protection and resource management (e.g. environmental specific services, environmental sole-purpose products, adapted goods, environmental technologies)

Figure 1: Employment in the environmental sector



Employment in environmental sector= A∪B
Employment created thanks to greening = A∪B∪D
Green jobs (Sub-component of employment in environmental sector that is decent)= (A∪B)∩C

Employment in the environmental sectors is defined comprising all persons who, during a set reference period,

- were involved in the production of environmentally desirable output (employment in environmental output).
- whose duties involve making their economic unit's production processes more environmentally friendly or make more efficient use of natural resources. This includes using methods, procedures, practices, or technologies that, for example reduce or eliminate pollution, reduce consumption of water and energy, minimize waste, or protect and restore ecosystems (employment in environmental processes).

This distinction takes into account the fact that environmental output not always produced by using environmental processes and technologies. These two components of employment in the environmental sector shed light on different ways of greening enterprises and economies and offer different entry points for policies.

Green jobs are defined as a subset of employment in the environmental sector that meets the requirements of decent work (e.g. offer adequate wages, safe conditions, workers' rights, social dialogue and social protection).[1]

[1] The decent work dimension of jobs may be measured according to relevant indicators selected from Guidelines for Producers and Users of Statistical and Legal Framework Indicators ILO Manual, second version. http://www.ilo.org/stat/Publications/WCMS_223121/lang--en/index.htm

Other related concepts include:

- Employment in the non-environmental sector created thanks to greening: This refers to employment in economic units that supply goods and services to the environmental sector.
- Employment in low carbon economic units and energy efficient enterprises: This refers to employment in units that have low carbon emissions (e.g. employment in green buildings) and to employment in enterprises that are more energy efficient than most of the enterprises within the same economic activity.
- 'Green work': This refers to all work involved in production of environmental goods and services. It includes employment, voluntary work and own-use production work4 to produce environmental goods and services.

3. Sources of data on employment in the environmental sector

Employment in the environmental sector and green jobs can be estimated by using data from inventories, regular statistical surveys and censuses; and specialized statistical modules, surveys and censuses, including subsample surveys.

Each source has its advantages and disadvantages. The suitability of each source should be evaluated on the basis of:

- size and importance of the environmental sector, or parts of this sector;
- efficiency of data collection (extent and level of detail needed for the analysis and relative cost in terms of resources and time to collect these data); and
- data quality (in terms of coverage, comprehensiveness and comparability).

Additional consideration to be taken into account are (a) Scope of the assessment, (b) Definitions, (c) type of jobs to be covered (e.g. direct, Indirect, Induced), (e) Net or Gross employment effect (job creation, substitution, job elimination, job transformation), etc.

In practice, different approaches may complement each other. Some parts of employment and economic activity in the environmental sector may be gauged by using data derived from existing regular statistical surveys (e.g., labour force surveys (LFS)) or administrative records (e.g., records maintained by industry associations or by government ministries or agencies), while others may require new data collection. Where the data compiled are incomplete,

statistical modelling techniques[2] may be needed – e.g., input–output analysis and social accounting matrices, dynamic macroeconomic models and other computable general equilibrium models.

Best use of existing data combined with modelling techniques can yield sufficiently exhaustive estimates at relatively low compilation cost without increasing the burden for data providers.

3.1 Specialized statistical modules, surveys and censuses, including subsample surveys

In order to collect comprehensive data regarding the employment in the environmental sector, specialized statistical modules, surveys and censuses are needed, because the on-going periodic statistical surveys such as LFS or establishment surveys based on existing classifications such as ISIC do not allow for exhaustive identification of all economic units carrying out environmental activities.

As comprehensive stand-alone surveys are not feasible because of time or resource constraints, information on employment in the environmental sector can be obtained by adding specific modules/questions to ongoing statistical surveys or censuses. A representative subsample rather than all sample of the ongoing survey may be surveyed for this module.

Albania [3] and Mongolia [4] are one of the first countries that have implemented the guidelines and estimated the number of jobs in the environmental sectors on the basis of modules for compiling separate information on employment in production of environmental outputs and on employment in environmental processes.

According to the results of the surveys in both countries approximately 1/3rd of all employed spend at least part of their working time on the production of environmental goods and/or services or using environmental processes and/or technologies. However, the percentage of those that spend more than half of their work time on production of environmental output or using environmentally friendly processes is below 3 per cent. In both countries, almost 40 per cent of enterprises face shortage of trained personnel with knowledge and skills in environmental activities/practices; and between 30

[2] See ILO. Methodologies for assessing green jobs: Policy brief. www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_ent/documents/publication/wcms_176462.pdf.

[3] Report on the pilot project towards developing statistical tools for measuring employment in the environmental sector and generating statistics on green jobs (ILO, 2014) http://www.ilo.org/stat/Publications/lang--en/index.htm.

[4] Employment in the environmental sector and green jobs in Mongolia, Pilot Study, September 2017.

and 50 per cent estimate that the level of environmental sensitivity of workers is satisfactory.

4. Tourism sustainability and the issue of measurement

Tourism plays an important role in many countries because of its economic and employment potential. The tourism industries makes a significant contribution to the overall level of economic activity and employment, and contributes greatly to the development of regions. Because of its reliance on foreign visitors, it proves to be more resilient to the economic crisis and provide jobs in particular for economically less advantaged socio-demographic group or regions. According to the World Tourism Organization (UNWTO) 1 out of 11 jobs are directly or indirectly related to tourism (UNWTO, 2018), and that its share in total employment is increasing.

With the increasing share of tourism in economic activities in most countries, it is acknowledged that (i) it is contributing more to the use of environmental resources and its impact on the natural environment was increasing, (ii) tourism activity might provide a path by which lower income countries and region might improve their standard of living.[5]

The potential of the tourism to make a significant contribution to the economic, social and environmental development, and to create decent jobs and generate economic growth, was emphasized by the international community in the 2030 Agenda for Sustainable Development6 (UN, 2015), in particular in Sustainable Development Goals (SDGs). There are three targets that relate directly to sustainable tourism, namely:

- Target 8.9: By 2030, devise and implement policies to promote sustainable tourism that creates jobs and promotes local culture and products.
- Target 12.b: Develop and implement tools to monitor sustainable development impacts for sustainable tourism that creates jobs and promotes local culture and products
- Target 14.7: By 2030, increase the economic benefits to small island developing States and least developed countries from the sustainable use of marine resources, including through sustainable management of fisheries, aquaculture and tourism

[5] FRAMING SUSTAINABLE TOURISM October 2016, Carl Obst, UNWTO Consultant and Director of the Institute for the Development of Environmental- Economic Accounting http://cf.cdn.unwto.org/sites/all/files/pdf/mst_issue_paper_1.pdf

[6] United Nations, 2015. Transforming our World: the 2030 Agenda for Sustainable Development. General Assembly resolution 70/1 (New York). https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

4.1 Measuring the sustainability of tourism

According to the UNWTO, Sustainable tourism is defined as "tourism that takes full account of its current and future economic, social and environmental impacts, addressing the needs of visitors, the industry, the environment and host communities"[7] (UNWTO/UNEP, 2005).

An important element in this definition is the fact that tourism has to address not only environmental issues but also to ensure inclusive economic growth and social development. In other words sustainable tourism should ensure positive contribution to environmental integrity while contributing at the same time to poverty reduction, social justice, decent work, gender equality, economic development, etc.

While comprehensive this definition does support identification of economic units within the tourism industries that might be considered sustainable, as opposed to traditional tourism. The second issues is related to the identification of economic units in tourism industries because they cannot be identified from the production side as other industries.

The Statistical Framework for Measuring Sustainable Tourism (SF-MST)[8] was developed to provide an organizing structure for integrating statistics on the economic, environmental and social dimensions of sustainable tourism. For each dimension it identifies a number of aspects and areas of focus relevant to sustainable tourism.

4.1 *Employment data for assessing the three dimensions of sustainability of tourism*

As tourism-related employment appears to be high on the agenda of the industry and policy makers and is identified as one of the indicators to monitor the progress towards achievements of SDG target 8.9, three is a need for establishing clear and measurable indicators, based on internationally agreed standards. The ILO statistical standards concerning employment, decent work and employment in the environmental activities and green jobs as well the System of Environmental - Economic Accounting (SEEA) Central Framework[9] and Tourism Satellite Account[10] provide measurement framework.

A recommended set of employment related indicators assessing the three dimensions of sustainability of tourism for may include:

---

[7] Making Tourism More Sustainable - A Guide for Policy Makers, UNEP and UNWTO, 2005.

[8] Statistical Framework for Measuring Sustainable Tourism- Draft prepared for initial round of consultation with the UNWTO Committee on Tourism Statistics and TSA and the Working Group of Experts on Measuring Sustainable Tourism
http://cf.cdn.unwto.org/sites/all/files/pdf/sf-mst_feb.pdf

[9] https://unstats.un.org/unsd/envaccounting/seearev/seea_cf_final_en.pdf

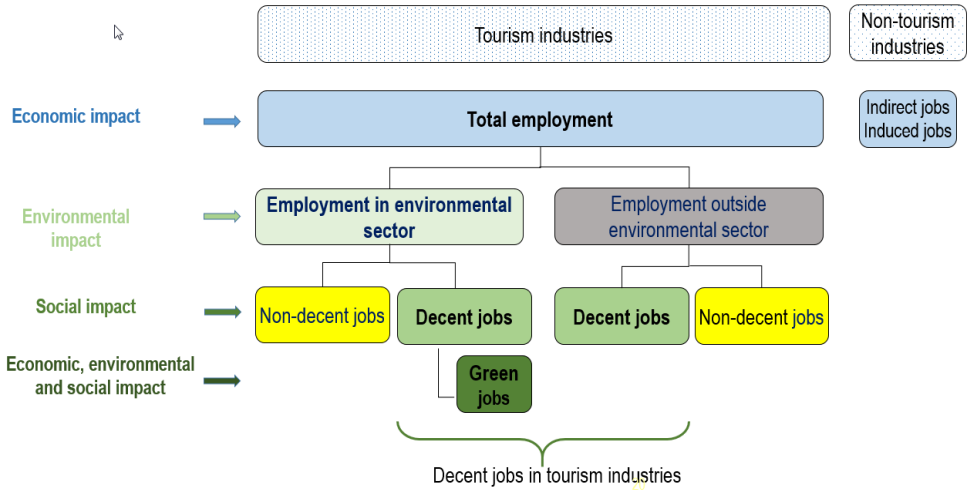
[10] https://unstats.un.org/unsd/publication/seriesf/seriesf_80rev1e.pdf

(i) % of employed in the tourism industries
(ii) % of employed in environmental sector of tourism
(iii) % of jobs in tourism industries that are decent
(iv) % of jobs in tourism industries that are green

While each of the listed indicators (i), (ii) and (iii) can be used for assessing one dimension of sustainability, the indicators (iv) can be used to assess all 3 dimensions of sustainable tourism A major advantage of counting 'green jobs' in tourism industries is that it focuses on all three dimensions the sustainable tourism: economic, environmental and social. It takes into account:

- Job in the tourism industries,
- Jobs that contribute to environmental protection and resource management in tourism,
- Quality of the jobs in tourism.

Figure 1: Employment data for assessing the three dimensions of sustainability of tourism



Other indicators that might be of interest to policy –makers include indirect and inducted employment create thanks to tourism. Employment in these groups will increase with the increase of the tourism activities. These jobs could be estimated through TSA or statistical modelling.

62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur
18-23 August 2019

Green economy: A conceptual overview

Margarita Rohr
University of Valencia, Spain

Abstract

The United Nations system has identified green economy as "investment in sectors such as energy efficiency technologies, renewable energy, public transport, sustainable agriculture, environment friendly tourism and sustainable management of natural resources, including ecosystems and biodiversity" aimed at generating new areas of production, quality jobs and an increase in income, while serving to mitigate climate change and protect biodiversity.

A greener economy, as a way to achieve sustainable development, is not optional for sustainable enterprises and labour markets, it is a necessity. Escalating natural resource use and pollution will compound the growing scarcity of fresh water and fertile land and accelerate the loss of biodiversity and climate change beyond tolerable – perhaps even manageable – levels. The overuse of natural resources, such as forests, fish and clean water, and the rising levels of pollution, including emissions of greenhouse gases (GHGs), are increasingly exceeding planetary boundaries.

In order to understand the nature of green jobs, it is important to understand the environment in which these jobs are being created. It is also true in the case of tourism industries where such jobs should be part of the green tourism economy. Many national and international users are interested in the size of the green economy (in terms of the number of establishments, the number of employees and the total turnover) and in its contribution to economic growth, especially to turnover, employment (number of people employed directly or indirectly, their level of skills and the specialist skills required), value added, investment, exports, etc.

The green economy paradigm is used in this paper both as background and a general framework of discussion.

Keywords

Green economy; Green growth; Sustainable development

1. Introduction

The existing model of economy has allowed a great growth of the world economy and that today millions of people enjoy high levels of wellbeing. Only in the last quarter of a century, the world economy has quadrupled, benefiting

hundreds of millions of people. However, economic growth in recent decades has been achieved by depleting natural resources, allowing the degradation and widespread loss of ecosystems and ignoring many people who, besides living in poverty, depend directly on these resources and systems. According to forecasts by the Organization for Economic Cooperation and Development (OECD), the world will lose in 2050 in comparison with 2000 from 61% to 72% of flora and fauna if maintains the same level of production and consumption, as well as 7.5 million square kilometres will be irreversibly destroyed.

The economic and environmental crisis have the same origin and are reciprocally enhanced due to the current economic model which seeks short-term benefits without considering ecosystems as scarce goods or the consequences generated on the environment and society.

In response to this problem, a new economic paradigm emerges, the green economy, which can contribute to obtaining material wealth without increasing environmental risks, ecological scarcity or social inequality. The concept of green economy is one of the global strategies to confront the economic and environmental crises faced by contemporary societies.

The concept of green economy does not replace the concept of sustainable development, but nowadays it is recognized that, in order to achieve sustainability, it is necessary to change the current brown economic model. Sustainability remains the vital long-term goal, but the green economy is describing a pathway to sustainable development.

In this context, the present work approaches methodologically the conceptualization, objectives, measurement and critiques to this new economic paradigm.

2. What is a green economy?

The concept of Green Economy is not completely a new concept. It was first introduced by the London Environmental Economics Centre in a publication "Blueprint for a Sustainable Economy" in 1989 authored by David Pearce, Anil Markandya, and Ed Barbier. However, the idea of a more sustainable economy appeared before in the report of the Club of Rome "The Limits to Growth" in 1972. At that time the concept did not receive wide acceptance. With the outbreak of the financial crisis in 2007 and the failure of most countries to move onto a sustainable development path, it has become evidently clear that the current development paradigm is not yielding the desired outcomes on all fronts: economic, social, and environmental.

The demand for a new model of sustainable development reappeared in 2009, when the United Nations Environment Program (UNEP) defined the green economy as "one that results in improved human well-being and social equity, and significantly reducing environmental risks and ecological scarcities". The rise and spread of the concept of the "green economy" has

stemmed from the identification of the need to address multiple issues in an integrated way, to overcome these existing interrelated crises and to better avoid any further ones. The green economy looks for the growth of the Gross Domestic Product (GDP) and jobs through shifting investments towards clean technologies and natural capital as well as human resources and social institutions. According to the UNEP, the basic principles of green economy are following:

- Justice and objectivity, both within a single generation and between generations;
- Coherence with the principles of sustainable development;
- A preventive approach to social and environmental impacts;
- Evaluation of natural and social capital, for example, the internationalization of external costs, green accounting, costs over the entire life cycle and improved management;
- Sustainable and efficient use of resources, consumption and production;
- The need to achieve existing macroeconomic goals through the creation of green jobs, poverty eradication, increased competitiveness and growth in key sectors.

Additionally, the OECD has developed and introduced the concept of "green growth", defining it as the maximum guarantee of economic growth and development, without affecting the quantity and quality of natural assets and using the growth potential that arises during the transition to a green economy. That is, "green growth" is GDP growth, which is subject to "green" conditions and focuses on "green" sectors as new growth engines.

It is important to say that there is no single internationally-accepted definition of the green economy until now. Besides the concepts mentioned above, other international organizations created their own definitions of green economy and green growth, highlighting the following:

- *Global Green New Deal (2009)*: "The economic crisis provides the opportunity to introduce a global green new deal, which consists in stimulating the economy towards the development of green sectors, green infrastructure and green jobs. A greening infrastructure is a process of transforming a business activity such that it reduces emissions of greenhouse gases and consumption of resources, produces less waste, and reduces social inequalities at the same time, ensuring the return on natural, human and economic capital".
- *World Bank (2012)*: "Green growth is an effective growth in terms of using clean resources, i.e. reducing pollution and environmental degradation, resistant to natural hazards and using environmental management to prevent other disasters".

- *Global Green Growth Institute (2012)*: "Green growth is the new revolutionary development paradigm that sustains economic growth while at the same time ensuring climate and environmental sustainability. It is aimed at reducing poverty, creating jobs, social integration and the sustainability of ecosystems, alleviating climate changes, supporting biodiversity, and providing access to clean energy and water".
- *European Environmental Agency (2012)*: "Green economy is an economy where environmental, economic and social policies and innovations support societies in the effective use of resources, while at the same time improving human well-being, accentuating social integration and protecting the natural systems which sustain life on the Earth".

The universal concept of green economy is still in the process of discussion and specification, and therefore there are numerous perceptions of it. It has been noted however that "while interpretations of the "green economy" vary to some degree, there is much common ground between the concepts employed by governments, businesses and international organisations globally. Basically, a green economy implies a departure from the "business as usual" economic paradigm, to one with regulatory measures and strong financial incentives for innovation, investments (for example, in green)" (Green Economy Report, 2010).

Building on UNEP's report 'Towards a Green Economy', in order to empower and achieve a green economy, an annual investment of 2% of global GDP is required. It would allow to maintain the current rate of economic growth and at the same time achieve changes towards sustainable processes. Additionally, it is necessary for countries to promote fiscal incentives in areas that stimulate a green economy (e.g. in technology, infrastructure or infant industries), establish control measures and introduce economic instruments to help conserve natural resources (Kumar, 2017). All the investments along with the political reforms, should promote the transformation of the sectors involved in the green economy, so that they acquire a competitive position in the long term (Gehring, 2016, Biswas & Roy, 2015). Depending on their current level of development, countries have different capacities to initiate and implement policy reform and cope with transformative change. Other supporting actions are therefore needed to increase capacity and strengthen institutions, provide training and skill enhancement to the workforce, and improve general education on sustainability.

Green economy is both a challenge and an opportunity for the labour market, which, in turn, is a major factor in potential green growth. Response dynamics and the good functioning of labour markets play a key role in facilitating the transition to a green and resource-efficient economy. The

transition to a sustainable economy leads to changes, some of them quite serious, in the employment structures and professional profiles of the workers.

3. Measuring progress towards the green economy

To ensure effective policy design supporting green economy and green growth, rigorous information and data on the environment and economy nexus are needed. Not only does following and assessing progress in green economy contribute to a better understanding of the determinants of green growth but can also point out further synergies between the environmental protection and economic growth. It is also important to have a solid information base to better communicate progress on green economy with citizens.

International organizations have taken numerous initiatives on measuring progress towards the green economy, including the following:

- The green growth indicators of OECD
- The green economy indicators of UNEP
- The United Nations Statistics Division's Project for "Strengthening the capacities of developing countries to measure progress towards a green economy", 2015-2016
- The International Labour Organization (ILO)
- The green industry initiative of the United Nations Industrial Development Organization (UNIDO)
- The United Nations Sustainable Development Goals (SDGs)
- The Latin American and Caribbean Initiative for Sustainable Development (ILAC).

The choice of indicators is one of the most important and, at the same time, the most difficult tasks since their quality has a direct impact on the reliability of the final classification and the accuracy of assessments based on the obtained results. The problem of proper selection of indicators is one of the most important factors deciding on the quality and reliability of assessments of the green economy.

The choice of indicators for the measuring a green economy is also key to the evaluation of its practical implementation. Usually, the list of available indicators is relatively long. The research dilemma lies in the choice of the most suitable subset, as the set of potential indicators is practically infinite.

Measurement should include both the assessment of the current environmental situation and the external impact resulting from human activities and governmental policies designed to promote a green economy. Each individual objective can be represented by several or, in some cases, several dozens of indicators.

The green economy indicators are a specific group. Much like the set of indicators of sustainable development, they should not only provide a good

representation of several distinctive areas of human life (economic, social, environmental), but also show the correlations between those areas, while at the same time providing a balanced and representative illustration of the fundamental aspects of the "greenification" process.

4. Critique of the green economy

Civil society groups and governments are critical of the transition to a green economy, taking into account that it does not adequately or clearly address the social, economic and ecological aspects, pillars of sustainable development (Geng et al., 2017). On the contrary, it can become a new framework for sustainable development, replacing the three pillars mentioned (Loiseau et al., 2016). From the United Nations Environment Program, it is stated that "the achievement of sustainability depends considerably on the adequacy of the economy", which makes it necessary to examine the concept of a green economy and the way in which the concept of a green economy would promote economic, ecological and social sustainability within this.

Another criticism that appears is the economistic character of the green economy, which, although based on the production of more sustainable sectors which reduce environmental problems, continues to have an economic pattern of accumulation and infinite growth (Lander, 2011; Karakul, 2016). For Droste et al. (2016), overcoming the current economic order, would imply the need to convert economic production to physical terms, so that the finite capacity of natural resources and assimilation of the waste of human activity on the planet becomes evident.

Unmüßig et al. (2012) and Diyar et al. (2014), state that the green economy is an inappropriate term, scientific and philosophical misunderstanding, which will not achieve sustainable development and the eradication of poverty. Also Montefrio & Dressler (2016) add that it was created from ambiguities, without scientific or philosophical support, and that on the contrary it will legitimize the opening of markets, create more tension with the ecological and cultural diversity of the planet and of humanity.

5. Conclusion

The green economy becomes a model which promotes growth, the creation of income and jobs, in particular "green jobs", which seeks to generate a change in the interaction between economic progress and environmental sustainability, mostly if wealth is measured account for natural assets and not just productivity. In addition, the green economy also contributes substantially to reducing social inequality among countries and eradicating poverty in the world.

It should be noted that the implementation of the green economy can achieve technological changes, which allow the adoption of environmentally

sustainable strategies, make use of natural resources in a responsible manner, as well as the waste of their activity can be reincorporated into the production process, decreasing the causes of pollution.

Nevertheless, to achieve the objectives of green economy it is necessary to accept and develop the proposed alternatives by both developed and developing countries, through the allocation of necessary economic resources, greater stringency in environmental regulations, creation of subsidies to environmentally friendly activities, as well as the optimization of the planning processes of the territory. Likewise, it is necessary to create a new economic framework that allows countries to coordinate on the same level, without losing sight of the fundamental premises of sustainable development.

Finally, the green economy seeks within its objectives the eradication of poverty and the inclusion of vulnerable social sectors, to achieve sustainable economic development in terms of maintenance of a healthy environment and the proper use of ecosystems both for the present generation and for future generations.

References

1. Biswas, A. & Roy, M. (2015). Green products: an exploratory study on the consumer behaviour in emerging economies of the East. Journal of Cleaner Production, 87, 463-468.
2. Diyar, S., Akparova, A., Toktabayev, A. & Tyutunnikova, M. (2014). Green Economy–Innovation-based Development of Kazakhstan. Procedia-Social and Behavioral Sciences, 140, 695-699.
3. Droste, N., Hansjürgens, B., Kuikman, P., Otter, N., Antikainen, R., Leskinen, P. & Thomsen, M. (2016). Steering innovations towards a green economy: Understanding government intervention. Journal of Cleaner Production, 135, 426-434.
4. Gehring, M. (2016). La Transición Legal a una Economía Verde. Revista de Derecho Ambiental, 6, 8-43.
5. Geng, R., Mansouri, S. A., Aktas, E. & Yen, D. A. (2017). The role of Guanxi in green supply chain management in Asia's emerging economies: A conceptual framework. Industrial Marketing Management, 63, 1-17.
6. Karakul, A. K. (2016). Educating labour force for a green economy and renewable energy jobs in Turkey: A quantitave approach. Renewable and Sustainable Energy Reviews, 63, 568-578.
7. Kumar, P. (2017). Innovative tools and new metrics for inclusive green economy. Current Opinion in Environmental Sustainability, 24, 47-51.
8. Lander, E. (2011). Informe: La economía verde: el lobo se viste con piel de cordero. Transnational Institute. 10p. https://www.tni.org/files/download/green-economy_es.pdf.

9. Loiseau, E., Saikku, L., Antikainen, R., Droste, N., Hansjürgens, B., Pitkänen, K. & Thomsen, M. (2016). Green economy and related concepts: An overview. Journal of Cleaner Production, 139, 361-371.
10. Montefrio, M. J. F. & Dressler, W. H. (2016). The Green Economy and Constructions of the "Idle" and "Unproductive" Uplands in the Philippines. World Development, 79, 114-126.
11. OECD (2011). Towards Green Growth, http://www.oecd.org/greengrowth/48224539.pdf
12. The Environment Report 2012 (2012). Responsibility in a Finite World, German Advisory Council on the Environment, http://www.umweltrat.de
13. The Global Green Economy Index (2012). Dual Citizen Inc., http://www.dualcitizeninc.com/ ggei2012.pdf.
14. Unmüßig, B., Fatheuer, T. & Sachs, W. (2012). Crítica a la Economía Verde Impulsos para un futuro social y ecológicamente justo. Ed: Fundación Heinrich Böll. 46p. https://mx.boell.org/sites/default/files/gruene_oekonomie_.pdf
15. UNEP, (2009), A Global Green New Deal, Geneva.
16. UNEP (2011). Towards a Green Economy: Pathways to Sustainable Development and Poverty Eradication, www.unep.org/greeneconomy.
17. World Bank (2012). Inclusive Green Growth: The Pathway to Sustainable Development, Washington, DC.

62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur
18-23 August 2019

Greening with jobs

Catherine Saget
International Labour Office (ILO)

Abstract

Twenty-four million new jobs will be created globally by 2030 if the right policie to promote a greener economy are put in place. Action to limit global warming to 2 degrees Celsius will result insufficient job creation to more than offset job losses of 6 million elsewhere. New jobs will be created by adopting sustainable practices in the energy sector, including changes in the energy mix, promoting the use of electric vehicles and improving the efficiency of buildings.

Ecosystem services, including air and water purification-soil renewal and fertilization, pest control, pollination and protection against extreme weather conditions-sustain, among others, farming, fishing, forestry and tourism activities, which employ 1.2 billion workers.

But projected temperature increases will benefit from net job creation: of the 163 economic sectors analysed, only 14 will suffer employment losses of more than 10,000 jobs worldwide. Only two sectors, petroleum extraction and petroleum refining, show losses of 1 million or more jobs. 2.5 million jobs will be created in renewables-based electricity, offsetting some 400,000 jobs lost in fossil fuel-based electricity generation.

Although measures to address climate change may result in short-term employment losses in some cases, their negative impact can be reduced through appropriate policies.

Urgent action to train workers in the skills needed for the transition to a greener economy needs to be taken and provide them with social protection that facilitates the transition to new a jobs, contributes to preventing poverty and the vulnerability of households and communities.

Low- and some middle- income countries still need support to develop data collection, and adopt and finance strategies towards a just transition to an environmentally sustainable economy and society that includes everyone from all groups of society.

Executive summary

1. Action to limit global warming to 2°C will create jobs

The long-term goal of the 2015 Paris Agreement is to keep the increase in global average temperature to less than 2°C above pre-industrial levels. The

agreement aims to help countries meet this target and strengthen societies' capacities to address the wide-ranging impacts of climate change. The employment projections in this report suggest that the net effect on job numbers will be positive. The transition to a green economy will inevitably cause job losses in certain sectors as carbon- and resource-intensive industries are scaled down, but these will be more than offset by new job opportunities. Measures taken in the production and use of energy, for example, will lead to job losses of around 6 million, as well as in the creation of around 24 million jobs. The net increase of approximately 18 million jobs around the world will be the result of the adoption of sustainable practices, including changes in the energy mix, the projected growth in the use of electric vehicles, and increases in energy efficiency in existing and future buildings. In order to ensure a just transition, efforts to promote the green economy must be accompanied by policies that facilitate the reallocation of workers, advance decent work, offer local solutions and support displaced workers.

2. A transition to agricultural sustainability and a circular economy will result in more and often better jobs …

The adoption of more sustainable agricultural policies can create wage employment in medium and large organic farms, and allow smallholders to diversify their sources of income through a transition to conservation agriculture. With complementary policies to support workers, adopting conservation agriculture can help sustain a structural transformation in developing countries. In parallel, embracing a circular economy that emphasizes the reuse, recycling, remanufacture and repair of goods will create around 6 million new employment opportunities across the world, as such policies replace the traditional model of "extract, make, use and dispose".

3. The transition is urgent, given the unsustainable pressure of current economic activity on the environment

Important progress was achieved during the period between 2000 and 2015 in the global economy and in the promotion of decent work, especially in the form of a reduction in working poverty and child labour. But wage growth has stagnated and, to a large extent, inequality has risen. Moreover, it is striking that, in a context of scarce resources and limited ability to absorb waste, current patterns of economic growth rely largely on the extraction of resources, manufacturing, consumption and waste. In 2013, for example, humanity used 1.7 times the amount of resources and waste that the biosphere was able to regenerate and absorb. Indeed, human activity has already caused irreversible environmental change on a global scale.

4. Jobs rely heavily on a healthy and stable environment and the services it provides …

From a jobs perspective, environmental sustainability is critical. In fact, the increasing frequency and intensity of natural disasters associated with human activity have already reduced productivity. Annually, between 2000 and 2015, natural disasters caused or exacerbated by humanity resulted in a global loss of working lives equivalent to 0.8 per cent of a year's work. Looking ahead, projected temperature increases will make heat stress more common, reducing the total number of working hours by 2.0 per cent globally by 2030 and affecting above all workers in agriculture and in developing countries. The damage associated with unmitigated climate change will therefore undermine GDP growth, employment, and working conditions. Local air, water and soil pollution and other forms of environmental degradation negatively affect workers' health, income, food and fuel security, as well as their productivity. The adoption of specific policy measures can reduce its negative impact, including occupational safety and health measures, social protection policies and other actions designed to adapt to a changing environment.

5. … which highlights the urgency of the transition to environmental sustainability for the world of work

Currently, 1.2 billion jobs rely directly on the effective management and sustainability of a healthy environment, in particular jobs in farming, fishing and forestry relying on natural processes such as air and water purification, soil renewal and fertilization, pollination, pest control, the moderation of extreme temperatures, and protection against storms, floods and strong winds. Environmental degradation threatens these ecosystem services and the jobs that depend on them. The effects of environmental degradation on the world of work are particularly acute for the most vulnerable workers. Workers from lower-income countries and Small Island Developing States, rural workers, people in poverty and other disadvantaged groups are affected the most by the impact of climate change. The transition to a green economy is not only urgent for the sake of the planet, but is also compatible with improvements in decent work. A key finding of this report is that many countries have succeeded in improving labour market outcomes while at the same time decoupling growth from carbon emissions.

6. Complementary policies can promote employment and mitigate the effects of climate change

Although climate change mitigation measures may result in short-term employment losses, their negative impact on GDP growth, employment and inequality can be reduced through appropriate policies. Climate change mitigation could reduce slightly the share of women in total employment, as

employment gains associated with the 2°C scenario create jobs in currently male-dominated industries (renewables, manufacturing and construction), unless action is taken to reduce occupational segregation. Coordination between the social partners can reduce inequality and promote efficiency gains, while coordination at the international level is necessary to achieve meaningful reductions in emissions. Certain mitigation policies (such as limiting the increase in temperature, for example by promoting renewable energy) may act as an incentive for enterprises to develop and adopt more efficient technology, thereby boosting employment in key occupations, as well as productivity. Adaptation policies (e.g. converting to resilient agriculture practice) can also create jobs at the local level.

7. The legal framework can provide incentives for the greening of the economy, while ensuring decent work

Legal standards can promote progress towards decent work during and beyond the transition to environmental sustainability. By virtue of their broad acceptance and universal relevance for workers, workplaces and the various sectors, international labour standards provide a social pillar for the green economy and can help to ensure that emerging sectors offer decent working conditions. In addition, ILO standards on occupational safety and health contribute to the preservation of the environment. The Indigenous and Tribal Peoples Convention, 1989 (No. 169), which requires environmental impact assessments to be carried out in relation to development activities that may affect that population, the Prevention of Major Industrial Accidents Convention, 1993 (No. 174), and the Employment and Decent Work for Peace and Resilience Recommendation, 2017 (No. 205), among others, address environmental issues directly.

Multilateral environmental agreements (MEAs), which are binding agreements between States dealing with environmental matters, increasingly include labour dimensions, such as the importance of environmental rights at work, employment protection and promotion. They place particular emphasis on occupational safety and health standards. At the national level, environmental legislation and policies are increasingly incorporating labour issues. In 19 of the 26 national legal frameworks reviewed for this report, climate change policies contain labour considerations, including complementary skills policies and job creation. Sector-specific environmental legislation also tends to cover employment and decent work issues. The strong links between environmental regulation and labour issues are also more and more evident in sub-Saharan Africa in the renewable energy and waste management sectors.

8. Social dialogue contributes to ensuring that the green transition is a just transition

Social dialogue has contributed to making environmental governance more labour-friendly by promoting frameworks, legislation and policies that include both labour and environmental concerns. This illustrates the priorities established by the UN Agenda for Sustainable Development and the principles embedded in international labour standards, including the importance of consultation and collective bargaining. At the international level, international framework agreements (IFAs) are voluntary agreements between multinational enterprises and global union federations. Of the 104 IFAs reviewed for this report, 61 include environmental provisions on such issues as respect for the environment as a corporate responsibility and waste management measures, particularly in the manufacturing, energy, mining and automotive industries. At the national and enterprise level, while the number of collective agreements containing green clauses is still limited, they are used by employers and workers to reconcile social and economic objectives with environmental concerns. Emerging examples indicate that workers and employers, through social dialogue, have identified areas where the environmental impact could be mitigated without reducing or negatively affecting employment or working conditions. In the longer term, the protection of environmental rights at work could also be strengthened in national policies and legislation.

9. Synergies between social protection and environmental policy can support both workers' incomes and the green transition

Social protection systems are the first line of protection against the negative effects on income of different risks, including those stemming from climate change and local environmental degradation. They support the economy by stabilizing household incomes. Four policy areas offer particular synergies between social protection and environmental sustainability: unemployment protection, cash transfer programmes, public employment programmes (PEPs) and payments for ecosystem services (PES).

Unemployment protection schemes and cash transfer programmes play a critical role in supporting workers facing job loss related either to the transition to environmental sustainability or to a natural disaster. They facilitate the transition to new jobs, particularly when combined with skills development and job placement or relocation measures. In addition, access to safe and regular labour migration opportunities can foster economic diversification and increase adaptive capacity through remittance and skills transfer. Cash transfer programmes contribute to preventing poverty and reducing the vulnerability of households and communities.

PEPs too can be powerful tools to address the impact of climate change on workers and their incomes, while also enhancing mitigation. Half of the 86 PEPs in 62 countries surveyed include an environmental component. They often provide health care, education and other benefits. Similarly, PES, although originally conceived with an environmental objective, can provide effective support for household incomes in specific circumstances.

A policy mix comprising cash transfers, stronger social insurance and limits on the use of fossil fuels could lead to faster economic growth, stronger employment creation and a fairer income distribution, as well as lower greenhouse gas emissions.

10. Although skills development programmes for enterprises and workers facilitate the transition to a green economy, they are yet to be mainstreamed in policy discussions

Skills development programmes are crucial to the achievement of a just transition. Of the 27 countries surveyed, about two-thirds have established platforms to anticipate skills needs and the provision of training in general, but they are not all used to discuss the skills implications of the green transition. The active participation of social partners is useful in identifying skills gaps, implementing training provisions, emphasizing that higher skills translate into higher pay, and recognizing the skills acquired on the job. However, social partners are not always involved in the relevant discussions; this is especially the case of workers. Where they exist, specific bodies to discuss skills for the green transition have led to positive changes in training for the sectors directly involved in the transition (such as renewable energy and waste management), but they have comparatively little influence on the greening of the economy as a whole.

National environmental legislation increasingly refers to skills development but the provisions are often limited to specific skills policy areas (such as the identification of skills needs), target groups (e.g. youth), sectors (especially energy) or regions. Consensus has not yet been reached in many countries on the definition of skills for the green transition and the capacity is lacking to collect relevant data for reliable skills identification. As a result, skills development policies for the green transition tend to adopt a short-term and fragmented approach. Greater awareness of environmental issues and their mainstreaming in skills policy discussions are required to ensure that identification of skills needs and implementation of training programmes respond to labour market needs.

11. Institutions, policy-making and effective implementation are key for a just transition

Social dialogue, the elimination of discrimination in employment and occupation, and good governance are the foundations of an effective and just transition. For example, the involvement of central and local governments, social partners and NGOs in debates on climate change at the national level has led to the integration of economic, social and environmental objectives. Tax reform can support the transition to a green economy, while at the same time facilitating employment creation.

Low-income and some middle-income countries need support to develop data collection, identify and adopt best practices, strengthen implementation and finance both mitigation and adaptation strategies in order to achieve a just transition to environmentally sustainable economies and societies for all. A just transition requires identifying and implementing policy solutions to some of the most pressing challenges to the future of work that also affect climate change, such as employment and working conditions in the rural economy, demographic shifts and globalization.

12. A just transition offers enhanced potential for decent job creation through the integration of labour and environmental issues

This ILO report quantifies job losses and job creation in the transition to a green economy, based on projections to 2030 founded on the agreed policy goal of limiting global warming to 2°C. More generally, it finds that the greening of economies can have a positive overall effect on growth and jobs. Positive employment outcomes will also probably apply in the 1.5°C scenario, as encouraged by the Paris Agreement.

The report shows that environmental laws, regulations and policies that include labour issues offer a powerful means of integrating elements of the Decent Work Agenda with environmental objectives. This is true for social protection programmes, skills development programmes, macroeconomic policy and the legal framework. Though some degree of integration is observed in all these domains, it is not yet systematic and not yet universal. For example, while environmental legal frameworks can be effective in combining some elements of the Decent Work Agenda with environmental objectives, the respective provisions often focus on particular groups of workers (such as additional support for local communities, training in areas that are key for the transition, and the protection of workers in specific sectors). The transition affects all workers, however; the universality of rights and protection therefore remains important in order to ensure that the transition delivers inclusive growth and decent work.

Sustainable grassroots tourism through green jobs: Measurement issues

Igor Chernyshev[1], Florabela Carausu[2*]
[1] *GJASD International*, Chancy (Geneva), Switzerland
[2] *GOPA Luxembourg S.A R.L.*, Luxembourg

Abstract
Sustainable tourism takes full account of its current and future economic, social and environmental impacts, addressing the needs of host communities, visitors, the environment and the industry.
The grassroots or local perspective is a high profile focus of sustainable tourism discussion. Therefore, the area of common interest is whether a host community is heavily affected due to the negative impact of tourism activity on its habitat including fauna and flora. Improving life quality of local populations at communal level is considered as a solution to protect natural environment and local bio-diversity.
For the successful and sustainable development of tourism at the local level in settlements adjacent to or located directly in the territory of national parks, plantations, protected forests and riverbanks, etc., it is necessary to encourage the local population to take care of natural resources and biodiversity. A practical way to achieve this goal could be the greening of tourism and creation of green jobs for the ingenious population. Thus, the ratio of green jobs could serve as an indicator of tourism suitability at the local and national levels.

Keywords
Sustainable tourism; green economy; employment; decent work; green jobs

1. Introduction

At the 2016 G20 Tourism Ministers Meeting, G20 leaders underlined that "tourism is one of the main sectors driving economic globalization, interconnection, integration and socio-economic development". It is a "driving force for social inclusion with a particular potential to advance employment and economically empower groups which are more vulnerable to social and economic risks, including, but not limited to, women, young people, migrants, indigenous and tribal peoples and rural residents", G20 China (2016).

Consequently, tourism can lead to the reduction of poverty and to the promotion of socio-economic development and decent work. However, if tourism does not respect local cultures and is uncontrolled, unsustainable or

not socially accountable, it can also have a negative impact on local communities, their heritage and environment, exacerbating inequalities.

2. Methodology

2.1 Sustainable tourism and employment

According to the UNWTO, *sustainable tourism takes full account of its current and future economic, social and environmental impacts, addressing the needs of host communities, visitors, the environment and the industry*. It should generate local prosperity, decent work, promote environmental awareness, conserve and protect the environment, respect wildlife, flora, biodiversity, ecosystems and cultural diversity, and improve the welfare and livelihoods of local communities, including those of women, by supporting their local economies and the human and natural environment as a whole, UNWTO and UNEP (2005).

An important element in the UNWTO definition of sustainable tourism is the fact that tourism is about addressing not only environmental issues, but also to ensure inclusive economic growth and social development. Consequently, the social dimension is a key element in tourism planning and management: poverty, employment, wages, education, skills, changes in host populations, living conditions, characteristics of tourism employees' households, are relevant issues for tourism sustainability. Stemming from the above, the social dimension of sustainable tourism embraces employment. One more important element is directly associated with the above definition: decent work.

Decent work sums up the aspirations of people in their working lives. It involves opportunities for work that is productive and delivers a fair income, security in the workplace and social protection for families, better prospects for personal development and social integration, freedom for people to express their concerns, organize and participate in the decisions that affect their lives and equality of opportunity and treatment for all women and men (ILOb).

2.2 Green jobs and the ILO

Statistics on green jobs is of interest to a wide variety of users: the general public, media and civil society, decision and policy makers concerned with economic growth, job creation, environmental protection, climate change and sustainability, as well as analysts, experts and advisors, academics, training institutions, government officials and international agencies.

In order to provide a clear statistical definition of green jobs that would facilitate the production of internationally comparable data, the *Guidelines concerning statistical definition of employment in the*

environmental sector and green jobs were endorsed by the 19th ICLS, ILO (2013).

In the broader definition currently used by the Green Jobs Programme of the ILO: "Jobs are green when they help reduce negative environmental impact ultimately leading to environmentally, economically and socially sustainable enterprises and economies. More precisely green jobs are decent jobs that: reduce consumption of energy and raw materials; limit greenhouse gas emissions; minimize waste and pollution; and protect and restore ecosystems", (ILOa).

2.3 Green jobs and Statistical Framework for Measuring Sustainable Tourism

Despite the long-standing interest and discussion of sustainable tourism and the important advances in tourism statistics, there is as yet no standardised basis for the collection of relevant information, at either the national or subnational level. This is a significant gap, and one that limits the potential for the development of policies directed at advancing sustainable tourism. In order to fill this gap, UNWTO, with the support of the UN Statistics Division (UNSD), has initiated the project Towards a Statistical Framework for Measuring Sustainable Tourism (SF-MST). The SF-MST aims at expanding the measurement of tourism, which currently is mostly focused on economics, to include also environmental and social considerations (employment in particular).

The SF-MST Sub-Group on Tourism Employment focusing on tourism employment such as:

- Green jobs
- Human capital (availability of skills and experience)
- Decent work
- Others

In short, the Sub-group identifies interrelationships exiting among green economy, green jobs and sustainable tourism.

2.4 Green Economy

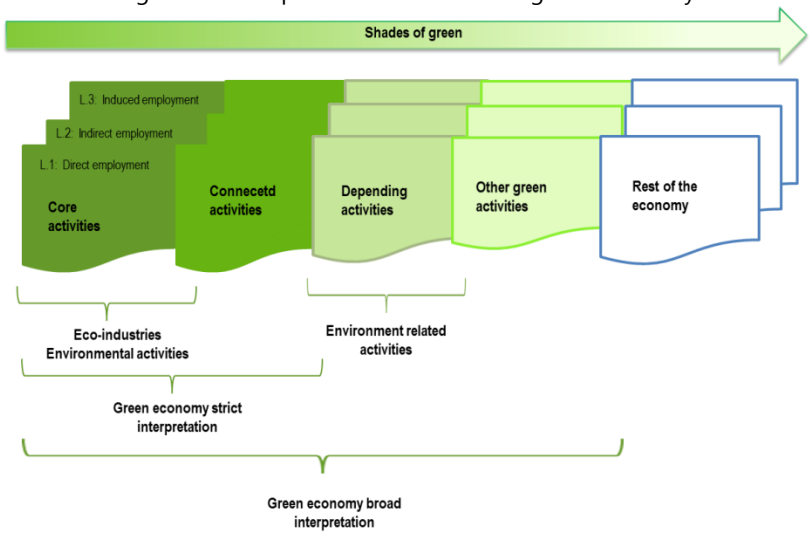
This paper discusses the issue of measuring sustainable grassroots tourism through green jobs that directly relates with the environmental dimension of sustainability, which is a multifaceted phenomenon. Green economy is a subset element of the environmental dimension of tourism sustainability.

According to Jon Rynne, the economy is an ecosystem which consists of green or eco-industries, Rynn (2007). He argues that green jobs will reinvigorate the economy, creating entirely new green-collar job sectors. EUROSTAT and OECD define an eco-industry as: "activities which produce

goods and services to measure, prevent, limit, minimize or correct environmental damage to water, air and soil, as well as problems concerning waste, noise and eco-systems. This includes technologies, products, and services that reduce environmental risks and minimise pollution and the use of resources" OECD/EUROSTAT (1999).

The conceptual framework of the green economy shown in Figure 1 contains the shades of green that are based on the intended impact on the environment.

Figure 1: Conceptual framework of the green economy



Source: IDEA Consult.

Particular attention requires the impact of green economy on employment Level 1: direct activity; Level 2: indirect activity; and Level 3: induced activity.

In turn, these three types of activities refer to the three types of employment in the tourism industries: (a) direct employment[1]; (b) indirect employment in the sectors supplying inputs to the tourism industries; and (c) induced effect on employment as a result of subsequent rounds of spending.

2.5 Grassroots or local perspective of tourism sustainability through green jobs creation

In the context of this paper, we put the grassroots or local perspective as a high profile focus of sustainable tourism discussion. Therefore, the area of common interest is (a) whether a host community is heavily

[1] Jobs in tourism industries that can be attributable to tourism spending plus jobs in non-tourism industries that can be directly attributed to tourism spending

affected due to the negative impact of tourism activity on its habitat including fauna and flora; (ii) how to involve a host community in tourism related activities as a means to bring them out of poverty and stimulate participation in preserving their habitat and the nature.

Improving life quality of local populations at communal level is considered as a solution to protect natural environment and local bio-diversity. In many cases, the rural population live in self-built houses lacking basic facilities, tools, access to financial resources and sustainable infrastructure. In order to improve the quality of life of local people, it is essential providing them with assistance in reconstruction of their habitat in order to make these areas attractive for eco- or green tourism.

According to the UNWTO, sustainable tourism should ensure viable, long-term economic operations, providing socio-economic benefits to all stakeholders that are fairly distributed, including stable employment and income-earning opportunities and social services to host communities, and contributing to poverty alleviation (UNWTO).

A possible solution supporting successful and sustainable development of tourism at the local level in settlements adjacent to or located directly in the territory of national parks, plantations, protected forests and riverbanks, etc., would to encourage the local population to take care of natural resources and biodiversity. A practical way to achieve this goal could be the greening of tourism and creation of green jobs for the ingenious population. Thus, the ratio of green jobs can serve as an indicator of tourism suitability at the local and national levels.

To recall, jobs are green when they help reduce negative environmental impact ultimately leading to environmentally, economically and socially sustainable enterprises and economies.

2.5.1 Measurement issue

Employment in the environmental sector and green jobs can be estimated by using data from inventories, regular statistical surveys and censuses; and specialized statistical modules, surveys and censuses, including subsample surveys (Stoevska, 2019).

In addition to the methodology recommended by the ILO, we suggest considering a simplified approach to the measuring green jobs among the rural and ingenious population involved in activities of the tourism sector with the objective to stimulate environmental protection and poverty reduction. As a first step and as a compensation for their activities supporting agro- or eco-tourism they could be offered within the framework of relevant projects such as amenities as:

- Improving the thermal insulation of family homes, installation of heating systems with renewable energies, solar windows and panels, bioenergy (eco-efficient cooking stoves), etc.
- Installation of equipment for energy efficiency and production of renewable energies such as solar cells (solar kits), solar fuels or fuel cells.
- Promotion of closing the loop eco-principle solutions such as agro-food biomass and waste reduction, etc.

The above efforts would make it possible to develop sustainable tourism facilities, routes and individual itineraries.

Given that in their great majority the above activities would contribute to the environmental protection, or in aliis verbis would help reduce negative environmental impact ultimately leading to environmentally, economically and socially sustainable [local] economy, they could be considered as green jobs.

In our view, most jobs referred to above would also contribute to the economic dimension of sustainable tourism and by their nature they predominantly would be counted as indirect or induced green jobs in the tourism industries.

Evidently, it would not be easy under circumstances to distinguish between decent and not decent green jobs.

2.5.2 Data collection sources and methods

The main sources of information that may be used to assess how many green jobs exist in tourism industries (activities) are censuses, labour force and establishment surveys, administrative records (although the latter may not be sufficiently efficient in this particular case). To optimise the resource used, it is suggested that, as far as possible, the required data be collected by extending existing surveys rather than by initiating totally new ones. The possibility of incorporating new questions or modules in existing, on-going or planned surveys should be explored in order to fill the data gaps.

Depending on national priorities, data collection could be focused on key tourism activities (e.g., the largest in terms of their contribution to the provision of environmentally-friendly goods and services) and/or on those that have the greatest potential to change. A pragmatic approach could be to focus on some resource management subsectors (green tourism, agro-tourism, ecotourism, green resorts and green or eco-hotels, national parks, sports and recreation facilities, environmentally–friendly transport and catering, etc.) where clear benchmarks exist (e.g., specific labels).

3. Case of Vanuatu: employment and environmental sustainability

Vanuatu is an archipelago of 83 islands stretching over more than 1,000 km in the southern Pacific Ocean. Its population is mostly rural and growing. A large majority of employment is medium-skilled occupations (ILO 2017).

In 2009, 60.5 per cent of employment was in the agricultural, forestry and fishing sector (Fig. 8). The transition to a low-carbon and resource efficient economy will require a significant expansion of employment in a number of green economic activities, such as those related to resource management or environmental services (for example, waste management and reforestation.

In 1994, the country's municipal solid waste generation was 3.28 kg per capita per day. A significant proportion of the waste is organic (at 71 per cent), followed by recyclable material, such as paper, glass and plastics (at 22 per cent). There are opportunities to create decent work with "safe" composting and recycling for local communities.

In 2014, only 16 per cent of the population relied primarily on clean fuel and technology, in the sense that they do not create indoor pollution within the home. The share of renewable energy in total energy consumption has fluctuated since 2000, peaking in 2006 at 69.6 per cent before dropping to 32.4 per cent in 2014. Renewable energy generation increased between 2011 and 2013 In 2009, employment in the electricity, gas, steam and air conditioning supply sector was 0.2 per cent. There is a notable green tourism potential and improvements in utilities can potentially provide benefits on three fronts: community health, environmental health) and the economy, with increased employment opportunities.

Better data on green and decent jobs is particularly needed to assess the impact of climate change and climate-related policies on social inclusion. Without better data, it will be difficult to determine what policy changes are needed to assure a just transition to environmental sustainability and to monitor progress going forward.

4. Conclusion

The case of Vanuatu is just one example of a country whose indigenous population could be involved with certain investments in preservation of environment and protection of bio-diversity providing them steady income and economic development through, inter alia, eco-tourism. As mentioned in earlier sections, people involved in this work could be considered employed in green jobs.

We agree that this approach may look quite simplistic, but it has the advantage of direct application with minimal costs using the traditional methods of data collection that could already be in place.

Finally, we would also like to note that in countries such as Vanuatu, rapid assessments of the potential for creating green jobs are needed to develop

realistic economic and social policies. Given the structure of the economy and the country's capabilities, eco-tourism could become a sector capable of improving life quality of local populations in a relatively short period of time.

References

1. G20 China (2016). 7th G20 Tourism Ministers Meeting Beijing, China, 20 May 2016. Declaration "Sustainable Tourism – An Effective Tool for Inclusive Development; available at: http://www.mofa.go.jp/mofaj/files/000205641.pdf
2. OECD/EUROSTAT (1999). The Environmental Goods and Services Industry. Manual for Data Collection and Analysis, Paris, 1999; available at: https://unstats.un.org/UNSD/envAccounting/ceea/archive/EPEA/EnvIndustry_Manual_for_data_collection.PDF
3. International Labour Organization (ILO, 2013). Report of the Conference. 19th International Conference of Labour Statisticians (Geneva, 2–11 October 2013.) Guidelines concerning a statistical definition of employment in the environmental sector; see available at: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_234124.pdf
4. International Labour Organization (ILO, 2017). Vanuatu employment and environmental sustainability fact sheet 2017; available at: https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/documents/publication/wcms_627570.pdf
5. International Labour Organization (ILOa). The Green Jobs Programme of the ILO; available at: https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_ent/documents/publication/wcms_371396.pdf
6. International Labour Office (ILOb). Decent Work; available at: http://www.ilo.org/global/topics/decent-work/lang--en/index.htm
7. Rynn, Jon (2007). The economy is an ecosystem; available at: http://grist.org/article/rebuild-the-economy-by-building-green-industries/
8. UNWTO and UNEP (2005). Making Tourism More Sustainable – A Guide for Policy Makers. UNWTO, Madrid and UNEP, Paris 2005; available at: http://www.unep.fr/shared/publications/pdf/DTIx0592xPA-TourismPolicyEN.pdf
9. UNWTO. Sustainable Development of Tourism; available at: http://sdt.unwto.org/content/about-us-5

Central limit theorem and bootstrap procedure for Wasserstein's variations with an application to structural relationships between distributions

Eustasio del Barrio[1], Paula Gordaliza[1], Hélène Lescornel[2], Jean-Michel Loubes[2]
[1]IMUVA, Universidad de Valladolid
[2]Institut de mathématiques de Toulouse

Abstract
Wasserstein barycenters and variance-like criteria based on the Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. We propose the estimation of the quantiles of the empirical process ofWasserstein's variation using a bootstrap procedure. We then use these results for statistical inference on a distribution registration model for general deformation functions. The tests are based on the variance of the distributions with respect to their Wasserstein's barycenters for which we prove central limit theorems, including bootstrap versions.

Keywords
Central Limit Theorem; goodness-of-fit; wasserstein distance

1. Introduction

Analyzing the variability of large data sets is a difficult task when the inner geometry of the information conveyed by the observations is far from being Euclidean. Indeed, deformations on the data such as location-scale transformations or more general warping procedures preclude the use of common statistical methods. Looking for a way to measure structural relationships within data is of high importance. Such issues arise when considering the estimation of probability measures observed with deformations; it is common, e.g., when considering gene expression.

Over the last decade, there has been a large amount of work dealing with registrations issues. We refer, e.g., to [3, 5, 29] and references therein. However, when dealing with the registration of warped distributions, the literature is scarce. We mention here the method provided for biological computational issues known as quantile normalization in [10, 22] and references therein. Recently, using optimal transport methodologies, comparisons of distributions have been studied using a notion of Fréchet mean for distributions as in [1] or a notion of depth as in [11].

As a natural frame for applications of a deformation model, consider J independent random samples of size n, where for each j$\in \{1, \ldots . J\}$, the real-valued random variable X_j has distribution μ_j and, for each $i \in \{1, \ldots . , n\}$, the

ith observation of X_j is such that

$$X_{i,j} = g_j(\varepsilon_{i,j}),$$

where the $\varepsilon_{i,j}$s are iid random variables with unknown distribution μ. Assume that the functions $g_1,\ldots,g_J$ belong to a class $\mathcal{G}$ of deformation functions, which model how the distributions $\mu_1, \ldots, \mu_J$ are warped one to another.

This model is the natural extension of the functional deformation models studied in the statistical literature for which estimation procedures are provided in [23] and testing issues are tackled in [12]. Note that at the era of parallelized inference where a large amount of data is processed in the same way but at different locations or by different computers, this framework appears also natural since this parallelization may lead to small changes with respect to the law of the observations that should be eliminated.

In the framework of warped distributions, a central goal is the estimation of the warping functions, possibly as a first step towards registration or alignment of the (estimated) distributions. Of course, without some constraints on the class $\mathcal{G}$, the deformation model is meaningless. We can, for instance, obtain any distribution on $\mathbb{R}^d$ as a warped version of a fixed probability having a density if we take the optimal transportation map as the warping function; see [35]. One has to consider smaller classes of deformation functions to perform a reasonable registration.

In cases where $\mathcal{G}$ is a parametric class, estimation of the warping functions is studied in [2]. However, estimation/registration procedures may lead to inconsistent conclusions if the chosen deformation class $\mathcal{G}$ is too small. It is, therefore, important to be able to assess the fit to the deformation model given by a particular choice of $\mathcal{G}$. This is the main goal of this paper. We note that within this framework, statistical inference on deformation models for distributions has been studied first in [21]. Here we provide a different approach which allows to deal with more general deformation classes.

The pioneering works [16, 26] study the existence of relationships between distributions F and G by using a discrepancy measure $\Delta(F,G)$ between them which is built using the Wasserstein distance. The authors consider the assumption $\mathcal{H}_0 : \Delta(F,G) > \Delta_0$ versus $\mathcal{H}_a : \Delta(F,G) \leq \Delta_0$ for a chosen threshold Δ_0. Thus when the null hypothesis is rejected, there is statistical evidence that the two distributions are similar with respect to the chosen criterion. In this same vein, we define a notion of variation of distributions using the Wasserstein distance, W_r, in the set $\mathcal{W}_r(\mathbb{R}^d)$ of probability measures with finite rth moments, where $r \geq 1$. This notion generalizes the concept of variance for random distributions over $\mathbb{R}^d$. This quantity can be defined as

$$V_r(\mu_1, \ldots, \mu_J) = \inf_{\eta \in \mathcal{W}_r(\mathbb{R}^d)} \left\{ \frac{1}{J} \sum_{j=1}^{J} W_r^r(\mu_j, \eta) \right\}^{1/r},$$

which measures the spread of the distributions. Then, to measure closeness to a deformation model, we take a look at the minimal variation among warped distributions, a quantity that we could consider as a minimal alignment cost. Under some mild conditions, a deformation model holds if and only if this minimal alignment cost is null and we can base our assessment of a deformation model on this quantity.

As in [16, 26], we provide results (a Central Limit Theorem and bootstrap versions) that enable to reject that the minimal alignment cost exceeds some threshold, and hence to conclude that it is below that threshold. Our results are given in a setup of general, nonparametric classes of warping functions. We also provide results in the somewhat more restrictive setup where one is interested in the more classical goodness-of-fit problem for the deformation model. Note that a general Central Limit Theorem is available for the Wasserstein distance in [19]. This works is published in a long version in [9].

2. Wasserstein variation and deformation models for distributions

Much recent work has been conducted to measure the spread or the inner structure of a collection of distributions. In this paper, we define a notion of variability which relies on the notion of Fréchet mean for the space of probabilities endowed with the Wasserstein metrics, of which we will recall the definition hereafter. First, for any integer $d \geq 1$, consider the set $\mathcal{W}_r(\mathbb{R}^d)$ of probabilities with finite rth moment. For μ and ν in $\mathcal{W}_r(\mathbb{R}^d)$, we denote by $\prod(\mu,\nu)$ the set of all probability measures π over the product set $\mathbb{R}^d \times \mathbb{R}^d$ with first (respectively second) marginal μ (respectively ν) . The L_r transportation cost between these two measures is defined as

$$W_r(\mu,\nu)^r = \inf_{\pi \in \prod(\mu,\nu)} \int \|x-y\|^r d\pi(x,y).$$

This transportation cost makes it possible to endow the set $\mathcal{W}_r(\mathbb{R}^d)$ with the metric $W_r(\mu,\nu)$. More details on Wasserstein distances and their links with optimal transport problems can be found, e.g., in [27, 35].

Within this framework, we can define a global measure of separation of a collection of probability measures as follows. Given $\mu_1, \ldots \mu_J \in \mathcal{W}_r(\mathbb{R}^d)$, let

$$V_r(\mu_1, \ldots, \mu_J) = \inf_{\eta \in \mathcal{W}_r(\mathbb{R}^d)} \left\{ \frac{1}{J} \sum_{j=1}^{J} W_r^r(\mu_j, \eta) \right\}^{1/r}$$

be the Wasserstein r-variation of $\mu_1, \ldots, \mu_J$ or the variance of the μ_js.

The special case $r = 2$ has been studied in the literature. The existence of a minimizer of the map $\eta \mapsto \{W_2^2(\mu_1,\eta) + \cdots + W_2^2(\mu_J,\eta)\}/J$ is proved in [1], as well as its uniqueness under some smoothness assumptions. Such a minimizer, μ_B, is called a barycenter or Fréchet mean of $\mu_1, \ldots, \mu_J$. Hence,

$$V_2(\mu_1, \dots, \mu_J) = \left\{ \frac{1}{J} \sum_{j=1}^{J} W_2^2(\mu_j, \mu_B) \right\}^{1/2}.$$

Empirical versions of the barycenter are analyzed in [8, 25]. Similar ideas have also been developed in [6, 15].

This quantity, which is an extension of the variance for probability distributions is a good candidate to evaluate the concentration of a collection of measures around their Fréchet mean. In particular, it can be used to measure the fit to a distribution deformation model. More precisely, assume as in the Introduction that we observe J independent random samples with sample $j\epsilon\{1, \dots, J\}$ consisting of iid observations $X_{1,j}, \dots, X_{n,j}$ with common distribution μ_j. We assume that $\mathcal{G}_j$ is a family (parametric or nonparametric) of invertible warping functions and denote $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_J$.

Then, the deformation model assumes that

there exists $(\varphi_1^*, \dots \varphi_J^*) \in \mathcal{G}$ and iid $\left(\varepsilon_{i,j}\right)_{1 \le i \le n, 1 \le j \le J}$ such that for all $j \in \{1, \dots, J\}, X_{i,j} = \left(\varphi_j^*\right)^{-1}\left(\mathcal{E}_{i,j}\right)$. (1)

Equivalently, the deformation model (1) means that there exists $(\varphi_1^*, \dots, \varphi_J^*)\epsilon\mathcal{G}$ such that collection of $\varphi_j^*\left(X_{i,j}\right)s$ taken over all $j \in \{1, \dots, J\}$ and $i\epsilon\{1, \dots, n\}$ is iid or, if we write $\mu_j(\varphi_j)$ for the distribution of $\varphi_j(X_{i,j})$, that there exists $(\varphi_1^*, \dots, \varphi_J^*)\epsilon\mathcal{G}$ such that $\mu_1(\varphi_1^*) = \cdots = \mu_J(\varphi_J^*)$.

We propose to use the Wasserstein variation to measure the fit of model (1) through the minimal alignment cost

$$A_r(\mathcal{G}) = \inf_{(\varphi_1, \dots, \varphi_J) \in \mathcal{G}} V_r^r\left\{\mu_1(\varphi_1), \dots, \mu_J\left(\varphi_J\right)\right\}. \quad (2)$$

Let us assume that $\mu_1(\varphi_1), \dots, \mu_J\left(\varphi_J\right), (\varphi_1, \dots, \varphi_J) \in \mathcal{G}$ are in $\mathcal{W}_r(\mathbb{R}^d)$. If the deformation model (1) holds, then $A_r(\mathcal{G}) = 0$. Under the additional mild assumption that the minimum in (2) is attained, we have that the deformation model can be equivalently formulated as $A_r(\mathcal{G}) = 0$ and a goodness-of-fit test to the deformation model becomes, formally, a test of

$$\mathcal{H}_0 : A_r(\mathcal{G}) = 0 \quad \text{vs.} \quad \mathcal{H}_a : A_r(\mathcal{G}) > 0. \quad (3)$$

A testing procedure can be based on the empirical version of $A_r(\mathcal{G})$, namely,

$$A_{n,r}(\mathcal{G}) = \inf_{(\varphi_1, \dots, \varphi_J) \in \mathcal{G}} V_r^r\left\{\mu_{n,1}(\varphi_1), \dots, \mu_{nJ}(\varphi_J)\right\}, \quad (4)$$

where $\mu_{n,j}(\varphi_j)$ denotes the empirical measure on $\varphi_j\left(X_{1,j}\right), \dots, \varphi_j\left(X_{n,j}\right)$. We would reject the deformation model (1) for large values of $A_{n,r}(\mathcal{G})$.

As noted in [16, 26], the testing problem (3) can be considered as a mere sanity check for the deformation model, since lack of rejection of the null does not provide statistical evidence that the deformation model holds. Consequently, as in the cited references, we will also consider the alternative testing problem

$$\mathcal{H}_0 : A_r(\mathcal{G}) \geq \Delta_0 \qquad \text{vs} \qquad \mathcal{H}_a : A_r(\mathcal{G}) < \Delta_0, \tag{5}$$

where $\Delta_0 > 0$ is a fixed threshold. With this formulation the test decision of rejecting the null hypothesis implies that there is statistical evidence that the deformation model is approximately true. In this case, rejection would correspond to small observed values of $A_{n,r}(\mathcal{G})$. In subsequent sections, we provide theoretical results that allow the computation of approximate critical values and p-values for the testing problems (3) and (5) under suitable assumptions.

3. Bootstraping Wasserstein's variations

We present now some general results on Wasserstein distances that will be applied to estimate the asymptotic distribution of the minimal alignment cost statistic, $A_{n,r}(\mathcal{G})$, defined in (4). In this section, we write $\mathcal{L}(Z)$ for the law of any random variable Z. We note the abuse of notation in the following, in which W_r is used both for the Wasserstein distance on $\mathbb{R}$ and on $\mathbb{R}^d$, but this should not cause much confusion.

Our first result shows that the laws of empirical transportation costs are continuous (and even Lipschitz) functions of the underlying distributions.

Theorem 1. Set v, v', η *probability measures in* $\mathcal{W}_r(\mathbb{R}^d)$, $Y_1, \ldots, Y_n$ iid *random vectors with common law* $v, Y_1', \ldots, Y_n'$, iid *with law* v' *and write* v_n, v_n' *for the corresponding empirical measures. Then*

$$W_r[\mathcal{L}\{W_r(v_n, \eta)\}, \mathcal{L}\{W_r(v_n', \eta)\}] \leq W_r(v, v').$$

The deformation assessment criterion introduced in Section 2 is based on the Wasserstein r-variation of distributions, V_r. It is convenient to note that $V_r^r(v_1, \ldots, v_J)$ can also be expressed as

$$V_r^r(v_1, \ldots, v_J) = \inf_{\pi \in \prod(v_1, \ldots, v_J)} \int T(y_1, \ldots, y_J) d\pi(y_1, \ldots, y_J), \tag{6}$$

where $\prod(v_1, \ldots, v_J)$ denotes the set of probability measures on $\mathbb{R}^d$ with marginals $v_1, \ldots, v_J$ and

$$T(y_1, \ldots, y_J) = \min_{z \in \mathbb{R}^d} \frac{1}{J} \sum_{j=1}^{J} \|y_j - z\|^r .$$

Here we are interested in empirical Wasserstein r-variations, namely, the r-variations computed from the empirical measures $v_{n_j,j}$ coming from independent samples $Y_{1,j}, \ldots, Y_{n_j,j}$ of iid random variables with distribution v_j. Note that in this case, problem (6) is a linear optimization problem for which a minimizer always exists.

As before, we consider the continuity of the law of empirical Wasserstein r-variations with respect to the underlying probabilities. This is covered in the next result.

Theorem 2. *With the above notation,*

$$W_r^r\left[\mathcal{L}\left\{V_r(v_{n_1,1},\ldots,v_{n_J,J})\right\},\mathcal{L}\left\{V_r(v'_{n_1,1},\ldots,v'_{n_J,J})\right\}\right]\leq\frac{1}{J}\sum_{j=1}^{J}W_r^r(v_j,v'_j).$$

A useful consequence of the above results is that empirical Wasserstein distances or *r*-variations can be bootstrapped under rather general conditions. To be more precise, in Theorem 1 we take $v'=v_n$, the empirical measure on $Y_1,\ldots,Y_n$, and consider a bootstrap sample $Y_1^*,\ldots,Y_{m_n}^*$ of iid (conditionally given $Y_1,\ldots,Y_n$) observations with common law v_n. We will assume that the resampling size m_n satisfies $m_n\to\infty$, $m_n=o(n)$ and write $v_{m_n}^*$ for the empirical measure on $Y_1^*,\ldots,Y_{m_n}^*$ and $\mathcal{L}^*(Z)$ for the conditional law of Z given $Y_1,\ldots,Y_n$. Theorem 1 now reads

$$W_r\left[\mathcal{L}^*\left\{W_r\left(v_{m_n}^*,v\right)\right\},\mathcal{L}\left(W_r\left\{v_{m_n},v\right)\right\}\right]\leq W_r(v_n,v).$$

Hence, if $W_r(v_n,v)=O_{Pr}\,(1/r_n)$ for some sequence $r_n>0$ such that $r_{m_n}/r_n\to 0$ as $n\to\infty$, then using the fact that $W_r\{\mathcal{L}(aX),\mathcal{L}(aY)\}=aW_r\{\mathcal{L}(X),\mathcal{L}(Y)\}$ for $a>0$, we see that

$$W_r[\mathcal{L}^*\left\{r_{m_n}W_r\left(v_{m_n}^*,v\right)\right\},\mathcal{L}\left\{r_{m_n}W_r\left(v_{m_n},v\right)\right\}]\leq\frac{r_{m_n}}{r_n}r_nW_r(v_n,v)\to 0$$

in probability.

Assume that, in addition, $r_nW_r(v_n,v)\rightsquigarrow\gamma(v)$ for a smooth distribution $\gamma(v)$. If $\hat{c}_n(\alpha)$ denotes that αth quantile of the conditional distribution $\mathcal{L}^*\{r_{m_n}W_r\left(v_{m_n}^*,v\right)\}$, then

$$\lim_{n\to\infty}\Pr\{r_nW_r(v_n,v)\leq\hat{c}_n(\alpha)\}=\alpha\,; \tag{7}$$

see, e.g., Lemma 1 in [24]. We conclude in this case that the quantiles of $r_nW_r(v_n,v)$ can be consistently estimated by the bootstrap quantiles, $\hat{c}_n(\alpha)$, which, in turn, can be approximated through Monte Carlo simulation.

As an example, if *d*=1 and *r*=2, under integrability and smoothness assumptions on *v*, we have

$$\sqrt{n}W_2(v_n,v)\rightsquigarrow\left[\int_0^1\frac{B^2(t)}{f^2\{F^{-1}(t)\}}dt\right]^{1/2},$$

as $n\to\infty$, where f and F^{-1} are the density and the quantile function of v, respectively; see [18]). Therefore, Eq. (7) holds. Bootstrap results have also been provided in [20].

For the deformation model (1), statistical inference is based on $A_{n,r}(\mathcal{G})$, introduced in (4). Now consider $A'_{n,r},(\mathcal{G})$, the corresponding version obtained from samples with underlying distributions μ'_j. Then, a version of Theorem 2 is valid for these minimal alignment costs, provided that the deformation classes are uniformly Lipschitz, namely, under the assumption that, for all $j\in\{1,\ldots,J\}$,

$$L_j=\sup_{x\neq y,\varphi_j\in\mathcal{G}_j}\frac{\|\varphi_j(x)-\varphi_j(x)\|}{\|x-y\|} \tag{8}$$

is finite. **Theorem3**. If $L=\max(L_1,\ldots,L_j)<\infty$, with L_j *as in* (8), *then*

$$W_r^r\left[\mathcal{L}\left[\{A_{n,r}(\mathcal{G})\}^{1/r}\right],\mathcal{L}\left[\{A'_{n,r}(\mathcal{G})\}^{1/r}\right]\right]\leq L^r\frac{1}{J}\sum_{j=1}^{J}W_r^r(\mu_j,\mu'_j).$$

Hence, the Wasserstein distance of the variance of two collections of distributions can be controlled using the distance between the distributions. The main consequence of this fact is that the minimal alignment cost can also be bootstrapped as soon as a distributional limit theorem exists for $A_{n,r}(\mathcal{G})$, as in the discussion above. In Sections ?? and ?? below, we present distributional results of this type in the one-dimensional case. We note that, while general Central Limit Theorems for the empirical transportation cost are not available in dimension $d > 1$, some recent progress has been made in this direction; see, e.g., [30] for Gaussian distributions and [32], which gives such results for distributions on $\mathbb{R}^d$ with finite support. Further advances along these lines would make it possible to extend the results in the following section to higher dimensions.

References

1. M. Agueh, G. Carlier, Barycenters in the Wasserstein space, SIAM J. Math. Anal. 43 (2011) 904–924.
2. M.Agulló-Antolín, J.A.Cuesta-Albertos, H.Lescornel, J.-M.Loubes, A parametric registration model for warped distributions with Wasserstein's distance, J. Multivariate Anal. 135 (2015) 117–130.
3. S. Allassonnière, Y. Amit, A. rouvé, Towards a coherent statistical framework for dense deformable template estimation, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 69 (2007) 3–29.
4. P.C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, C. Matrán, Trimmed comparison of distributions, J. Amer. Statist. Assoc. 103 (2008) 697–704.
5. Y. Amit, U. Grenander, M. Piccioni, Structural image restoration through deformable template, J. Amer. Statist. Assoc. 86 (1991) 376–387.
6. J. Bigot, T. Klein, Characterization of barycenters in the Wasserstein space by averaging optimal transport maps, ESAIM: Probability and Statistics, in press (2018).
7. S. Bobkov, M. Ledoux, One-dimensional empirical measures, order statistics and Kantorovich transport distances, Memoirs of the American Mathematical Society, in press (2018).
8. E. Boissard, T. Le Gouic, J.-M. Loubes, Distribution's template estimate with Wasserstein metrics, Bernoulli 21 (2015) 740–759.
9. DelBarrio,EustasioandGordaliza,PaulaandLescornel,Hélèneand Loubes, Jean-Michel, Central limit theorem and bootstrap procedure for Wassersteinâs variations with an application to structural relationships between distributions, Journal of Multivariate Analysis,169,341–362.

10. B.M. Bolstad, R.A. Irizarry, M. Åstrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.
11. V.Chernozhukov, A.Galichon, M.Hallin, M.Henry, Monge–Kantorovich depth, quantiles, ranks, and signs, Ann.Statist.45(2017)223–256.
12. O. Collier, A.S. Dalalyan, Curve registration by nonparametric goodness-of-fit testing, J. Statist. Plann. Inf. 162 (2015) 20–42.
13. M. Csörgó, Quantile processes with statistical applications, CBMS-NSF Regional Conference Series in Applied Mathematics 42, SIAM, 1983.
14. M. Csörgó, L. Horváth, Weighted Approximations in Probability and Statistics, Wiley, New York, 1993.
15. M.Cuturi, A.Doucet, Fast computation of Wasserstein barycenters, Proceedings of the International Conference on Machine Learning 2014, JMLR W&CP 32 (2014) 685–693.
16. C.Czado, A.Munk, Assessing the similarity of distributions–finite sample performance of the empirical Mallows distance, J.Statist. Comput. Simul. 60 (1998) 319–346.
17. E. del Barrio, P. Deheuvels, S. van de Geer, Lectures on Empirical Processes: Theory and Statistical Applications, European Mathematical Society, Zürich, Switzerland, 2007.
18. E. del Barrio, E. Giné, F. Utzet, Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances, Bernoulli 11 (2005) 131–189.
19. E. Del Barrio, J.-M.Loubes, Central Limit Theorem for empirical transportation cost in general dimension. ArXiv e-prints 1705.01299, 2017.
20. J. Ebert, V. Spokoiny, A. Suvorikova, Construction of Non-asymptotic Confidence Sets in 2-Wasserstein Space ArXiv preprint 1703.03658, 2017

Completing the securities picture: Integrating official securities Statistics with regulatory trading data[1]

David Buckmann, Tobias Cagala, Alena Wabitsch
Deutsche Bundesbank, Frankfurt, Germany

Abstract

With an ongoing integration of financial markets, the interconnectedness of market participants and corresponding risks to financial stability become increasingly important for policymakers and regulators. To develop a more comprehensive picture on exposures to risks and relationships between market participants, we integrate two data sources: Official statistics on securities holdings of banks in Germany (SHS) and regulatory transaction-by-transaction trading data that comprises millions of securities transactions per day (MiFID). Because the datasets provide information on stock markets from different perspectives – the SHS data show securities portfolios whereas the MiFID data show securities transactions – integrating the datasets is a challenge. To overcome this challenge, we combine supervised and unsupervised machine learning algorithms and develop a simple and transparent set of rules for the integration of the datasets. We find that, in combination with expert heuristics, our data driven approach allows for a successful isolation of subsamples that can be matched accurately between the datasets. For these subsamples, the integration of both data sources allows us analyse the exposure of banks to portfolio risks at any point in time. This is a considerable advancement from the monthly information on portfolios in SHS and the lack of information on portfolios in MiFID data.

Keywords

Securities Transactions; Securities Holdings; Banking Networks; Data Integration

JEL: L14, C8

1. Introduction

Financial markets play a central role in the efficient allocation of scarce resources. The downside of their pivotal role is that failures of financial markets pose a threat to economic stability. To safeguard financial stability and to improve our understanding of the ways in which financial markets work, central banks exploit statistical and regulatory data sources. Often, these data

[1] The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

sources provide different perspectives on the stock market. In this paper, we show the potential and caveats for integrating heterogeneous statistical and regulatory data on stock markets.

We focus on two data sources, German Securities Holdings Statistics (SHS) and data that are collected on the basis of the Markets in Financial Instruments Directive (MiFID). The datasets differ in three respects. First, whereas the MiFID data record every trade on German stock exchanges, the SHS data provide a monthly snapshot of the composition of investments in securities in Germany. Second, there are differences with respect to the level of granularity of information on the investor: In the SHS dataset, the investor positions are aggregated to the level of the economic sector of the investor. Exceptions to this rule are the reporting banks' own investments in securities, which are not aggregated with investments of other actors in the financial sector. The MiFID data, on the other hand, provide granular information on trades on a counterparty-by-counterparty level. The third, and most fundamental difference between the datasets, is that they provide diverse perspectives on the stock market. Whereas the SHS data show aggregated portfolios of investors and thus provide a portfolio perspective, MiFID data show flows between market participants. The datasets are similar in the sense that they both contain security-by-security information, i.e. the level of granularity is the same regarding the issuer and the unique identifier of the security.

An integration of the datasets allows us to exploit the strengths of both data sources: Whereas the SHS data provides a complete picture of banks' securities portfolios, MiFID data provides more timely and more granular information on investment decisions. To integrate the datasets, we proceed in two steps. In the first step, we use machine learning methods for a data-driven development of a matching algorithm. Specifically, we combine supervised segmentation of the data with unsupervised association rule discovery. In the second step, we refine the discovered rules with expert heuristics to develop a comprehensive set of rules for matching the data. We show that this approach provides a good performance for the integration of a large subset of data points. For this subsample, the MiFID data allow us to update the end-of-month portfolio composition, that banks report to SHS, continuously and analyse portfolio risks in-between the monthly SHS reporting dates.

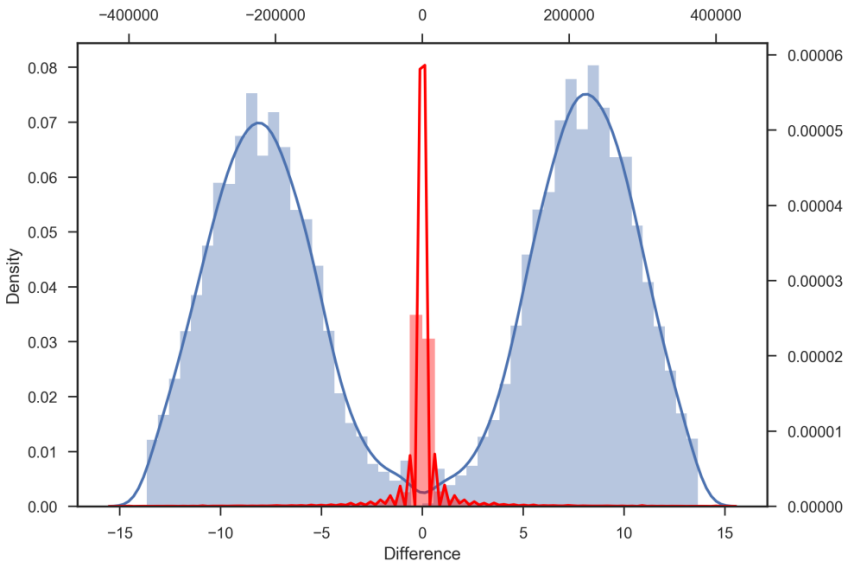
The paper is organized as follows. Section 2 outlines our methodology. Section 3 analyses our results. Section 4 discusses our results and concludes.

2. Data

In this paper, we use data on banks' own investments in equity securities (SHS) and banks' transactions of equity securities (MiFID) from 2014 to 2015. For a matching between the datasets to be feasible, we first have to transform the data so that both datasets show the same perspective on the stock market

(either flows or stocks). To this end, we derive transactions – a flow-perspective – from the SHS data by taking the first difference between the monthly reported stocks. To adjust for the difference in reporting frequencies between the datasets, we then aggregate transactions in the MiFID data to their monthly sum, netting purchase and selling transactions. The result are two transformed datasets (SHS* and MiFID*) that show aggregated monthly transactions of banks on a security-by-security basis. This leaves us with 764,713 data points. Figure 1 shows the distribution of the differences between the datasets (red bars and upper axis) and the distribution of normalized differences (blue bars and lower axis).[2] Because of the symmetry of the distribution around zero, on an aggregate level, the positive and negative deviations cancel each other out. Thus, both data sources show the same change in banks' aggregate stock of equity securities. Turning to the more granular security-by-security level, we find an exact match of the transactions for 26% of the data points. For 81% of the data points, the absolute difference is below EUR 10,000 (the average volume of a transaction in the SHS* data is EUR 42,685).

Figure 3: Distribution of Differences between the Datasets



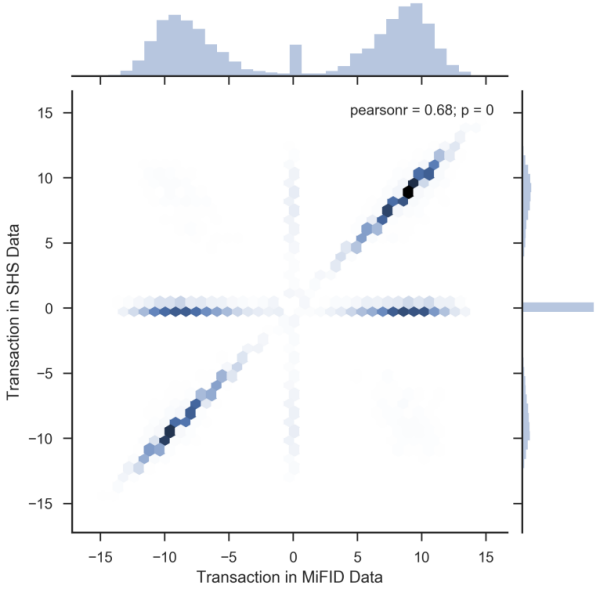
Note: The figure shows the distribution of the differences between the datasets (red and upper axis) and the distribution of normalized differences (blue and lower axis). For the normalized differences, we exclude differences of zero.

If the mismatches in **Figure 3** have a structural underpinning, we can use machine learning methods to mine rules for isolating transactions that can be

[2] For the normalization, we use the inverse hyperbolic sinus scaling function.

matched. **Figure 4** provides an example for a structural mismatch between the datasets. The figure shows the relationship between the transactions (normalized) in both datasets for securities that were issued by Canadian issuers. The darkness of the hexagons is proportional to the number of observations. We find that a large fraction of the differences is due to transactions only showing up in one dataset (data points on the horizontal and vertical axes). For securities of Canadian issuers, this absence of transactions is more prevalent in the SHS data. Because the structure of the mismatch correlates with observable features of the data, namely the issuer country of the security, there is a chance that a learning algorithm can successfully isolate groups of transactions that can be matched accurately.

Figure 4:Transactions in MiFID and SHS data for Securities with Canadian Issuers



Note: The figure provides an illustrative example of a structural mismatch between the datasets. The Figure shows the relationship between the transactions (normalized) in both datasets for securities that were issued by Canadian issuers.

3. Methodology

We proceed in two steps. First, we use a two-tier approach to derive rules for matching the datasets that combines supervised learning and unsupervised association rule discovery. Second, we develop a set of heuristic rules based on of the first-step results.

For the discovery of rules that allow us to integrate the datasets, we use decision trees (supervised) and association rules (unsupervised). Our goal is to find subsamples, for which an integration of the MiFID* data with the SHS* data

does not result in a mismatch of transaction volumes. To train the algorithms, we generate a binary outcome variable for each transaction *i* that takes the value one if there is a successful match and zero otherwise with:

$$y_i = \begin{cases} 1 & if\ \Delta T_i < \theta, \\ 0 & otherwise, \end{cases}$$

where $\Delta T_i = |MiFID_i^* - SHS_i^*|$ and θ is the threshold for a match. By altering θ, we can adjust the deviation that we accept for a transaction that we consider a match.[3]

To learn rules with a **decision tree**, we use y_i as an outcome. We further provide the tree with a number of features in the form of a dummy variable for data points that do not appear in the SHS data and dummies for the currency of the stock and the issuer country. To allow the tree to consider information on the investors, we include an indicator variable for the cluster to which a k-means (k=3) clustering algorithm assigns the investor on the basis of her aggregated volume of transactions (bank cluster). On the basis of these features, the tree successively splits the data into subsamples by selecting the variable and cutoff rule that maximizes the homogeneity of the subsamples in terms of the outcome. This way, the tree provides sample splits that produce groups of transactions *g* for which an accurate match is possible ($\sum_j^{n_g} \frac{y_j}{n_g} \to 1$) and groups of transactions that do not allow for an accurate matching of transactions ($\sum_j^{n_g} \frac{y_j}{n_g} \to 0$).

Unsupervised learning of **association rules** complements the supervised approach. The goal of association rule learning is to find groups of feature values that are common (support) in the data and indicate a high propensity for a successful matching of the datasets (confidence).[4] Because the algorithm provides us with a broad set of association rules and does not specifically isolate rules that allow for the integration of the datasets, we have to proceed in two steps. In the first step, we mine association rules. In the second step, we filter rules that associate feature values with the indicator variable for a successful matching of the data (y).

To derive the final heuristic ruleset, we combine the results from the decision tree with the association rules. We then exclude splits (rules) that intuitively do not make sense. This way, we ensure that the heuristic ruleset has a foundation in the experience of domain experts.

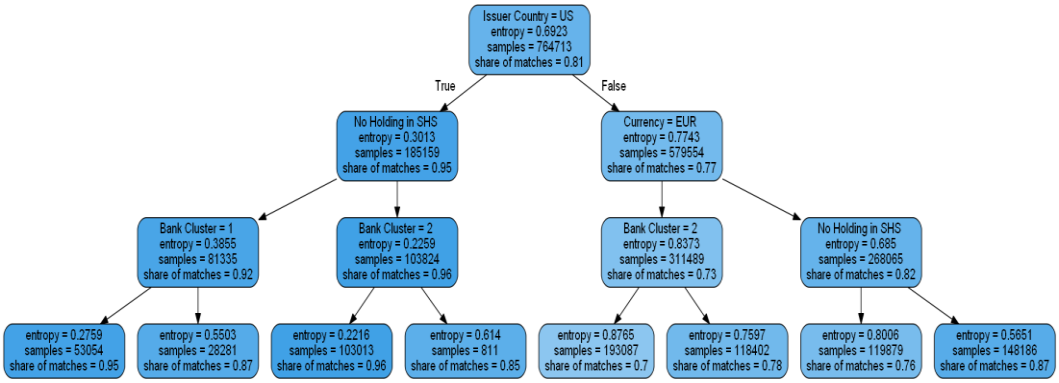
[3] We set the threshold to EUR 10,000.

[4] Because we filter the resulting set of rules for association rules that have the matching indicator *y* as a consequence, the share of matches in the subsample that satisfies the rule equals the confidence of the rule.

4. Results

Figure 3 illustrates the result of the supervised learning with a decision tree. For each node, the figure shows the feature of the split, the number of samples, and the share of data points that can successfully be matched in this node. To avoid overfitting and a level of complexity that renders the derived ruleset difficult to understand, we limit the maximal depth of the tree to three. We find that the issuer country and the bank cluster are important for defining subsamples with accurate matching. One example for a rule that isolates a subsample of transactions that we can integrate accurately is: Select transactions of securities with US issuers by banks in the second bank cluster if a corresponding investment was reported in the SHS data (96% of transactions can be matched in this subgroup).

Figure 5: Rules for Matching Datasets from a Decision Tree



Note: The figure shows the result of the supervised learning with a decision tree.

Table 1 shows the results of the **association rule mining**. In general, the intersection of the set of association rules and the set of rules that we derive from the decision tree is large. However, as we can see in the third row of the association ruleset, there are rules that we discover with the association rule mining that extend the set of rules that we derive from the decision tree.

Table 2: Results of the Association Rule Mining (=> y_i)

	Rule	Support	Share of Matches
1	(Bank Cluster = 1) & (Issuer Country = US)	0.07	0.95
2	(Bank Cluster = 1) & (Issuer Country = US) & (No Holding in SHS = True)	0.07	0.95
3	(Currency = USD) & (Issuer Country = US)	0.23	0.95

Note: The table shows three exemplary rules that result from the association rule mining that include the indicator for a successful matching of the datasets

(*y*) as a consequence. Out of the much larger body of rules, we select the rules that result in the highest share of matches between the SHS and the MiFID data.

On the basis of expert domain knowledge, we combine the results of the supervised approach (decision tree) and the unsupervised approach (association rules). The combined ruleset comprises ten rules for isolating subsamples that allow for an accurate matching of SHS with MiFID data.

5. Discussion and Conclusion

To develop a more complete picture of exposures to risks and the interconnectedness of market participants, we integrate an official statistic on securities holdings and regulatory transaction-by-transaction securities trading data. We develop a simple and transparent set of rules that allows us to integrate the datasets despite the stark conceptual differences between the data sources, by combining supervised and unsupervised machine learning algorithms. We find that, in combination with domain knowledge, this data driven approach allows for a successful isolation of subsamples that can accurately be matched between the datasets. One benefit from integrating the datasets is that, for the successfully integrated subsamples, we are able to analyse portfolio risks at any point in time, rather than only on a monthly basis. With the implementation of the new Market in Financial Instruments Regulation (MiFIR, MiFID II) in the beginning of 2018, we will be able to further improve the matching. This will allow for an even more accurate integration of both data sources.

The fire-sale channels of universal banks in the European sovereign debt crisis*

Giulio Bagattini*, Falko Fecht*, and Patrick Webery†
*Frankfurt School of Finance and Management
†Deutsche Bundesbank, DG Statistics

Abstract

We use a unique security-level data set to analyze correlations in bond trading of banks, their respective retail customers and their affiliated mutual funds. Matching banks' proprietary holdings with the holdings of their funds and their retail customers for the period 2009-2016 at the security level, we find evidence that banks sold off risky euro-area sovereign bonds to both their retail customers and their affiliated mutual funds (particularly their public funds) during the European sovereign debt crisis. Overall, this enabled banks with affiliated mutual funds to sell off larger amounts of their risky sovereign bond holdings, while bank-affiliated mutual funds acquired more risky sovereign bonds compared to their unaffiliated peers. The larger the risky sovereign bond position a fund acquired from its parent bank, the lower are the fund's short-term raw returns controlling for the risky bonds the fund overall acquired. Our findings show that banks use their customers portfolio and their affiliated funds as liquidity provider when they sell off their risk bonds without paying the funds the ad-equate liquidity premium. On the one hand, this points to a severe conflict of interest between banks' own account trading and their asset and wealth management services. On the other hand, it highlights that the severity of fire-sale contagion depends on the organizational structure of the financial sector.

* We would like to thank Tarun Ramadorai, Linda Goldberg, Dragon Tang, Jens Christensen, Milos Bozovic, Corinna Woyand and Christian Buschmann (discussants), conference participants at the 11th LSE Annual Paul Woolley Centre Conference, the 18th Annual FDIC Bank Research Conference, the FSB-CBoI Research Workshop on Non-bank Financial Intermediation, the 14th Annual Central Bank Conference on the Microstructure of Financial Markets, the EFA 2018, the EEA 2018, the 21st Annual Conference of the Swiss Society for Financial Market Research, the Belgrade Young Economists Conference, the Universität Augsburg-Deutsche Bundesbank-Universität Wien 7th Workshop Banks & Financial Markets, and seminar participants at the Deutsche Bundesbank, the University of St. Gallen, the European Central Bank, the Banque de France, the University College Dublin, University of Hohenheim, the Central Bank of Ireland, and Frankfurt School of Finance & Management for helpful comments and suggestions. We would also like to thank Gabriele Meinert, Christoph Fricke and the Division Securities and Money Market Statistics. The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1. Introduction

Fire sales are considered as one of the major channels of financial contagion (see Shleifer and Vishny (2011) for a comprehensive survey). In the euro area, fire sales of sovereign bonds have been pointed out as a main driver of systemic risk in the financial system and a key vulnerability of the banking sector (see, for instance, Greenwood et al. (2015)). Fire sales of sovereign bonds by distressed banks are also seen as a key element in the vicious circle linking banking and sovereign debt crises and contributing to an inherently fragile financial system (see Cooper and Nikolov (2018)). As a consequence, regulators call for minimum capital requirements underlying banks' sovereign bond holdings (see, for example, European Systemic Risk Board (2015)) in order to mitigate fire-sale contagion and the doom loop between banking and sovereign defaults. At the same time, though, recent research highlights that a bank can opportunistically steer its customers' portfolios towards assets which the bank intends to sell off from its proprietary trading portfolio (see Fecht et al. (2018)). This suggests that banks which dispose of a large customer base and/or manage considerable wealth on behalf of customers might be able to mitigate fire-sale pricing by pushing those sovereign bonds that the bank intends to liquidate to bank-affiliated mutual funds or directly to their retail customers.

In this paper, we test this hypothesis using a unique dataset from the Deutsche Bundesbank that allows us to match for the period 2009Q3–2016Q1 security-level data on all German banks' proprietary sovereign bond holdings with the respective security holdings of the bank's affiliated mutual funds (if it has any) as well as the holdings of its retail customers. As a proxy for the time-varying riskiness of a particular country's sovereign, we use credit default swap spread data from Markit at maturities matched to those of the individual sovereign bond.[1]

In a fist set of panel regressions, we find that whenever a bank sells a risky sovereign bond during the crisis the changes in the bank's holdings are negatively correlated with both its retail customers' and its affiliated mutual funds' holdings of the same bond. This negative correlation increases the riskier the respective sovereign bond. These findings hold even if we fully saturate the model with time-varying security, time-varying bank (or fund) and bank (or fund)-security fixed effects to account for market wide changes in funds' (households') risky bond investments, changes in a mutual funds' (bank customers') overall bond purchase and persistent differences in fund (bank customers') specific investments in certain bonds. Our findings are particularly

[1] We use the CDS on senior debt of the country with six different maturities (1y, 2y, 3y, 5y, 7y and 10y. In a robustness check, we also use the official credit ratings from S&P, Moody's and Fitch.

pronounced for public fund, in contrast to specialized funds that cater other financial institutions and are presumably more closely monitored. Interestingly, these results are robust if we also control for the fact that banks might sell particularly illiquid bonds (as proxied by the bid-ask spreads obtained from Bloomberg) to their customers and mutual funds to mitigate market impact. However, we do not find that a bank's sales of risky and illiquid sovereign bonds are more correlated with the bank customers' and funds' purchases than for liquid risky bonds. As regards bank characteristics, especially banks that experience a severe drop in their equity ratio (and presumably therefore have to deleverage fast and on a larger scale) tend to sell risky sovereign bonds to their customers.

In a second step, we compare the portfolio dynamics of funds that are affiliated to a bank with changes in the security holdings of independent mutual funds. Controlling for time-varying security and fund fixed effects, we find that bank-affiliated mutual funds increased their risky sovereign bond holdings significantly more than their unaffiliated peers. Similarly, when a fund has a parent bank and the parent bank reduced its holdings of a risk sovereign bond, we see that the affiliated fund purchases more of the respective risky bond than its peers, again taking time-varying fund and security fixed effects into account. Overall, we find that from the beginning of the sovereign debt crisis to its peak the portfolio share of risky sovereign bonds increased more at bank-affiliated mutual funds compared to the unaffiliated peers. This difference is the more pronounced the riskier the sovereign bond. These findings suggest that affiliated funds did not or could not offset the acquisition of risky bonds from their parent bank by reducing relatively their portfolio holdings of other risky sovereign bonds.

We next turn to the impact of a bank's fire sales of risky sovereign bonds to affiliated funds on the performance of those bank-affiliated funds. When we compare the raw returns of funds, we find that a fund's short-term performance is significantly lower if it has a parent bank and seemingly acquired more risky bonds from its parent bank. This holds even if we include time and fund fixed effects and control for a fund's overall risky bond holdings and acquisitions. This suggests that bank-affiliated funds provided price support when purchasing risky sovereign bonds sold off by their parent bank. In turn this compressed the liquidity premium those funds obtained compared to other funds that purchased risk bonds at fire sale prices in the market.

Finally, we study whether having a mutual fund also allowed banks to reduce their portfolio share of risky sovereign bonds during the sovereign debt crisis. When regressing for each bank the changes in the portfolio share of the different bonds, we find that banks with an affiliated fund were able to reduce their holdings of risky sovereign bonds significantly more than banks without an asset management company. This effect is robust to the inclusion

of time-varying bank and security fixed effects and appears stronger the riskier the respective bond.

Our findings have important implications. First, they suggest that there is a conflict of interest between banks' own account trading and the asset and wealth management services they offer to retail investors, potentially calling for better consumer protection. The EU regulation Mifid II rolled out in January 2018, which requires trading prices for certain fixed-income instruments to be published, might be a step in that direction.[2] However, outstanding sovereign bonds are subject to the new rules only if the initial size of the offering was greater than €1 billion, which is the case for only a small percentage of them.

At the same time our findings also show that the severity of fire-sale contagion depends on the organizational structure of the financial sector. Universal banks, i.e. bank holding companies that comprise, besides proprietary trading, also asset management services for customers and asset management companies, might mitigate fire-sale contagion and contribute to a more resilient financial system.[3] Third, these findings also suggest that regulatory proposals suggesting a separation between bank proprietary trading and other bank activities – such as the Dodd-Frank Act in the U.S.[4], the Vickers Report in the U.K.[5], and the Liikanen Report in the EU[6] – might aggravate fire-sale contagion and lead to a more fragile banking system and a more severe doom loop between banking and sovereign defaults. As a consequence, with these institutional separations becoming effective, the need for minimum capital requirements covering banks' sovereign bond holdings becomes even more pressing.

The remainder of our paper is organized as follows. In the following section we discuss the related literature. Section 3 describes the institutional background that led banks to large-scale sovereign debt sell-offs. In section 4 we present our data set, sample and main variables. Section 5 derives, from a simple univariate analysis, first suggestive evidence of trading in risky sovereign bonds between banks and their affiliated mutual funds, as well as their retail customers. Section 6 uses a more sophisticated panel approach to

[2] In a study of OTC secondary trades in corporate bonds in the United States, Edwards et al. (2007) find that transaction costs are lower for bonds with transparent trade prices, and they drop when the TRACE reporting system starts to publicly disseminate their prices.

[3] It is interesting to note that, while these implications suggest that the opportunistic behavior of banks has redistributional effects between bank owners and bank clients, they also imply that the risky assets are immediately shifted to unleveraged market investors, which eliminates the risk of further knock-on effects.

[4] Dodd-Frank Wall Street Reform and Consumer Protection Act, enacted on July 21, 2010.

[5] Final Report of the UK's Independent Commission on Banking from 2011, chaired by John Vickers

[6] Final Report of the High-level Expert Group on reforming the structure of the EU banking sector, chaired by Erkii Liikanen and initiated by EU Commissioner Michel Barnier.

analyze the correlation. In section 7 we study whether bank-affiliated funds acquired more risky sovereign bonds than their unaffiliated peers during the sovereign debt crisis, and in section 8 we focus on whether banks with affiliated funds sold off more risky bonds during the crisis period compared to other banks. Section 9 reports results from various robustness tests and section 10 concludes.

2. Data and sample description

For our empirical analysis, we obtain two key data sets: the first is from the Deutsche Bun-desbank's securities holdings statistics (SHS) and reports the proprietary security holdings of each bank operating in Germany, as well as, for each bank, the aggregate portfolio of all retail customers at the security level. The second data set comprises the security holdings for each investment fund operating in Germany from the investment funds statistics (IFS).

The data set for the securities holdings statistics and the investment funds statistics lists the quarterly holdings of banks, its customers and mutual fund companies on a security- by-security basis for the time period Q3 2009 to Q1 2016.[7] For our analysis, we exclude affiliates of foreign banks operating in Germany, as well as special-purpose banks, such as development banks.

We focus on the holdings of government bonds from the 19 euro-area countries and exclude from our analysis bonds not denominated in euro.[8] These sovereign bonds only account for around 2% of the total, both in the banks' proprietary portfolios and in the investment funds' holdings.

The first sample we construct focuses on banks' and their affiliated mutual funds' sovereign bond holdings. We use a hand-collected matching list to match banks to their affiliated asset management companies, i.e. to asset management companies fully owned by the parent bank, and ultimately to the asset management companies' mutual funds. In doing so, we take into account changes in the ownership structure of asset management companies that occurred during our sample period. In total, 19 banks appear in the matched sample. As asset management companies typically own more funds, the median number of fund holdings matched with a single bank holding in the sample is 4, while the average is 7.77. Our data at the fund level also contain an indicator for whether the fund is public (open to retail investors) or special (dedicated to a specific institutional investor). In our sample of matched holdings, the observations that refer to public funds are just over

[7] Before September 2009 the investment funds statistics were not available at the security level.
[8] If we kept non-euro denominatade bonds in the original currency in our data, changes in the nominal holdings would have different magnitudes for different currencies. Alternatively we could convert them into euro. But then exchange rate fluctuations would introduce spurious correlations in the holdings that are unrelated with the trading activity of banks/funds. For these reasons, we drop securities not denominated in euro.

20% of the total. All the most important asset management companies in our sample own at least some public funds. The median number of public fund holdings associated to a single bank holding is 2, while the average is 3.4.

We match the bank and fund holdings on a security-quarter basis and drop observations when a bond only appeared in the bank's proprietary portfolio, but not in any of the bank's affiliated mutual funds' portfolios. Similarly, we disregard observations of sovereign bond holdings by a fund when the parent bank does not hold the same bond. Overall during the sample period, the average bank holds 329 distinct sovereign bonds that also appear in the sample of common bond holdings with its mutual funds, 170 of which are German Bunds and 70 of which are issued by one of the GIIPS countries. However, this number varies widely: the three most important banks in the sample hold on average 1148 distinct securities, while 7 banks have few bonds in common with their asset management arm, with no common holding at all in several quarters.

The 31 asset management companies that appear in the sample own as many as 3059 different funds, each of which holds on average 21 distinct bonds that the parent bank also has (median 11). The upper 10% hold from 47 to 396 distinct securities and the bottom 10% hold just one.[9]

The second sample focuses on banks' and their retail customers' holdings of sovereign bonds. Here no matching is required, because each German bank has to report besides their own security holdings the aggregate holdings of its retail customers on a security-by-security basis directly to the SHS. In total, 538 banks report at least one euro-area sovereign bond held both in the bank's and its customers' portfolio. We have on average 13 different securities for each of these 538 banks, out of which 45% are German and 38% are issued by the GIIPS countries: in particular, 24% are Greek bonds. Again, the distribution is extremely skewed: 41% of these banks have only one bond in common with their households customers, while the largest held a total of 990 distinct securities.

We use the two separate samples not only because analyzing bank-fund level correlations and bank-customer level correlations is interesting in its own right. The bank-fund level sample also has a much larger cross-section of bonds, while the bank-customer sample has a larger cross-section of banks allowing us to also study the effects of bank characteristics.

[9] The same funds' portfolios include overall (independently of whether they appear in the portfolio of the parent bank) an average of 40 distinct euro-area sovereign bonds over the sample period (median 24).

3. Conclusion

In this paper, we provide evidence suggesting that banks used both their affiliated mutual funds and their retail customers as an exit channel to sell off risky sovereign bonds. Some evidence indicates that banks did so to mitigate market impact: they seem to have particularly sold bonds with a relatively large bid-ask spread to their funds. But at the same time banks presumably pushed liquid risky bonds to their affiliated funds and retail customers. Admittedly, our test on whether banks used funds and customers as exit channel to mitigate market impact suffers from the fact that our proxy for market liquidity – the bid-ask spread – is not the best measure for market impact.

Our further analysis shows that bank-affiliated mutual funds not only increased their holdings of those bonds that their parent bank sold, they also increased their overall portfolio share of risky sovereign bonds during the euro-area sovereign debt crisis significantly more than their unaffiliated peers. This suggests that those funds ended up being riskier than funds without a parent bank. At the same time banks with affiliated mutual funds were able to reduce their holdings of risky and illiquid sovereign bonds more significantly during the sovereign debt crisis than comparable banks without an affiliated asset management company.

Although evidence indicates that funds did not underperform in the long term after piling up sovereign risk, this seemingly opportunistic behavior of banks might in general undermine the efficiency of their clients' investment decisions. On the other hand, it presumably helped banks to offload risky sovereign holdings with only limited market impact. As a consequence, this exit channel might have also helped to mitigate fire-sale pricing and thus fire-sale externalities.

Linking household survey data and aggregate statistics: The experience of Banca d'Italia

Andrea Neri
Banca d'Italia

Abstract

The financial crisis of 2008 and the following economic downturn have increased demand for timely, coherent and consistent distributional information for the household sector. The ideal approach to produce distributional indicators is to combine the information coming from the sample surveys with the one from national accounts. However, this is not an easy task since these two data sources usually don't provide coherent results. In the paper, I discuss the experience of Banca d'Italia in trying to reconcile the two sources of information and the methods used to compute distributional indicators for the household sector.

Keywords

Distributional national accounts; micro-macro linkage; household wealth, Pareto distribution, imputation methods

1. Introduction

The financial and economic crisis recently experienced by many European countries have increased demand for timely, coherent and consistent distributional information for the household sector.

The G20 data gap initiative has encouraged the production and dissemination of distributional information on income, consumption, saving, and wealth for the household sector. Eurostat and the European Statistical System have agreed in the "Vienna Memorandum" in 2016 to work towards the same objective as far as consumption, and income are concerned. In recent years, also the measurement of household wealth is becoming increasingly important for policy-making. This is especially the case in societies where job insecurity is growing and where the welfare state is no longer able to ensures acceptable living standards to all individuals. In these circumstances, household wealth becomes an important buffer to guarantee an economic wellbeing.

The measurement of household wealth is receiving high priority especially in the agenda of national central banks (NCBs). In fact, most of the NCBs in the in the Euro area, collect micro data on household finance and consumption (Eurosystem Household Finance and Consumption Survey, HFCS). Banca d'Italia conducts a wealth survey (SHIW, Survey on household income and

wealth) since 1965. The survey consists of a probabilistic sample of around 8,000 households that are representative of the Italian population. Starting from 2010, the SHIW is the Italian component of the HFCS.

Survey data are mainly used by NCBs for financial stability purposes to evaluate the households' ability to face their levels of indebtedness if some shock occurs (such as losing a job). Moreover, they help NCBs to have a better understanding of the effects of monetary policy on households' saving and spending decisions. Finally, survey data also enables central banks to estimate the effects of fiscal policies through simulation models.

National accounts (SNA) are another important source of (aggregate) information about the households' economic conditions. For instance, the financial accounts (FA) report the value of aggregate asset holdings and liabilities of all the resident households.

The ideal approach to produce distributional indicators would be to combine the information coming from the sample surveys with the one from national accounts. However, this is not an easy task since these two data sources usually don't provide coherent results, even after accounting for differences in definitions and concepts.

In 2015 the European Central Bank has launched the Expert Group on Linking Macro and Micro Data for the Household Sector (EG-LMM) with the aim to understand, quantify and explain the main differences between the Household Finance and Consumption Survey (HFCS, the the harmonized survey collecting micro data on household finance and consumption at the Euro area level) and the financial accounts (FA). Building on its experience, the group is currently working on a method to produce distributional information combining micro and macro data. Banca d'Italia is actively involved in this project. In this paper I will present its experience in trying to combine the two sources of information.

2. The main differences between survey data and the National accounts

Survey data and National accounts are computed for different aims and are based on different definitions and concepts. The aim of the HFCS is to gain more insight into the economic behavior of households as well as into the distribution of wealth and liabilities among households and household groups. The valuation of assets and liabilities is based on the households' self-assessment. Some countries complement interview data with administrative or estimated data for individual wealth or income items. The HFCS data collection is based on a set of common definitions and descriptive features according to an output-oriented approach.

The aim of FA is to provide timely macroeconomic information on the balance sheets as well as financing and investment of the entire household

sector (including nonprofit institutions serving households). FA do not exclusively focus on the household sector but rather describe relations between all institutional sectors. The definitions of instruments, sectors and concepts such as valuation are given by the ESA 2010 and are mandatory in all EU countries.

The two sources of information present differences in the definitions of the household sector, in the periodicity, the timeliness, the reference periods and in the valuation criteria.

In theory, once such notional differences are taken into account the two sources of information should provide consistent information. Unfortunately, this is not often the case.

Two major reasons for this discrepancy relate survey data. The first one is unit nonresponse that happens when some groups of the selected households refuse to participate in the survey. This is usually the case for very rich households (in most cases they are even difficult to contact to negotiate the interview). Unfortunately, rich households tend to concentrate a large share of wealth in their hands and therefore such missing wealth will not show up in survey data. Second, since wealth surveys include both complex and sensitive items, respondents are not always able or even willing to report the correct value of an item. The underreporting behavior generally differs across the groups of the population or across wealth components. For instance, rich households may be prone to underreport their wealth because of social desirability bias. Moreover, financial assets tend to be more underreported, whereas rich households tend to have larger portfolio shares of these assets.

The combination of nonresponse and underreporting leads to a general missing wealth problem in surveys. Figure 1 shows the survey estimates of total financial assets as a percentage of the national accounts. For most countries, there is considerable discrepancy between the two sources of information.

3. The experience of Banca d'Italia

Banca d'Italia has a long standing interest in the comparison of micro and macro statistics relating household wealth. This interest was initially motivated by the need to assess the quality of the survey results by comparing them with external benchmarks.

Cannari and D'Alessio (1990) compared SHIW estimates of real estate wealth with those derived from the Census. They found that the number of residential properties was quite well estimated in the SHIW, but that the number of rented and vacation homes was severely underestimated. In the same paper the authors proposed a method for correcting the survey estimate of the number of dwellings owned.

Similar results were found by Neri and Monteduro (2013) who carried out an adjustment of housing wealth based on the aggregate distributions of ownership from tax records. The authors found that SHIW underestimates both the number of taxpayers who own just one and those who own more than five units of housing. Correcting the SHIW data by aligning the sample data with the administrative data increases total housing wealth by about a quarter.

As far as financial wealth is concerned, Cannari, D'Alessio, Raimondi and Rinaldi (1990) and Cannari and D'Alessio (1993) performed a statistical matching of the financial assets declared by SHIW respondents with data provided by a sample of commercial bank clients from a survey carried out by the bank. The authors used statistical matching to model non-reporting and under-reporting behavior and to adjust SHIW data. A similar approach was used (D'Aurizio et al., 2006). The adjusted estimates of financial assets average more than twice the original figures, reaching 85 percent of the aggregate coming from financial assets. The paper also adjusted financial liabilities, whose corrected values are on average about 40 percent higher.

Other studies were mainly focused on the analysis of the differences in definitions and concepts between the micro and macro sources (Antoniewicz 2005 and Bonci et. Al 2005).

More recently, D'Alessio and Neri (2015) conducted several adjustment experiments on SHIW data, combining different imputation and calibration techniques, in order to produce estimates consistent with the macro-economic information available from other sources. The study shows some results are robust to the adjustment method applied. For instance, whatever the method is used, the adjusted estimates of the Gini concentration indexes of both income and wealth are always higher than the unadjusted one.

Yet, the authors also show that results are strongly affected by the auxiliary information available. Without any external information, they are basically driven by the choice of the assumptions behind the adjustment models.

The main limitation of all these studies is that they were not based on an exact matching of survey data with administrative records. This was mainly due to the existence of legal constrained that made it impossible for data producers to share their administrative records with Banca d'Italia. Given such limitation, the only solution available was to use statistical matching techniques whose goal is to find the most similar observations in two different databases (based on the observable characteristics). Banca d'Italia is currently working to change this situation for the future.

4. The current approach and the road ahead

Banca d'Italia is currently working in collaboration with the EG-LMM group and with the University of Perugia (professor Giovanna Ranalli) to develop a new adjustment method that builds on the previous experiences.

The methodology focuses on the two major reasons for the macro-micro gap (once differences in definitions are addressed): the low probability of rich households to be captured by the survey and the underreporting behavior. The baseline method works under the assumption that the only external information available are the National accounts totals and a list of the richest people in the country such as the Forbes World Billionaires list. Yet, it can be easily extended to incorporate external information (when available).

The preliminary step consists in harmonizing as much as possible definitions and concepts across the two data sources. For instance, the total wealth held by non-profit institutions serving households is estimated and then removed from FAs. Then we apply a set of sequential adjustments some of which are iteratively repeated until a convergence criterion is met. The first step is to split the survey sample in two groups: the "rich" households and the "non-rich" households. The two groups are then adjusted using a different methodology. For rich households we use the Pareto method (Vermeulen, 2016). The Pareto is a highly right-skewed distribution with a heavy tail which has already been shown to fit the upper tail of the distribution. It requires a preliminary estimation of the share of rich households and the choice of a wealth threshold (above which households are classified as rich). As a result of the method it is possible to estimate the total wealth held by rich households. We then subtract it from the total household wealth and distribute the remaining share among "non-rich" households using imputation methods (proportional adjustment can be seen as a particular case). We repeat the process until convergence (the difference between the share held by the Rich does not change significantly over iterations).

The iterative procedure makes sure that the adjustment of the two groups is done simultaneously and that the final results do not depend on the initial choices relating the threshold and the parameter of the Pareto distribution. We call this approach "Simultaneous Pareto-calibration allocation".

Table 1 shows some preliminary results relating four countries. The method is compared with a simple proportional allocation which consists in estimating for each wealth component the ratio between macro and micro estimate and then in multiplying the amount declared in the survey by respondents to this ratio. The proportional method is quite used as a benchmark since it is very easy to apply and preserves the univariate distributions. The cons are that it assumes that the underreporting behavior is equal for all households and that it does not adjust either for missing wealthy at the top or for noreporting.

The two methods produce similar results in qualitative terms. They both estimate a higher number of rich households and a higher level of inequality compared to survey unadjusted data. Moreover, the level of indebtedness (compare to financial assets) is lower for rich household than the one resulting from unadjusted data.

However, the punctual estimates are quite different, with the Simultaneous Pareto-calibration allocation generally suggesting a higher level of inequality. Without auxiliary information, the results are likely to depend on the modelling assumptions and it is difficult to assess the reliability of the final statistics. The ideal situation would be to have access to administrative records (such as credit registers or tax data) matched to survey data at the individual level. Such information can shed light to the magnitude of the missing wealth due to nonresponse and measurement error and how to distribute it among households (giving for instance a larger share of the gap to those that are estimated to be more prone to underreporting).

In Italy, we recently had the opportunity to match survey data with credit register data. This fact gives us the opportunity to assess the goodness of our model. Table 2 shows the distribution of household debt by gross wealth quintiles. According to credit register data, about 64% of total debt is hold by households in the highest wealth class and only 0,1% is held by the poorest households. The results in the table show that our method gives promising results.

5. Conclusions

In recent years there has been an increasing demand for incorporating microeconomic heterogeneity in the aggregate statistics relating household income and wealth coming from National Accounts. Yet, the production of distributional national accounts is still in its infancy.

The main difficulty to overcame is the existence of a sizable gap between survey data, which contain distributional information and the national accounts. When this is the case, like in the HFCS survey, the challenge is to find a sound methodology to fill the gap between these two sources of information.

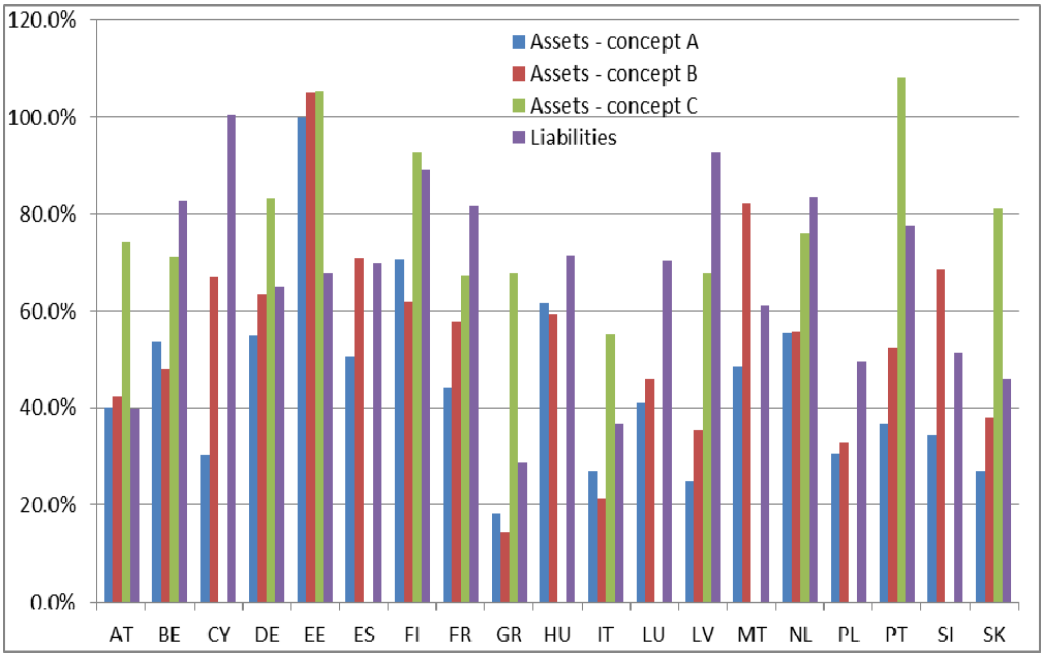
At present there is no clear, transparent methodology that is accepted in the literature for combining different data sources and to evaluate the quality of results.

Banca d'Italia has a long tradition in trying to reconcile micro and macro statistics on household income and wealth. Drawing on our past experience we are currently working on a new flexible method to allocate the macro-micro gap that enables to incorporate all the external information available (if any). The most favorable scenario would be to have survey data matched with administrative records. This would enable very detailed assessments of the

underreporting behavior and of the nonresponse bias. Yet, even if register data are available only at the aggregate level, that would provide beneficial information. Where register data can be disaggregated by some household groups with classifications available in the HFCS data, assessments on the degree of reporting bias across various household groups can be made. Yet, administrative sources are not perfect either. The main limitation is that they have not been designed thinking of a statistical use. Moreover, admin data may suffer of undercoverage problems (i.e. personal tax data may not include individual below a given threshold) or may be available with some delay (i.e. personal data are generally available with a 2-years lag from reference period). Finally, admin data are likely to use different concepts and definitions from the ones used in the survey.

Administrative records are likely to be the key to produce reliable distributional indicators. Yet, more research is still needed to use them in combination with survey data. This is probably one of the main challenges for the near future.

Figure 1. Financial Assets: comparison between survey data and FA
(HFCS wave 2 – lhs for FA (HH+NPISH) and HFCS, rhs for coverage ratios)



Financial assets: deposits, debt securities, loans, listed shares and investment fund shares.
Source: EG LMM final report 2019.

Table 1 – comparison between results of alternative estimation methods

		FR	IT	DE	FI
HFCS[1]	N. households above w_0(1)	861,196	650,893	1,168,432	49,560
	Top 10% (2)	49.4%	42.3%	57.5%	41.4%
	debt to fin. assets (top quintile)	0.99	0.29	0.62	0.81
	Gini index	0.66	0.60	0.74	0.59
Proportional method	households above w_0 (1)	1,789,261	1,457,439	1,656,712	37,101
	Top 10% (2)	52.3%	52.2%	58.9%	51.5%
	debt to fin. assets (top quintile)	0.49	0.19	0.44	0.60
	Gini index	0.67	0.64	0.72	0.70
Simultaneous Pareto-calibration allocation	households above w_0 (1)	1,686,232	1,261,837	1,694,376	46,000
	Top 10% (2)	60.1%	62.3%	67.4%	39.2%
	debt to fin. assets (top quintile)	0.50	0.22	0.44	0.57
	Gini index	0.71	0.73	0.78	0.67

(1) Estimated number of households after the wealth threshold, by net wealth. Pareto wealth threshold 0 set at 1mln EUR. (2) Wealth share of the top 10 percentile. Top % shares are computed with the HFCS total as the reference (which, in the Pareto adjusted survey, is very close to the FA total). Source: Final report EG-LMM

Table 2. Share of total debt owned by wealth quintile

Method	Quintiles of gross wealth				
	1°	2°	3°	4°	5°
HFCS-IT	2%	7%	23%	30%	39%
Proportional allocation	2%	10%	26%	29%	33%
Simultaneous Pareto-calibration allocation	0.1%	4%	14%	25%	56%
Administrative data*	0.1%	4%	11%	20%	64%

*Credit register; provisional data

[1] Based on the period in which HFCS interviews have been conducted, the data have been compared with FA data at end 2014 for DE, FR and IT, and end-2013 for FI.

References

1. Antoniewicz R. R. Bonci, A. Generale, G. Marchese, K. Maser, P. O'hagan, (2015), Household wealth: comparing micro and macro data in Cyprus, Canada, Italy and United States. 10.13140/rg.2.1.1558.9928.
2. Bonci R., G. Marchese, A. Neri (2005). Financial wealth in the financial accounts and in the Survey of Household Income and Wealth, Temi di discussione (Economic working papers) 565, Bank of Italy, Economic Research and International Relations Area.
3. Cannari L., G. D'Alessio (1990), Housing Assets in the Bank of Italy's Survey of Household Income and Wealth, in Dagum e Zenga (editor), "Income and Wealth Distribution, Inequality and Poverty", Springer Verlag, Berlino, p. 326-334.
4. Cannari L., G. D'Alessio, G. Raimondi, A.I. Rinaldi (1990), Le attività finanziarie delle famiglie italiane, Banca d'Italia, Temi di discussione, n. 136.
5. Cannari L., G. D'Alessio (1993), Non-reporting and Under-reporting Behavior in the Bank of Italy's Survey of Household Income and Wealth, in "Bulletin of the International Statistical Institute", vol. LV, n. 3, Pavia, p. 395-412.
6. Expert Group Linking Marco and Micro Data for the Household Sector (EG LMM) 2019, Final Report, June 2019.
7. Neri A., T. Monteduro (2013), La ricchezza immobiliare delle famiglie italiane: un confronto fra dati campionari e censuari, Questioni di Economia e Finanza, n. 146 – January.
8. Vermeulen, P. (2016). Estimating the Top Tail of the Wealth Distribution, American Economic Review, 106 (5): 646-50. DOI: 10.1257/aer.p20161021.

Linking micro data sets for firms' FX risk monitoring database

Burcu Zühal İman Er, Özgül Atılgan Ayanoğlu
Central Bank of the Republic of Turkey, Ankara, Turkey

Abstract

The purpose of this paper is to present the experiences of the Central Bank of the Republic of Turkey (CBRT) in establishing a data hub which provides a comprehensive data set for FX risk monitoring of non-financial corporations (NFCs). The increasing net FX open position of NFCs, coupling with depreciating TRY, has arisen the need to compile micro level data on the FX positions of real sector companies. CBRT established the Systemic Risk Data Monitoring System to collect liabilities and assets information from NFCs which have high FX bank loans. In order to be able to have a comprehensive understanding of the situations of NFCs, CBRT initiated the Real Sector Data Hub project where company level data from various sources are compiled and linked to the data collected from companies. Credit database from the Banks Association of Turkey, financial tables from Revenue Administration, employment information from Social Security Institution, export and import data from Ministry of Trade are among the many sources which are planned to be integrated in the data hub. We believe that this project will enable decision makers to rely on solid micro level analysis while taking policy actions.

Keywords

FX risk, Systemic Risk Data Monitoring System, Real Sector Data Hub

1. Introduction

The problems in the American mortgage market triggered the most severe financial crisis in the United States in 2007 and led to the "Global Financial Crisis". Several developments led to the construction of this crisis process, including the complex securities and unstandardized derivative contracts, high leverage and inadequate risk management. These factors together created systemic risk which can be defined as the risk associated with the collapse or failure of a company, industry, financial institution or an entire economy. One of the most important features of systemic risk is that the risk spreads from unhealthy institutions to healthy ones through a transmission mechanism. It is an endogenous factor in a market system and therefore, it is deemed unavoidable. Emerging markets experienced particularly aggravating effects through contagion arising from the interconnectedness of economies (BIS, 2016).

The systemic concerns about the bankruptcy of several large financial firms in the U.S. forced FED to enact expansionary monetary policy in the upcoming years of the crisis and other developed countries followed counterpart policies as their economic situation was very much alike. The quantitative easing policy worked in two transmission channels. First channel is direct and it involves a central bank buying up long-term public and private debt in massive amounts resulting in increasing money supply. (BIS, 2014; IMF, 2015) Since the start of the crisis, the four central banks of the United States, the United Kingdom, the Euro zone, and Japan have injected trillions of liquidity into their economies, pushing interest rates to very low levels. The second channel is semi direct; the decline in bond market yields resulting from central bank bond purchasing, lowers the costs of financing and triggers demand. As a result, lower policy rates and expansion in central bank balance sheets in the advanced economies created excess liquidity for emerging markets (Shin 2013). This has eased financing conditions and provided FX borrowing opportunities for emerging market companies. Lower interest rates of FX loans and moderate exchange rate levels increased their foreign currency borrowing. Aggregate foreign currency liabilities in the balance sheets of nonfinancial companies have therefore risen sharply since 2010.

In December 2013, FED announcement for tapering off its asset purchases from January 2014 had a massive effect on markets. Tapering was on a progressive basis and the asset purchase program was concluded in October 2014. In October 2017, the FED took a crucial step by initiating the process to reduce its balance sheet. This policy was implemented by reducing FED's reinvestment of payments made by issuers of securities it holds. The ECB continued its purchasing program in 2016 and started going for a normalization of its balance sheet in 2017.

The decisions to end quantitative easing constituted a challenge for emerging markets as they had significant vulnerabilities arising from currency mismatches in the balance sheets of NFCs that may aggravate market volatility. Waves in international capital flows deteriorate the exchange rate exposure of nonfinancial companies through exchange rate depreciations, leverage decisions and corporate financial distress. The studies in the literature point out the fact that firm size is related to corporate distress and, further, that currency depreciations amplify the impact of leverage on financial vulnerability for large firms during a crisis. There is also a granularity effect in that large firms are systemically important (Gabaix, 2011). Relatedly there is evidence that the sales growth of large firms with higher leverage is more adversely impacted by exchange rate shocks (Alfaro et.al. 2017). Hence, the empirical studies proves the importance of micro level analysis in order to see the big picture.

2. Data Gaps Initiative

Global financial crisis urged the actors of the system to take proactive measures for better monitoring and preventing the failures of the markets. In this context, the IMF and the FSB presented a report that would launch the Data Gaps Initiative (DGI) at the Pittsburg summit in 2009 to identify major financial and economic information gaps that needed to be filled, along with the set of recommendations to be implemented in the years to come. The aim of this initiative is supporting enhanced policy analysis. The first phase of the DGI was successfully ended in autumn 2015 and the second phase of the initiative, DGI – 2, was endorsed by the G-20 Finance Ministers and Central Bank Governors. The new or revised 20 recommendations require inland or cross border data sets which provide necessary information for monitoring the risk in financial sector and analyzing vulnerabilities, interconnections and spillovers. Moreover, communication of official statistics to serve the key objective of implementing regular collection and dissemination of comparable, timely, high quality and standardized statistics for policy use is encouraged.

As a member of FSB, Turkey is actively involved in the DGI-2. The initiative has been included in the Official Statistics Program for 2017-2021 as a separate statistics area and the CBRT has been designated as the coordinator of the institutions which are responsible for the studies concerning the aforementioned set of statistics. CBRT, TURKSTAT, Ministry of Treasury and Finance, Banking Regulation and Supervision Agency and Capital Markets Board are contributing to these statistics by conducting their own studies about the related recommendations. In Turkey, there has been progress in priority areas of coordinated portfolio investment and government finance statistics with the contribution of the earlier studies. There have been significant improvements in the statistics produced within the scope of the DGI with respect to timing and content and new statistics, such as sectoral accounts and real estate price indexes, have started to be compiled and produced. For other priority areas, action plans have been drawn up for the institutions to complete the necessary improvements in a specified calendar.

3. Systemic Risk Data Monitoring System

During the last ten years, the FX open position of non-financial corporations in Turkey has increased dramatically. This, in turn, brought up the issue of balance sheet imbalances deriving from currency mismatches. This created vulnerability leading to the fact that depreciation of TRY would deteriorate balance sheets of NFCs which could result in bankruptcies. Therefore, monitoring FX position of NFCs has become more critical from a macro-prudential perspective.

In this context, the CBRT has taken a major step in terms of constructing a data set that enables micro level analysis by initiating the "Systemic Risk Data Monitoring System". To increase the resilience of Turkish economy against exchange rate volatility, Turkey's Financial Stability Board decided to establish a new regulatory framework of FX risk management based on a company specific data set where risks can be monitored in detail. For this purpose, the duty of collecting detailed data about foreign currency positions of non-financial firms has been assigned to the CBRT. At first, non-financial firms, which have FX loans over 15 million US dollars constitute the scope of the system. The aim is to conduct analysis of detailed company level data and use the results as an input in the surveillance process in order to allow more precise calculation of risks and effective policy making.

According to the CBRT Law, the Bank is authorized to collect data on economic issues and produce statistics by compiling and processing data. In order to establish the legal basis to collect data for purposes other than producing statistics, amendments to the Central Bank Act. Law Nr 1211 are made. With the amendments, the CBRT has gained the authority to request all kinds of information and documents from real and legal persons to monitor their transactions affecting their foreign exchange positions. Central Bank determines the scope of the requested information and documents, method of collecting and monitoring data, supervision of accuracy and all the other principles of implementation. Judicial fine shall be imposed against companies that do not give the required information and documents, or give incomplete or incorrect information. Regulation on the Principles and Procedures regarding the Monitoring of Transactions Affecting Foreign Exchange Position by the CBRT has been enacted and announced in the Official Gazette dated 17/2/2018. The regulation rules that firms report the data forms specified by the CBRT quarterly and annually. The annual reports are subject to independent auditing process in order to maintain a high quality database. Firms are allowed to submit their quarterly reports in two months, and annual reports in three months. Besides, firms are allowed for two additional months for independent auditing process.

The system is based on a comprehensive data set to monitor FX position, cash flow and derivative instrument utilization of non-financial companies which has FX debt over 15 million US dollars. 15 million US dollars of reporting scope threshold has been set due to the fact that the majority of FX debt (%80) has concentrated on ca. 2,000 companies with FX debt of 15 million US dollars or over. 20,000 companies with FX debt under the specified threshold account for %20 of the total FX debt.

After completing the legal basis, the system has been established by the CBRT in cooperation with Credit Registry Bureau of Turkey for installation of system infrastructure, software, maintenance, and operation and support

services. Firms in the scope of the system are obliged to report their company identity, FX assets and liabilities, cash flows, income statements and derivative transactions directly to the CBRT by logging in to the internet address, www.tcmbveri.gov.tr.

All the documents related to the system can be found on the web site. Legal framework, data form and instructions and user's guide are presented in order to meet the users' needs in their process of adaptation to the system. There is also a call center for system users which is operated by Credit Registry Bureau on behalf of the CBRT. This «Central Bank Call Center» platform, where questions about the operation of the system are answered and technical assistance is provided, is an exceptional example among the central banks. Detailed training on the system is provided by CBRT to the call center employees. Since it is aimed to compile high quality data, technical support is provided to the companies in the reporting process and all information requests by the companies are met.

Systemic Risk Data Monitoring System presents the first comprehensive data set in Turkey that contains information not only about FX liabilities, but also about FX assets in the same data set. As a result, the system provides a new regulatory framework for FX risk management based on a company specific data set where risks can be monitored in detail. The system covers not only non-financial companies but also public institutions like municipalities. Broader coverage of the system helps to oversee the overall situation of the real sector.

Reports are collected quarterly aiming a timely monitoring of FX position, balance sheet and income statements of the companies. IFRS reporting format is obligatory so as to provide better understanding of the financial position and performance of the companies. Harmonization with international reporting standards presents the highest quality of the financial tables in terms of comparability. Besides annual independent external auditing is required to increase the data quality.

The system is based on individual financial statements of companies in order to analyze the real FX debt burden on individual companies and their distress in managing the debt.

The system provides comprehensive FX credit information including direct loans from abroad that gives the opportunity to compile integrative and coherent data on liability side. FX assets and liabilities are recorded in original currency distinction, which is one of the most important features of the system. Additionally, FX assets and liabilities are recorded in cash flow distinction (0-3, 3-6, 6-12 months, over 1 year) which enables policy makers to foresee liquidity requirement of the market. Export and import information back to three years is collected in order to calculate the net export amounts and natural hedge positions. In addition, inventory information based on market

value of the inventories in FX assets is compiled. The firms are also obliged to report a separate derivative transactions data form which includes very detailed information on transactions both on organized platforms and over the counter markets.

The data set is kept under the highest security procedures and it is only available to the CBRT's authorized personnel. Instant data set update is provided and cross check with other credit databases is set in order to prevent misreporting or system avoidance. Additionally, dashboards are made available with the intent of data visualization and are updated regularly to enable policy makers to monitor the analyses based on firm level, sectoral level and total level data.

4. Real Sector Data Hub

In order to be able to have a comprehensive understanding of the situations of NFCs, CBRT initiated the Real Sector Data Hub project where company level data from various sources are compiled and linked to the data collected directly from the companies through the Systemic Risk Data Monitoring System.

There are already two main micro level data sets within the CBRT related to the FX liabilities of the companies. The first one is Credit Registry data which is under the directorship of the Bank Association of Turkey Risk Center. This data provides information about all FX and FX indexed loans granted by domestic banks and foreign banks through domestic banks. The main shortage about this data set is that it doesn't cover external loans granted directly by foreign financial institutions. Nonetheless, all companies are obliged to report their direct external loans and payments to the CBRT. This data set within the CBRT provides the necessary information about the direct external loans but it is based on declaration of firms and misreporting is a challenging issue in calculation of total FX debt.

In addition to these data sets of company loans, the CBRT compiles balance sheets and income statements of a sample of real sector companies for the purpose of producing "Company Accounts" statistics. This micro level data set provides annual information about financial situation of real sector firms on solo basis. However, this data set covers only a limited number of firms and is based on the data prepared according to Turkish Tax Regulations and reported on a yearly basis.

Moreover, banks report their customers' derivative transactions to the CBRT. This data set is based on on-balance transactions and provides information about type of the transaction, currency and amount details. However, this data set does not include over the counter transactions. The global financial crisis revealed that OTC derivative markets contribute to the build-up of systemic risk through contagion arising from the

interconnectedness of OTC derivatives market participants and limited transparency of counterparty relationships. One main recommendation as part of the DGI was to accelerate the implementation of strong measures to improve transparency and regulatory oversight of OTC derivatives in an internationally consistent and non-discriminatory way. In Europe, a large part of the G20 data gap reform initiative was formalized in 2012 in the European Market Infrastructure Regulation (EMIR). EMIR introduces the mandatory reporting of all derivative contracts to trade repositories since February 2014.

As a member of FSB, Turkey has established an important part of the necessary regulatory framework concerning these reforms. Takasbank (Central Clearing Party) has been working on a project that would allow OTC derivatives to be centrally cleared. Regulation on operating principles of trade repositories was published and entered into force on 19 September 2018. OTC derivatives reporting has started in September 2018. Additionally, reporting obligation fully compatible with EMIR is planned to enter into force and cover all OTC derivative products in June 2019. All asset classes are planned to be reportable beginning from the second half of 2019. Although this data set will cover all the derivative transactions, it doesn't serve the purpose of analyzing NFCs' natural hedges to cover its foreign exchange risk exposure as of today.

Although there is particular information on the liability side of the balance sheets, the asset side of the balance sheet cannot be analyzed in detail. On the other hand, firms which have massive FX liabilities may have specific features which enable them to reduce their FX risk exposures. The risks may be eliminated through hedging operations or exports constitute a natural hedge for companies which has FX loans. The available data sets however, are needed to be coherently integrated and some of them are by themselves not enough to calculate the real net FX positions of the real sector due to recording standards, time lags and currency conversions. This brings forward the need of a data hub which provides a comprehensive data set for FX risk monitoring of non-financial corporations.

There are ongoing attempts to obtain additional data related to the NFCs' financial conditions. In this regard, trade registry data, credit registry data, public financial statements, foreign trade and investment incentive data, media and text analytics, external credit scoring, financial statements reported to Revenue Administration, employment data, economic tendency surveys, spot exchange transactions are planned to be compiled and integrated in a 360 degree manner and the aim is to constitute a comprehensive data set which provides all the necessary micro level information regarding a company's financial position. A major part of this "360 degree systemic risk data monitoring system" study has been completed. So far, integration of the trade registry data, credit database of the Banks Association of Turkey, financial tables from Revenue Administration, CBRT's economic tendency

surveys and spot exchange transactions data are accomplished. There is an ongoing communication with related institutions to make protocols enabling access to employment information from Social Security Institution and export and import data from Ministry of Trade. As mentioned above, DGI initiative regarding derivatives is planned to be completed in the second half of 2019. The reports will be shared with the CBRT and along with other company specific information, derivatives data set will constitute another main part of the systemic risk data monitoring system. Once the data hub is completed, it will allow comprehensive analysis regarding the financial positions of the NFCs.

5. Discussion and Conclusion

As a conclusion, the global financial crisis of 2008 forced all the central banks to start to compile comprehensive, coherent, timely and high-quality micro data sets which enable them to analyze the systemic effects deriving from changing economic conditions. In this respect, the CBRT has made a significant progress in collecting a wide-range of data and producing comparable, consistent and timely statistics. The statistics produced within the scope of the DGI-2 and the steps taken towards establishing a comprehensive real sector data hub which allows timely monitoring of transactions that will potentially affect the FX position of firms is very important for detecting risks at the micro level and taking the necessary measures.

References

1. Alfaro, L., Chari, A., Asis, G., Panizza, U. (2017) "The Real Effects of Capital Controls: Firm-Level Evidence from a Policy Experiment." Journal of International Economics 108: 191– 210.
2. Bank of International Settlement (2014), "Buoyant Yet Fragile?" BIS Quarterly Review, December 2014.
3. Bank of International Settlement (2016), Debt Securities Data Base. http://www.bis.org/statistics/secstats.htm
4. Financial Stability Board, International Monetary Fund (2017), "The Financial Crisis and Information Gaps" Second Phase of the G-20 Data Gaps Initiative Second Progress Report, September 2017.
5. Gabaix, X., (2011): "The Granular Origins of Aggregate Fluctuations." Econometrica, 79, 733–72.
6. IMF (2015) "Corporate Leverage in Emerging Markets—A Concern?" in Vulnerabilities, Legacies, and Policy Challenges Risks Rotating to Emerging Markets, Global Financial Stability Report, October. Washington D.C.

7. Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories Text with EEA relevance
8. Shin, Hyun Song (2013) "The Second Phase of Global Liquidity and Its Impact on Emerging Economies" Proceedings of the Asia Economic Policy Conference, Federal Reserve Bank of San Francisco

The integration of micro-data sets into a macro-prudential regulatory landscape, exemplified by the AnaCredit regulation

Olivia Hauet
Nordea Bank, Stockholm, Sweden

Abstract

By introducing the AnaCredit regulation, the European Central Bank (ECB) has taken the concept of data-driven regulatory reporting to an unprecedented level. In addition to an ever-increasing burden of traditional template-based reports serving both supervisory and statistical purposes, credit institutions of the Euro area are requested to deliver micro-data on a loan-by-loan and borrower-by-borrower level on a monthly and quarterly basis. The data are compiled according to precise definitions and organised into an entity-relationship model.

This innovative approach to data collection within the banking sector is premised on the following principles: any aggregated data can be obtained from granular data, but the reverse is not true – granular data enabling both drill-down and roll-up analysis; in order to be comparable, data need to be harmonised and standardised by way of common definitions and standards; data are of better quality when they are organised in a model reproducing business reality and input systems to the extent possible; top-down indicators calculated from granular data are not necessarily known in advance as new indicators can emerge from correlations observed between the underlying data.

Ultimately, this approach should alleviate the reporting burden put on credit institutions since new macro-prudential indicators will be possible to compute from already collected data. In the longer term, and to the extent that rules for data dissemination across the national and European authorities allow, it could also establish a bridge between statistical and supervisory needs.

This article explores, by illustrative examples, how some of the caveats of the AnaCredit regulation, from its conception to its application, may impede the fulfilment of its double objective, with regards to the collection of high-quality granular credit data across the Euro area and to the reduction of the reporting burden. It concludes by the necessity to break the remaining silos within banks but also within national and European authorities, and to pursue the work initiated on a larger scale, through initiatives such as the Banks' Integrated Reporting Dictionary (BIRD) and the Integrated Reporting Framework (IReF).

Keywords

granular data, credit data, standardisation, data model, data dictionary

1. Introduction

In the aftermath of the financial crisis of 2008, the lack of comparable data relative to lending exposures in Europe became obvious and urgent to resolve. Notwithstanding the existence of central credit registers in some European countries, the data collected by national authorities were too heterogeneous to be effectively used by policymakers. The interconnectedness of financial markets has put a strain on the ECB to establish a new form of credit data collection which would primarily serve monetary policy, risk management and financial stability purposes. The initial consultations regarding a European granular database containing credit data were launched back in 2011, well before the establishment of the banking supervision in 2014. The AnaCredit project was born – AnaCredit being the acronym for 'Analytical credit data sets' – and the formal decision to go ahead was taken by the Governing Council in 2014. With its 10 interconnected data sets and 88 unique attributes, the AnaCredit regulation became final on 18 May 2016. The first transmissions to the ECB occurred in September 2018, and by 31 March 2019, the full scope of AnaCredit has been implemented.

2. Scope and limitations

The scope of AnaCredit, nonetheless, contains some intrinsic limitations. Initially conceived as a stepwise implementation split in different stages, the AnaCredit project is currently restricted to credit granted by resident credit institutions to legal entities, whereby the following scope limitations apply:

- Credit institutions and branches of credit institutions are resident in a reporting Member State i.e. in a country of the Euro area.
- The reporting is done on a solo basis. Foreign branches of credit institutions resident in a member state report both on a home and host basis, in their own right and through their head office, to the National Central Banks (NCBs) of both countries. This creates situations of double reporting.
- Less significant institutions and foreign branches may obtain derogations from their relevant NCB, by application of the proportionality principle.
- Lending to natural persons is excluded. The exclusion comprises sole proprietors and deceased estates belonging to S.14 (households).
- Credit derivatives, such as loan commitments and financial guarantees are not eligible for reporting until they are converted to a loan.
- Only loans which total commitment amount (sum of outstanding nominal amount and off-balance sheet amount) exceeds 25,000 EUR are considered.

Extending the scope of AnaCredit would require a new decision from the Governing council 'on each subsequent stage at least two years prior to its implementation'. The question of the collection of household data is politically

sensitive, especially after the enforcement of the EU General Data Protection Regulation (GDPR) on 25 May 2018. Household data may be obtained by way of sampling and would in any case be anonymised prior transmission. On a national level, however, NCBs may already apply an extended scope in the primary reporting, which is filtered out in the secondary reporting to the ECB, by including households, non-resident foreign branches, credit derivatives and by removing the threshold.

A realistic mid-term objective of AnaCredit could be to replace some of the aggregated figures required for the Balance Sheet Items (BSI) and MFI interest rate statistics (MIR) by sums obtained from granular credit data. Per contra, an incomplete scope implies an increased reporting burden since banks will need to maintain both types of frameworks, granular and aggregated, in their IT architecture, to report loan data.

3. Data model, normalisation and granularity: an abstract and simplified representation of complex business phenomena

The originality of the AnaCredit data sets is that the data required are organised in a relational model, in which entities are connected through unique identifiers, instead of traditional templates where data points represent aggregates. A data model is a conceptual representation of a target-system. In the process of data modelling, abstraction is the fundamental step1. In its simplest form, the AnaCredit model represents loans granted by a credit institution to borrowers and secured by collaterals or guarantees at a reference date: the financial and accounting facts of the loans are connected to dimension tables describing the customers, the protections and the instruments. For a given reference date, records of every data set are uniquely identified by a compound primary key consisting of the reporting agent identifier, the observed agent identifier (except in the counterparty reference data set), and the counterparty, contract-instrument, or protection identifier.

An important technical feature of the AnaCredit logical data model is that it has a low level of normalisation. Notably, some attributes are applicable on a different level of granularity from the primary key's, and will therefore contain 'NULL' values on a high number of records: in the Counterparty reference data set, the attributes 'Immediate/Ultimate parent undertaking identifier', 'Legal form', 'Status of legal proceedings', 'Enterprise size', 'Annual turnover', 'Balance sheet total' and 'Number of employees' are applicable only if the counterparty is a legal entity in its own right, not if it is a foreign branch. In a de-normalised database, it is difficult to distinguish whether 'NULL' values are due to non-applicability or non-availability, which can complicate the data quality assessment.

Granularity is usually defined as the "extent to which an object or model is broken down into smaller elements'2. In other words, it is the level of detail

contained in the data sets: the greater the granularity, the deeper the level of detail. The definition of the grain is crucial as it conditions future drill-down possibilities. In AnaCredit, the grain of the various data sets is determined as follows:

- The concept of 'Institutional Unit' applies to counterparties: a counterparty is either a legal entity in its own right or a foreign branch. There can be only one foreign branch by country.
- An instrument is defined as the lowest level of agreement within a deal or a facility. It can belong to an umbrella contract which represents the facility.
- A protection is defined as a financial asset or pool of financial assets, a physical asset such as a real estate property, an insurance or a guarantee.

It should be noted that the choice between a borrower-by-borrower vs a loan-by-loan data collection was the object of deep consultations in 2013 and 2014. In spite of higher costs involved, the loan-by-loan approach was selected as there was a clear consensus that such granularity would allow more flexibility in the analysis.

As mentioned above, any data model is a simplified representation of a target-system. In several cases, the grain defined in AnaCredit is not reflecting complex behaviours of the business reality and may lead to summarising errors. Examples of such errors or misinterpretations of the data follow:

- In the instrument data set, the 'Commitment amount at inception' shall be reported only on lump-sum instruments whenever the instrument belongs to a cross-limit structure, due to the fact that the commitment amount at inception is fixed over the lifetime of the instrument by definition. In the case of instruments with an off-balance sheet component such as an overdraft facility, the commitment amount at inception would not remain constant if it was allocated between several instruments. However, since it is expected that the off-balance sheet amount is allocated on the instruments with an off-balance sheet component if all instruments of the same facility are eligible as per AnaCredit criteria, it means that the total commitment amount at inception of the facility level will not be reported until a non-lump-sum instrument is drawn from the limit structure, resulting into misleading off-balance sheet estimations on contract level.
- 'Accumulated impairment amount', 'Accumulated write-offs' and 'Cumulative recoveries since default' are required on instrument level in the Accounting data set. It is common practice to book this information on customer level, especially in the case of individual assessment. The AnaCredit manual suggests redistributing the amount on instrument level in this case. Such allocations are technically difficult to implement as they require the approval of internal models considering the level of collateral and need to browse through all exposures, within and outside the AnaCredit subset, which can create design architectural challenges. The

sum of the allocated amounts is not easily reconcilable, and a better quality would have been achieved by enabling a reporting of these attributes on customer level.

- The modelling of the protection does not cater well for cases where the protection secures loans across different entities of the same reporting agent or group. In this case, the protection is expected to be repeated in the reporting of the different observed agents. In general, the model used for Protection in AnaCredit is too simplistic to represent real estate collateral relationships in many European countries, where the loans are in fact secured by one or many mortgage deeds which in turn are connected to one or several immovable properties. While the mortgage deeds are not directly modelled in AnaCredit, they are distributed over the combination of every loan and asset to which they are connected in the Instrument-protection data set. The distribution is called the 'Protection allocated value' and shall exclude senior liens. Cases where the reporting credit institution has different relative lien priorities with other banks which mortgage deeds pledge the same property are not straightforward to report: in general, a conservative approach is retained, assuming the worst priority in such a scenario.
- The joint liabilities data set covers in fact both cases of joint liabilities, in which co-debtors are fully liable for the whole debt in solido, and of several liabilities in which the co-debtors are liable to the extent of the amount or proportion of the debt they have committed to. There is no indication of nature of the co-liability, being joint or several as part of the data set. If used without the addition of a percentage or without selection of the main debtor, this data set can easily create summarising errors in aggregations due to the repetition of outstanding nominal amounts in the case of solidary debts.
- By introducing a common key for syndicated loans structures (called 'Syndicated contract identifier'), the ECB expects to be able to consolidate the information relative to the different shares of the syndication. However, in the case where some participants of the syndicated loan are not reporting to AnaCredit, the lead bank reporting to AnaCredit is expected to gather the shares of non-resident participants as separate instruments as part of the reporting. Such information is not likely to be easily available and the request itself if questionable. In the case where the lead arranger itself is not reporting to AnaCredit, the shares of creditors not residing in an AnaCredit country will not be reported by any reporting agent, leading to an incorrect consolidation on Syndicated contract identifier level. In the absence of technical and global standards, the quality of this attribute will remain doubtful.

- The attributes' granularity is not atomic, as some attributes mix different concepts, and the domain values are not always disjoint. For instance, the 'Status of Forbearance and renegotiation' in the accounting data set mixes two distinct concepts, where the forbearance shall take precedence. This means that performing loans which have not yet exited the two-year forbearance period, and which are renegotiated are not reported as renegotiated, which can create reconciliation issues with some MIR breakdowns.

In summary, granularity is beneficial when the grain is appropriate. As part of the IReF consultations3, it is worth mentioning that granularity is envisaged on transactional level (instead of balance level in AnaCredit).

4. Standards and semantics: fundaments of a 'single version of the truth'

Truth is traditionally defined as 'the adequation of things and intellect' (Thomas Aquinas). A representation is true if it conforms to the reality it describes. Representations of a unique reality are in fact not always unique, either when standards are missing or concurring, when a perspective is introduced (for example, the Probability of default may be reported for a customer by one entity and not reported by another entity within the same group, if the second entity applies a standardised approach), or when data circulate in parallel flows. Notably, foreign branches affected by double reporting rules in AnaCredit (reporting both on host and home basis through their reporting agent) may be in a situation where the host NCB sends part of the data to the ECB, while the remaining is transmitted by the NCB of the head office. This set up can easily create data integrity issues between the data sets, particularly when the threshold conditions or remittance dates differ between the NCBs (as early as Business Day 6 in Germany instead of Business Day 15 for the majority).

Concurring standards are exemplified by the identification of the counterparties, which, granulated as institutional units, cannot all be identified through the Legal Entity Identifier (LEI), as not all borrowers have recorded an LEI. Initially, the LEI's scope was restricted to legal entities; it has, however, been expanded to foreign branches4 since 2017, provided that the head office has been granted an LEI. Due to this coverage limitation, AnaCredit mostly relies on national business registers to identify counterparties. The golden de-duplicated records are created in the Register of Institutions and Affiliates Database (RIAD) under a unique RIAD code. Similarly to the LEI database, RIAD contains metadata about legal persons reported through AnaCredit but also the Centralised Securities Database (CSDB) and the Securities Holdings Statistics Database (SHSDB), and for which the NCB of their country of location is responsible for the quality assurance. The 'single version of the truth' is obtained by enrichment from different sources: the LEI database, the national

business registers, the European business register and the credit institutions. By doing so, the ECB creates a new repository which data will not necessarily be fully synchronised with the LEI database for the common attributes, such as the legal form and the legal hierarchy. This leads to the next questions: in case of divergence, which database shall be considered as the truth? Will credit institutions benefit from a feedback loop from RIAD in order to update their customer information and thereby improve their data quality? A lack of transparency regarding the content of RIAD would mean that the credit institutions will not be in full control of the reconciliation checks performed by the NCBs and the ECB between their AnaCredit and BSI data: such reconciliations are done within the banks on the basis of their customer data such as institutional sector or NACE code, which may be different from the golden record data.

Another major issue in the process of establishing a single version of the truth is the semantics. Definitions used in the financial industry and by the regulators are still not harmonised, despite major efforts done in this area in the recent years, and AnaCredit has not avoided the 'similar but different' pitfall. A definition is a statement which establishes a connection between a term and a thing or a concept. From a linguistic perspective, AnaCredit uses both intensional (describing the intrinsic properties of the concepts) and extensional (enumerating member values) definitions. For example, 'Amortisation type' is defined as: 'Type of amortisation of the instrument including principal and interest' and through its code list consisting of 'French, German, Fixed amortization schedule, Bullet, and Other'. By doing so, it introduces new taxonomies — for example, 'French', which definition is 'Amortisation in which the total amount — principal plus interest — repaid in each instalment is the same' means "Constant annuity'.

Regulatory definitions are not always harmonised due to different phenomena that will be elaborated in the next sections: two definitions are identical except in their temporal dimension; some concepts are not strictly regulated and leave room for interpretations or refer to internal models; several words can apply to the same concept; lastly, but most importantly, one word can relate to different concepts, even slightly.

The first phenomenon occurs with some of the metrics such as 'Accumulated write-offs', which need to be reported in AnaCredit (as part of the Accounting data set) and in FINREP according to the same definition, but with a different 'until when' condition. They are reported in AnaCredit until the next quarter following a full write-off, or until debt forgiveness if the debtor has other credits which are not fully written-off and which fulfil the threshold condition. This differs from the FINREP condition by which 'these amounts shall be reported until the total extinguishment of all the reporting institution's rights by expiry of the statute-of-limitations period, forgiveness or other

causes, or until recovery'5. Consequently, amounts reported in AnaCredit will be considerably smaller.

The second point can be illustrated with the definition of default: in case of default assessment on obligor level, the AnaCredit regulation does not specify whether to apply the contagion effect to the group of connected clients, or in a joint liability, in the case where the contagion is triggered according to the internal default policy6. The reference of the default definition is the Art. 178 of the Capital Requirements Regulation (EU) No 575/2013 (CRR) which does not contain any condition of contagion to a group of connected clients. Due to the new default definition from the European Banking Authority (EBA) being applicable on 1 January 2021, banks are currently in the process of refining their default policy and it is likely that AnaCredit data will not be fully comparable until then; likewise, the 'Protection allocated value' in AnaCredit refers to internal risk models and is calculated irrespective of CRR eligibility, without any requirement to cap the distributed amount with the loan carrying amount as it is expected in FINREP. It is nevertheless to be anticipated that some credit institutions will have opted for an allocation logic similar to the one reported in FINREP, to ease the reconciliation between frameworks. Figures stemming from different banks will be difficult to compare due to differences in interpretation.

The third point – one concept covered by different terms – is best illustrated by the non-performing loans. AnaCredit requires the data from three different frameworks: the prudential definition of default from the Art. 178 of the CRR in the Counterparty default data set, the accounting concept of impairment from IFRS 9 in the Accounting data set (if the accounting standard used by the reporting agent is IFRS) and the supervisory definition of non-performing exposure in the Annex V to Implementing Regulation (EU) No 680/2014. As explained by the ECB Banking supervision in its 'Guidance to banks on non-performing loans'7, the three definitions are at present aligned for the vast majority of exposures subject to impairment – the non-performing definition being slightly broader than the default due to the one-year cure period, a second forbearance and the pulling effect.

Last, and this is more problematic, widespread cases where one business term is used to describe different realities, within AnaCredit, between AnaCredit and regular contractual terms, or between AnaCredit and another regulatory framework:

- Several attributes of the Instrument data set, such as 'Interest rate spread/margin', have a definition deviating from regular contractual terms: thus, this attribute is only applicable on top of a reference rate, not on fixed-rate loans. In practice, business rules are required to transform contractual data so as to comply with the AnaCredit definition.

- The household sector, which is not in scope for AnaCredit, corresponds to S.14 according to the ESA 2010, with the possible addition of some boundary cases (such as deceased estates). However, the S.15 ('Non¬profit institutions serving households') is included in AnaCredit. The boundary between household and non-household counterparties is different in FINREP, which includes this latter category in the household sector8.
- The 'Date of past due of the instrument' in AnaCredit is required as a separate attribute of the Financial data set, which needs to be consistent with the 'Arrears for the instrument', and calculated without application of any materiality threshold; whereas the 'Days past due' used to determine the reason for the Default status of the counterparty or the instrument shall be calculated once the materiality threshold is reached. By 31 December 2020, banks will have implemented the (EU) Regulation 2918/1845 'in relation to the threshold for assessing the materiality of credit obligations past due'. This means that two different sets of date of past due need to be monitored in the source systems, in cases where initial past due amounts are below thresholds (both absolute and relative).
- The 'Enterprise size' attribute refers to the Commission recommendation 2003/361/EC as a function of 'Number of employees', 'Annual turnover', 'Balance sheet total' and of the enterprise's autonomy. Small and medium-sized enterprises (SMEs) are split between 'Micro', 'Small' and 'Medium-sized' while 'Large' is the residual category. This definition differs from the Article 501(2)(b) of the CRR, by which only the annual turnover determines whether an enterprise classifies as SME; in addition, only the legal entity consisting of the head office and its branches are considered in the calculation in AnaCredit, as opposed to a consolidated perspective implied in the CRR.
- The definition of the 'Commercial real estate' in AnaCredit is based on the CRR, by which the commercial real estate is defined negatively as real estate which is not residential real estate; further, it is split into 'Offices and commercial premises' and 'Commercial real estate' (CRE). Shortly after the publication of the AnaCredit regulation, the European Systemic Risk Board (ESRB) issued its "Recommendation on Closing real estate data gaps' (ESRB/2016/14) on 31 October 2016. In the Recital of the recommendation, the ESRB explains why the AnaCredit data collection is not sufficient to close the real estate data gaps due to its limited scope and definitions. The ESRB's definition of the CRE excludes social housing, property owned by end-users, and buy-to-let housing. Some European Financial Supervisory Authorities have started to enforce the recommendation and to request ad-hoc or regular CRE granular and aggregated loan data from banks whereas the AnaCredit implementation is still ongoing. The promise made

by the ECB statisticians to the banks that AnaCredit would reduce the reporting burden through fewer ad-hoc requests has not yet come true...

5. In conclusion, AnaCredit is a precursor to micro-data-driven reporting, which needs to be completed by other initiatives on a wider scale.

In his visionary speech at the Seventh ECB Statistics Conference, Mario Draghi opened the way for the integration of statistical and supervisory data collection: 'Statistics produced by Central Banks and supervisory data have so far lived in different realms. They capture similar phenomena but often using somewhat different concepts and different reporting frameworks. This will need to change. We cannot afford to have two, somewhat truncated and somewhat incompatible views of the world. It is detrimental to policy making, it is costly to the reporting agents and it undermines the trust to the financial system.'

The AnaCredit regulation has induced huge costs for the credit institutions. The banking industry has made all necessary investments to complete its data landscape and has operated a cultural change to establish robust data quality and governance processes in centralised flows. The considerable work done as part of AnaCredit can be beneficial both for banks and regulators, on the condition that definitions are truly harmonised between European and national frameworks, no matter their statistical or supervisory nature, and that data dissemination allows the reporting of the same phenomena only once to the diverse regulatory actors, using granular data as input.

The coming months will be crucial for the future of regulatory reporting in Europe. Some pioneering countries show the way: Croatia and its unique dictionary serving both statistical and supervisory purposes9, Austria with the Aurep platform co-owned by 7 Austrian banks, by which reporting data is pushed from a basic and redundancy-free cube to the Austrian National Bank (OeNB), Italy with Puma 2, but also outside the European Union, Rwanda10 and its 'data-pull approach', have been successful in designing different solutions to streamline the banks' reporting processes and thereby enhance the use of data. On European level, the banking industry and the regulators need to increase and speed up their collaboration in order to finalise the work started on establishing the BIRD11. The BIRD is an interesting transitory attempt to reduce the reporting burden by listing the data requests in a modelled input layer, and by delivering transformation schemes to frameworks of different nature. However, the BIRD does not have the legal status to own definitions, nor supersede/precede the regulatory demands; hence, being output-driven in the sense that the output frameworks determine the input and intermediary layers by a 'reverse engineering' logic, the BIRD currently comforts the template-based approach of the EBA and the

different codifications, handled through transformation mappings to the reference layer, instead of addressing the root cause of the misalignments. Therefore, in a second step, or rather in parallel, it will be essential to establish a proper joint committee, in collaboration with the banking industry, which would define all regulatory requests into a common, single repository, with full alignment of concepts and taxonomies, preferably granular and data-driven.

In this regard, it is too early to depict AnaCredit as the symbol of a paradigm shift in regulatory reporting practices, but it is most likely a precursor of the regulatory landscape's evolution to come on a wider scale, which has enabled the banks to prepare themselves by establishing best practices in data governance. No doubt the future will open for more data, but also better data, and the journey is probably only starting.

References

1. Fishwick P. A. (1988). The role of process abstraction in simulation.
2. Hehenberger P. (2014). Perspectives on hierarchical modeling in mechatronic design. Advanced Engineering Informatics.
3. Qualitative stock-taking questionnaire on the integrated reporting framework. Analysis of high-level considerations and high-priority technical aspects' (2019). ECB.
4. LEI ROC statement of purpose (2016). Including data on international/foreign branches in the Global LEI System.
5. Regulation (EU) 2017/1443 of 29 June 2017 amending Implementing Regulation (EU) No 680/2014 laying down implementing technical standards with regards to supervisory reporting of institutions according to Regulation (EU) No 575/2013 of the European Parliament and of the Council.
6. Final Report on Guidelines on default definition (2017), paragraph 61. EBA-GL-2016-07.
7. Guidance to banks on non-performing loans (2017), section 5.1. ECB, Banking supervision.
8. Annex V to Implementing Regulation (EU) No 680/2014, paragraph 35 (f).
9. Basic I. (2017). Supervisory and statistical granular data modelling at the Croatian National Bank. Statistics paper series of the ECB.
10. Broeders D.. Prenio J. (2018). FSI Insights on policy implementation No 9 Innovative technology in financial supervision (suptech) – the experience of early users. Financial Stability Institute.
11. https://banks-integrated-reporting-dictionary.eu.

Measuring the data universe: The challenges of data integration in a time of exploding data worlds, successful approaches using Statistical standards, Bundesbank's experience

Dr. Patricia Staab
Deutsche Bundesbank

Abstract

We live in a time of exploding data worlds. Both the data available worldwide and the technical possibilities to work with them are growing constantly and rapidly. However, the task to extract information from the data has not become easier but more difficult. A big challenge is that the data, coming from many different sources, do not fit together from the start. They have to be integrated.

Data integration, whether in-house, cross-domain or international, is a task with many challenges. Obstacles may come from the IT industry, from data producers or data users, and from all other stakeholders. But it can succeed if one chooses the right course of action. One success factor is the use of standards; standardization is a central step in the data integration process.

Statistics' own data and metadata standard SDMX can form a reliable basis for successful data integration. Since its introduction in 2001, SDMX has been successfully used in international data exchange. But the true potential of SDMX lies in the underlying information model. It enables cross-divisional data usage and the creation of modern, generic analysis systems.

The Deutsche Bundesbank has been successfully using SDMX for years; it is the basis of their Central Statistical Infrastructure and the corresponding information system. This Central Statistical Infrastructure also hosts the Deutsche Bundesbank's House of Microdata (HoM), an institution-wide microdata based information and analysis system.

Keywords

Data Integration; Standardization; SDMX; Statistical Data and Metadata Exchange; Microdata

1. Introduction

The data universe is exploding. The global amount of digital data which is available online is growing constantly and rapidly. Information is not only generated by automatic process recording, e.g. via sensors ("Internet of Things") or search bots, but also eagerly provided by users of social networks and search engines, mobile phones and tablets.

The supply of data is met by an equally thriving demand: a sheer "numbermania", that is a fixation on quantitative information. We strive to

measure as much as possible – in our private life, this may mean sleeping hours, calories intake or step count. In business, there are new performance indicators or benchmark values every month.

Luckily, there are new technical and methodological developments which enable us to deal with the enormous volume of data. Additional computing power comes from technology innovations like Big Data, and new AI (Artificial Intelligence) driven analysis techniques like Machine Learning contend to help us get the answers we seek.

Yet, despite our "collectomania" we may still find ourselves in situations where we feel we don't have the data that is needed. This is partly because data are often not collected where it's needed, but simply where it occurs. Another reason is that the data universe lacks one ingredient which is crucial for connecting the dots and creating information out of data: Order. The IT world possesses neither a general system of order, nor a prominent generic data standard, nor a global identifier for information itself. The lack of a "barcode for information" leads to proprietary solutions for specific countries and industry branches. Many companies feel the need for a "global" order system for relevant information which is currently stored in various data silos across different business areas. This often leads to huge data integration projects, business intelligence projects, or data warehouse projects. In some instances, those projects are accompanied by the introduction of a Chief Data Officer, in the hope that a central management can create and maintain order.

However, ordering the data is not a means in itself but the prerequisite for the actual task at hand: to integrate the data.

2. Methodology

Data integration is the act of collecting data from different sources, combining it properly and thus creating a comprehensive data collection. The right combination of data naturally leads to new intelligence. However, the data integration projects of companies or institutions more often than not are started with high ambitions and fail to deliver on the expectations. To understand why this is the case one has to take a closer look at the process of integration itself.

From the user's perspective the difference between data integration and data processing is not always clear cut. But usually there are the following activities to observe (see Figure 1): First, the standardization, then the linking, and finally the simplification of the data. Standardization means transforming the data in a way that – independent of source or origin – the data describe the same things in the same way. Standardization enables the linking of the data, i.e. the data are merged to form a comprehensive and complex data structure. Usually only a specific part of this data structure is required for analysis, so in order to avoid having to deal with the complexity, the data

needed is extracted and simplified. Very often the linked "super structure" is not even realized or stored but remains a virtual entity.

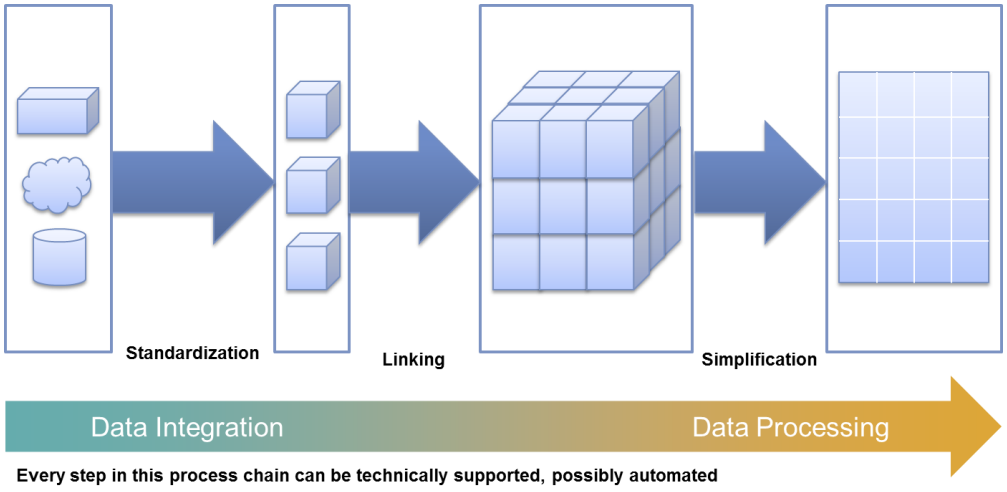


Figure 6: The production line of data processing and integration. R. Stahl, P. Staab (2018)

Now let's take a closer look at the data integration process, which transforms the original data into ready-to-be-linked data. This process can be divided into three steps (see Figure 2). First, the **logical centralization** is the act of storing the data (physically or virtually) in a common system. It's a very technical step, and it ensures that common procedures can be used e. g. for administration, authorization and access. Second, the **formal standardization** means the remodeling of the data according to a modeling standard, so that a uniform language – the same concepts and terms – is used to describe the data. This enables a rule-based (and automatable) treatment of the data. However, this does not yet ensure that the data describe the same things in the same way. This is done in the third step, the **semantic harmonization** of the ontologies, dictionaries, identifiers and classification systems used in different sources.

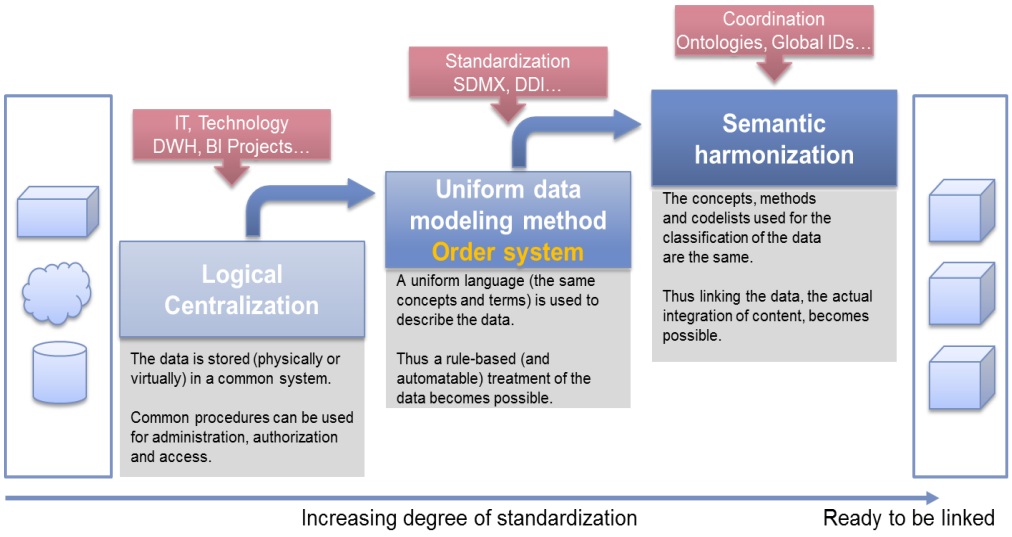


Figure 7: The three steps of data integration. R. Stahl, P. Staab (2018)

To sum up, data integration is a process which consists of multiple steps, and, as a consequence, often involves numerous stakeholders. So it does not come as a surprise that data integration endeavours are faced with various challenges:

Challenges might arise from the IT industry itself. The staggering innovation speed in the field of IT technology tends to leave little time for the establishment of comprehensive content-related standards. Also, the existing IT standards for data are not focused on data integration but on specific use cases. In companies, the pressure to introduce new technology faster than competitors has repeatedly led to a preference for projects which focus on IT features like Business Intelligence, Big Data, or AI, instead of specific business needs.

Furthermore, companies are very rarely homogenous. Usually they are divided into silos, making silo culture one of the biggest challenges for data integration. Operators of well-functioning silo systems often possess a special hybrid expertise, a compound of technology and business knowledge mixed with years of experience. So, not only do they hesitate to welcome a loss of their seniority, but they are also tough competitors to beat. Therefore, in order to be accepted by users, an integrated system has to be at least as good as the pre-existing silos. For IT engineers then again, the silos might have provided a quite high degree of individual freedom, creativity and efficiency, which makes it harder to inspire them for standardization.

On top of that, data users can form a sort of challenge for data integration. Naturally, analysts and researchers are not interested in the data production or data integration process in itself, but they need its result, namely a specific data set containing the information they need to perform their research.

Ideally, the data should be tailor-made according to the current need. This result, however, can in most cases not be automatically generated, even from the most integrated data worlds and well-arranged data structures. Thus, the final step of selecting and formatting the output still has to be done individually, even manually.

Other challenges arise from privacy and data protection regulations, leading to strict requirements regarding the confidentiality of data, which tend to become even stricter for linked data structures.

On the other hand, public data collecting institutions are prone to be underfunded and therefore inclined to focus on their actual tasks instead of fostering cross-domain standardization and data integration initiatives, with the notable exception of research data centres.

Despite the various mentioned obstacles, data integration is a highly important achievement that pays off in the long run. The central part of the integration process, its essence, is the standardization of the data. Since a large part of the resistance against data integration will be focused on this central endeavour, now some basic thoughts about standards.

Standards do not fall from the sky. It is important to be aware that a new standard almost always replaces something pre-existing – a quasi-standard or a singular solution –, and the migration to the new world means extra effort. This effort is especially annoying because **a standard is never the local optimum, even though it might become the general optimum.** Almost every specific business need could be covered better by an individual solution. So, for operators of pre-existing silo systems the migration to a new standard almost always means a deterioration of functionality or service. In order to win the stakeholders over, one must bear in mind that **standards are only accepted when they are used.** A standard can be established when it manages to get a significant "market share" of users. Then, standards can play to their strengths.

In order to successfully introduce data standardization and data integration, here are some best practices:

Start with the content	Always begin by understanding the content of your data. Define your integrated data model independent of products, platforms or technology. Only then think of the practical realization. Many projects neglect the first step and fast-forward to the second. But the correct sequence is crucial to avoid incomplete or useless systems.
Use global IDs and classification systems	The lack of a universal identifier can be a showstopper for any data standardization effort. Thus, when modeling your data, always use the most common, better global, identifiers and classification systems. Also, support the introduction of global identifiers and classification systems where they are missing.

Use technology wisely	The marketing promises of the IT industry suggest that any problem can be dealt with as long as the hardware and software are strong enough. But some challenges cannot be solved by money. IT cannot be a substitute for a well-thought model for data and processes. First, there must be an intelligent design, then the implementation on one or more IT platforms will follow.
Choose small steps	Data integration is a long-term-effort where eventual success will be reached by evolution rather than revolution; it calls for a step-by-step approach, a strategic directive, instead of an overambitious single project. Don't demand that all (!) company data enter the central data warehouse at once. Instead enable slow growth: create a space for integration, let different topics move in in their own time.
Treat stakeholders right	The dilemma of data integration is that data providers are expected to carry the burden, but have no direct benefit. Thus, have a clear understanding, to whom the placement of information in a central data warehouse offers a real added value, as well as of everybody's role. Then, do the necessary work to convince people and to justify the effort. This requires a sense of diplomacy, persuasiveness, but also tenacity and – most of all – patience.

Examples of successfully using standards for data integration can be found in the field of Statistics. Statistics is cross-domain by nature and serves as an auxiliary science in many fields. Therefore, it is statistical day-to-day business to integrate a wide range of information sources.

One standard that has been effectively used by the global statistical community for years is the SDMX (Statistical Data and Metadata Exchange) standard. SDMX is an ISO standard (ISO 17369) launched in 2001 by its sponsor organizations BIS, ECB, Eurostat, IMF, OECD, UN, and World Bank, to advance the international exchange of statistical data. The initiative was quite successful: Today the (previously bilateral and domain-specific) exchange of statistical financial and economic data between the sponsor organizations and the associated countries is SDMX-based and standardized.

But the true potential of SDMX lies in the underlying information model, which can be used for modeling data of any business domain. SDMX allows for a quite intuitive building block approach to data modeling (see Figure 4). Today, several institutions have thriving SDMX data collections.

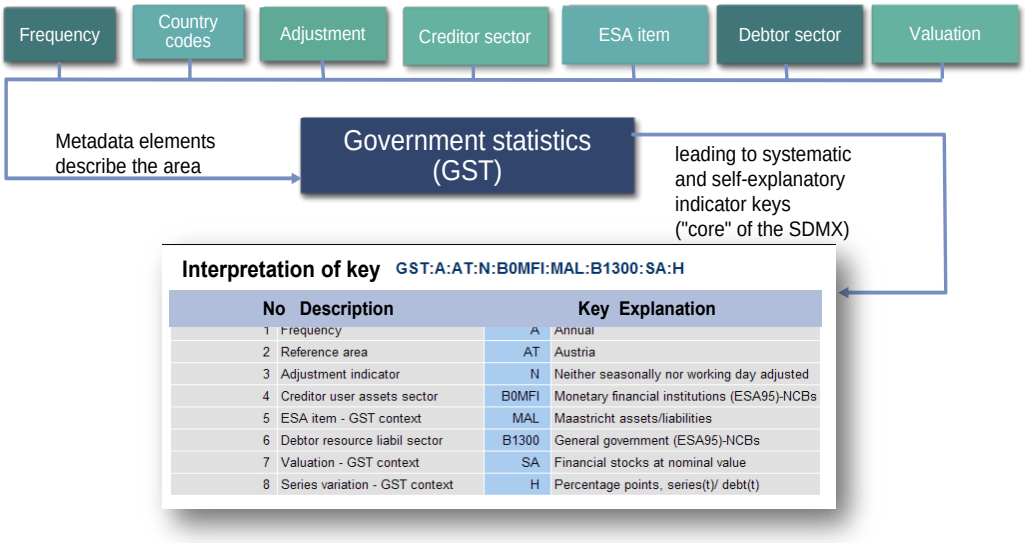


Figure 8: SDMX in a nutshell: Setup of a work area according to the building block approach

3. Results

The Deutsche Bundesbank has also been using the SDMX standard as a basis for data standardization and integration since 2003. It is the core of its Central Statistics Infrastructure, a data warehouse environment including a comprehensive statistical toolset for operational and analytical tasks. The Central Statistics Infrastructure is open to any kind of data and can therefore also be used for data pools outside the Directorate General Statistics. Standardized interfaces ensure that users can employ their own evaluation instruments rather than the statistical toolset provided by the Central Statistics Infrastructure.

Over the last years, more and more domains have been integrated, resulting in approx. 450 datasets, containing well over 160 million time series, and making the Central Statistics Infrastructure one of the most used data sources within the Bundesbank (see Figure 4).

As an internal system, the Central Statistics Infrastructure is open only to employees of the Deutsche Bundesbank and to connected institutions of the German government. Still, it manages more than 1,500 active users, 200 of which access it on an average day. Daily data traffic exceeds one million time series.

In addition to datasets from different Directorate Generals of the Deutsche Bundesbank, the Central Statistics Infrastructure contains datasets from national and international institutions such as the ECB, the German NSO, Eurostat, OECD, BIS, IMF, as well as licensed data.

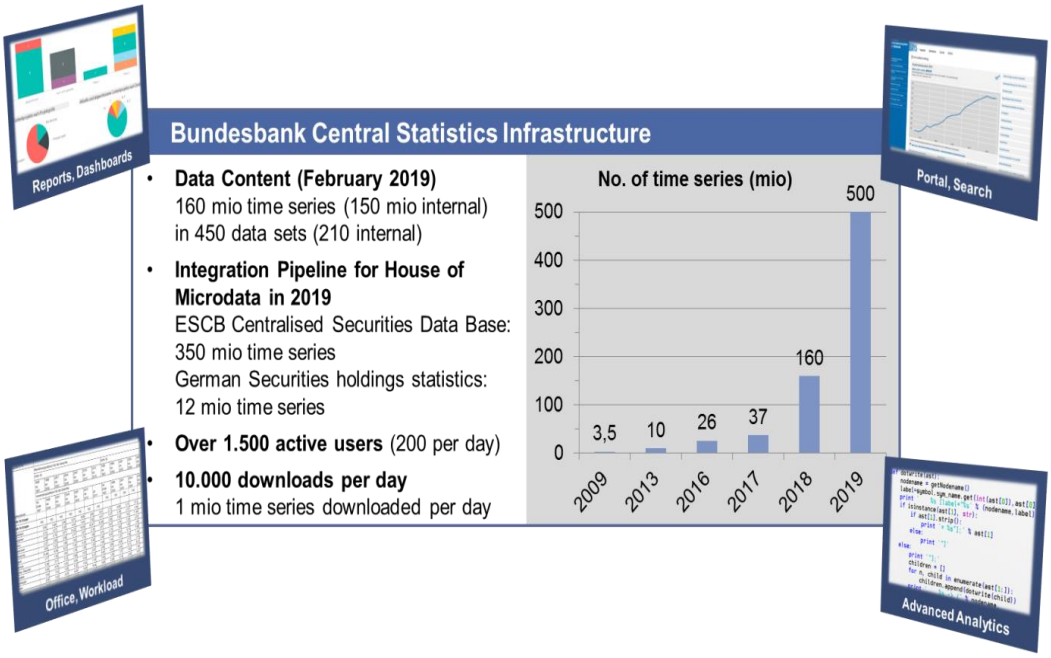


Figure 9: The Deutsche Bundesbank's Central Statistics Infrastructure

The SDMX standard provides a suitable framework for the integration of heterogeneous data. The multidimensional, generic approach of the SDMX information model offers an ideal means of linking and comparing data from different sources by using uniform code lists. So it does not come as a surprise that SDMX can also be used for microdata without restrictions. This aspect of SDMX was put to use in the Deutsche Bundesbank in 2013, when the Directorate General Statistics was mandated to establish an integrated interdepartmental information system for analytical and research purposes based on microdata for various user groups (financial stability, research, monetary policy, supervision). This goal was achieved by establishing a Research Data and Service Centre (RDSC) and developing a data warehouse for microdata called "House of Microdata" (HoM). The Directorate General Statistics created this "House of Microdata" on SDMX and the Central Statistics Infrastructure.

4. Discussion and Conclusion

In a time of exploding data worlds integrated data are the prerequisite to create knowledge from an abundance of information. However, the endeavour of data integration encompasses various stakeholders and faces a multitude of challenges.

A necessary component of the integration process is standardization, ie the introduction of a uniform data model in combination with an order system for data. Current efforts in the field of semantic harmonization of data, such

as dictionaries, repositories and ontologies, unfold their strength only when they are supported by a strong modeling standard. It can form the bridge to the technical implementation and support the realization of IT systems and automated workflows for the collection, storage and processing of data.

SDMX is such a standard. It has been successfully used in the international statistics community. In the Bundesbank, it drives the Central Statistics Infrastructure as well as the in-house microdata integration process.

The Statistics community could rely even more on this standard to drive further data integration efforts. That means to use SDMX not only for the sharing of statistical data between institutions, but for all the steps of the statistical business process. Furthermore to classify new data collections, even microdata, in SDMX.

Also, the sponsor institutions have been fostering the creation of (mostly open source) SDMX software. This should be intensified, in order to increase the amount of available software, programming libraries, consulting services, even platforms, for concrete use. After all, a standard is just a theoretical concept and lives only through practical implementations.

And last but not least, active marketing is necessary to raise the publicity of the SDMX standard beyond the statistics community and especially in the software industry. That includes the provision of not only more but more diverse documentation, which should not only address statistical information managers but also non-experts and deciders. Because the impact of a standard is driven less by the "genius" of its concept but by how intensively it is being used.

And, intensively used, SDMX can be a reliable basis for successfully measuring the data universe.

References

1. R. Stahl, P. Staab (2018). Measuring the Data Universe. Springer; 1st ed. 2018 (May 28, 2018)

Building a standardized taxonomy between financial reporting and macroeconomic statistics – A South African perspective

Lisa de Beer[1]
South African Reserve Bank; Pretoria, South Africa, Lisa.debeer@resbank.co.za

Abstract

Accounting and financial reporting is the "language" of corporations. This involves the recording of economic transactions within an entity/group of entities in a specific period based on general accounting and financial reporting principles. On the other hand, the macroeconomic statistics value chain commences with the collection of a wide variety of data, the cleaning and interpretation thereof and the compilation of statistics for user consumption based on international accepted methodology. A commonality is that the output of both are used to make inferences about the entity/ies they represent. In the case of statistics this extends to economic industrial and institutional sectors and the economy as a whole. In doing so, these two fields serve different applications and users, while inextricably linked to one another. Although many of the foundational principles in the two fields are uniform such as relevance, timeliness, reliability, comparability and consistency there are also significant and material differences. This duality is the essence of the paper, namely building a framework between the two to identify the common factors as well as the differences in semantics and structure as well as assessing how this bridging framework can be utilised to improve the statisticians understanding of the accounting/financial reporting[2] taxonomy and to utilise accounting and financial reporting micro data to compile macroeconomic statistics. The paper also provides some practical examples of work done at the South African Reserve Bank (SARB) to harness the commonalities in the two fields as well as to bridge the gap between them.

Keywords

Macroeconomic statistics; Financial reporting; Micro data; Respondents; Harmonisation

1. Introduction

In an ideal world underlying micro building block data should feed into the aggregated reporting frameworks of respondents, such as balance sheets,

[1] The views expressed in this paper are those of the author alone and not of the South African Reserve Bank.

[2] Although the author acknowledges that accounting and financial reporting are two related but different disciplines, in this paper these terms will be used interchangeably.

income statements, subsidiary ledgers, etc. as the ultimate source data for macroeconomic statistics. This micro-data largely reflect the national (and where applicable international) accounting, supervisory and taxation standards and as a result may be different from what is required for macroeconomic statistics. Although the income statement and balance sheets within the financial statements, as specified by the International Financial Reporting Standards (IFRS), and that used for the compilation of macroeconomic statistics have many commonalities, there are also very specific differences in concepts and terminology. In the absence of access to micro data, the challenge faced by economic statisticians is how to transform the data obtained from accounting/financial reporting records into macroeconomic statistics. The purpose of this paper is to present practical solutions and to recommend from a central bank perspective how to bridge the gap between financial/regulatory and macroeconomic statistical reporting.

2. Discussion

2.1 Reasons for differences between financial, regulatory and macroeconomic statistics reporting

Reconciling the three concepts requires acknowledging the different purposes which drives their designed. According to the IASB's Conceptual Framework for Financial Reporting, the objective of financial reporting is to provide financial information about the reporting entity that is useful to a wide range of users in making business decisions. The focus is generally on the examination of resource usage, cash flow, business performance and the financial health of an entity. IFRS aim to provide a single set of understandable, enforceable and globally accepted financial reporting standards which are based upon clear principles. Regulatory reporting on the other hand aims to ensure, amongst other factors, systemic stability in the financial system and consumer protection. This is normally backed by a sound legal framework. In contrast to this, macroeconomic statistics record non-financial assets as well as stock positions and flows of financial assets and liabilities between the various domestic sectors of the economy as well as between the domestic sectors of the economy and non-residents, with a particular focus on the relationship between the institutional sectors through macroeconomic aggregates. It is a key input for decision making in monetary and financial stability policy. The 2008 SNA is the overarching macroeconomic statistical framework for national accounts, external sector statistics, government finance statistics, and monetary and financial statistics. Since the three types of reporting are prepared for different purposes and according to different methodological guidelines that are only partially harmonised, differences exist

in data sources, valuation principles, the content and classification of financial instruments as well as in the treatment of certain economic events. Despite some of these differences, there is generally a closer alignment between regulatory and financial reporting as they both adhere to financial reporting standards where risk identification is important. By contrast macroeconomic statistics diverge from the aforementioned for a variety of very valid reasons.

2.2 Mind the gap - Identifying the differences between financial and macroeconomic statistics reporting

Valuable work has been done by international bodies e.g. the Joint Expert Group on Reconciliation of credit institutions' statistical and supervisory reporting requirements (JEGR) in an attempt to bridge elements of the statistical and supervisory reporting frameworks relating to monetary financial institutions and to identify possible reconciliations The International Monetary Fund (IMF) also offers insights on the relationship between monetary and financial statistics and the IFRS in the latest version of the monetary and financial statistics manual and compilation guide. These initiatives serve as a good basis for identifying conceptual differences and they also offer possible reconciliations. They are however specifically focussed on monetary and financial statistics and do not always provide sufficient guidance on the practical implementation within a macroeconomic statistical compilation environment. Apart from the conceptual differences, there are specific instances which generally result in variances between the different reporting environments and thus warrant explicit mentioning.

Trading and banking book distinction

In regulatory reporting, financial instruments are classified between the banking and trading books based on the intention to trade versus the intention to hold until maturity. By contrast, in macroeconomic statistics, financial instruments are classified in a specified instrument category based on its characteristic traits without any distinction based on intent. In order to overcome any inconsistencies which could arise as a result of the aforementioned difference, it is essential that respondents and compilers are aware of the specific requirements and any divergences in requirements. This process may need to be repeated at regular intervals as respondents generally do not have any statistical background.

Credit impairments

In macroeconomic statistics, loan asset values are presented on a gross basis and data on expected loan losses are included as memorandum items to ensure that the realizable values of loans can be calculated. However, under IFRS loan asset values are directly adjusted for impairment based on the

concept of the Excepted Credit Loss model which requires entities to take into account all information that is available, including information that is forward-looking, when determining impairments that should be raised. This fundamental difference will result in a substantial discrepancy in macroeconomic statistics if not adjusted. Possible solutions to this problem could be the use of bridging tables or accessing granular data, which is discussed in the next section.

Nominal versus market or fair value

Under IFRS, deposits and loans are valued at market or fair value, thus with amortized cost using the effective interest method. Amortized cost in IFRS can be defined as follows: "value at inception" minus "any repayments of principals (for loans)" plus "accrued interest" minus "reduction for impairments". In macroeconomic statistics deposits and loans are valued at nominal value which is the outstanding amount the debtor owes to the creditor. This comprises the outstanding amount including accrued interest (i.e., interest accrued but not yet paid). Nominal value is defined as the value as of date (equal to value at inception minus repayments) plus accrued interest. Deposit and loans shall not be netted against any other assets or liabilities.

Netting and off-setting

In macroeconomic statistics, respondent institutions are required to record the full amounts of asset and liability positions without netting. For example, in monetary statistics, loan and deposit amounts with respect to the same counterparties - if a bank's customer holds both a deposit balance and a loan balance, each of these positions should contribute in full to their total reported deposits and loans, opposed to reporting a net position contributing only to one side of the balance sheet. This adheres to the general reporting principles specified in the IMF manuals, namely that assets and liabilities should be recorded on a gross basis. Under IFRS, financial assets and liabilities can be offset, and the net amount reported, when an entity has a legally enforceable right to set off the amounts and intends either to settle on a net basis or simultaneously.

Unlisted shares

Under IFRS 9 and 13, unlisted shares are required to be measured at fair value. This is a significant problem area due to the inherent nature of unlisted shares dictating that they are not regularly traded and negotiable. It is often the case that a parent entity holds the unlisted shares of its subsidiaries, affiliates, etc. Due to the fact that there is no active market for unlisted shares with regular price quotes it creates problems in determining a market related price for these instruments. In very limited circumstances, IFRS 9 permits an

entity to use the cost as an appropriate estimate of the fair value of unlisted shares. Valuing unlisted shares under IFRS is complex and the input of valuation experts would likely be needed to determine fair value for subsequent measurement. The 2008 SNA and other statistical manuals encourage the valuation of unlisted shares at market prices and if not possible, recommend the use of a flexible approach. A number of alternative methods for the approximation of the market value of unlisted shares are suggested including recent transaction price, net asset value, present-value approach, market capitalization method, own funds at book value and apportioning global value. It is thus obvious that discrepancies will arise from the different valuation approaches used within the two reporting frameworks.

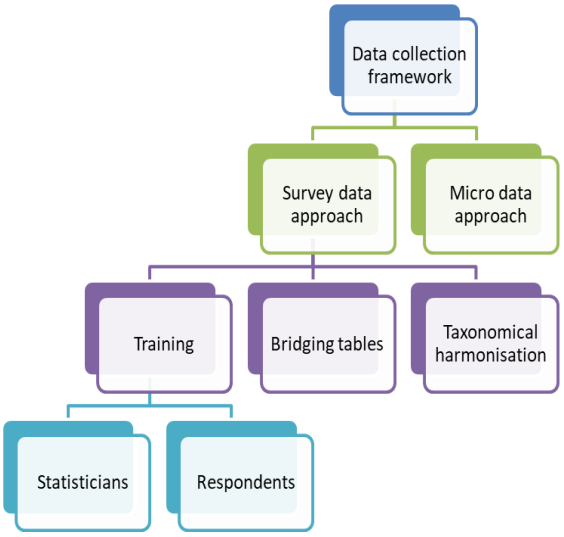
The aforementioned differences do not represent an exhaustive list as there are numerous others which have not been discussed in this paper.

2.3 Bridging the gap – from accounting to statistics

Data collection framework

A fundamental question that needs to be answered in bridging the gap between the accounting and statistical frameworks relates to the design of the data collection process. Figure 1 below provides a summary view of two data collection approaches – that of conventional survey data and micro data. The first approach is conventional sample surveying where data is collected using entity specific surveys that require data cell population of a pre-designed survey by respondents, usually designed for a specific macroeconomic statistical domain. This strategy presupposes that data aggregation will take place at the respondent's level and that it will be done correctly. In the traditional system of official statistics, reporting forms are generally used to obtain data from specific statistical domains. The data cells in the surveys are used to compile specific domain statistics in a pre-defined aggregation format. The alternative to this – and the approach recommended by the author – is that of micro data sourcing.

Figure 1: Choice of data collection framework



One fundamental requirement relating to the use of micro data is that it should be accompanied by detailed metadata so that the users are fully aware what the micro data means and how to interpret it. Once this requirement is met, micro data has numerous advantages. Micro data allows statistical compilers to identify and resolve inconsistencies between data compiled in different institutions, for example the inconsistencies referred to in this paper which arise due to differences in reporting frameworks, methodologies and terminology, while possibly also reducing the burden on respondents. Given the various possibilities, the micro data approach to compiling statistics may even replace the traditional system which would have a significant cost saving implication for respondents because they will only submit data once while the rest of the value chain process is done by the statistical agency. This implies that the aggregation cost is borne by the statistical agency. In addition, should stakeholders' needs change, the availability of micro data will allow statisticians to compile aggregates in flexible ways that meet changing needs without sending new data request to respondents. However, micro data initiatives require a significant amount of investment in terms of coordination, alignment and resources. In the absence of micro data and based on a continuance with conventional data sourcing, the following aspects need to be addressed to ensure that the gap is adequately bridged.

Training statisticians

The first focus in bridging the gap lies with training of statisticians in the art of data collection and statistics compilation. This type of training is a complex and life-long process which has both static as well as dynamic foundational aspects. Modern-day macroeconomic statisticians have to be multi-skilled with their attributes covering a number of specialisation fields.

These fields relate to statistics, economics, accounting/financial reporting and computer/data science. Due to the fact that the respondent data is generated from the accounting world, the obligatory requirement is that the statistician engaging with the respondent should be adequately trained in the field of accounting and financial reporting. This is needed to firstly comprehend the world of the respondent, to guide the respondent as to the transformation of accounting data to statistical data and lastly to stay abreast of occurring changes in financial reporting standards' implication for macroeconomic statistics. The most appropriate way to do this is to have targeted training interventions for statisticians that bridges the gap between accounting and statistics.

Educating respondents

In addition to the above mentioned aspect, the respondents responsible for providing source data also should have a working knowledge of both sets of standards and the dissimilarities between the respective reporting frameworks. Given the fact that the completion of surveys – even though compliance might be mandatory in some instances – will continuously be a cost item for respondents, the onus will always fall on the statistician to drive the quality assessment process. This should however always be done in conjunction with the respondent stakeholder and as a partnership rather than a one-directional relationship. This will require the development of a framework to assess the areas where differences between the two standards are most likely to occur, and investment in the training of respondents, coupled with statistical audits being performed. This initiative should be part of a larger stakeholder management framework throughout the statistical value chain. One approach could be to conduct random selections of respondents who will be subjected to a survey cell assessment. This assessment will only focus on cell data where known differences between the two frameworks exist. An advantage of this process is that common mistakes and interpretation anomalies can be detected and eliminated through targeted training of respondents. Proper treatment practices can then be shared with other respondents through methodological guidance notes to ensure progressive awareness and training of respondents.

Bridging tables

In the absence of micro data and with the use of conventional surveying techniques it is very important to be able to bridge the gap between data required for accounting and data required for statistics. One method which can be particularly helpful relates to the use of bridging tables. Although there may be differences between the aggregated data for the two frameworks, it is always important to be able to qualify the reasons for differences – especially

in times of increasing divergence. Bridging tables can potentially fill this gap – they can be used to derive statistical balance sheet information from accounting balance sheets by mapping asset and liability categories between the two types of balance sheets. This is not always an easy task due to various reasons including the lack of a one-to-one correspondence between items, differences in recording practices and valuation rules, etc.

Taxonomical harmonisation and standardisation

An important part of the above mentioned process of bridging the gap relates to the alignment of terminology between the respective reporting frameworks. One of the core contributory problems in incorrect data submission pertains to the incorrect interpretation of domain specific terminology. What is needed is a uniform and central taxonomical framework where all relevant components are classified, defined and taxonomical bridging between accounting and statistical terminology is achieved. One key feature of such a framework should be the self-service aspect and ease of usability from a respondent point of view. The framework should be housed within a modern technological platform, with user friendly interfaces. Underlying this outer body would be the database of terminology, concepts and metadata which can be accessed and interrogated with ease and accuracy. It would be ideal if this taxonomical framework could be international of nature with alignment between countries as well. This would facilitate more comparable statistics and where differences occur, underlying reasons could be sought based on taxonomical evidence. One statistical domain where this might have significant benefits relates to external accounts where multinational enterprise treatment across countries could potentially be more closely aligned resulting in less statistical discrepancies in international external accounts. Thus, such a taxonomical framework would not only facilitate the bridging between accounting and statistics but also between statistics of different countries.

2.4 Practical examples from a central bank perspective

Raising awareness of the differences

The SARB is currently in the process of re-assessing various components of its statistics compilation framework. Two noteworthy areas in this regard relate to the revamping of its surveys and training of its staff. Currently all of the units responsible for specific statistical domain data will revamp their surveys to be aligned with the guidance manuals and in the process they will compile comprehensive user friendly respondent guidelines that accompany each survey. Apart from providing guidance on how to complete the required survey forms, these guidelines will aim to address the major taxonomical

components between accounting and statistics and will guide the respondent as to the mapping between the two frameworks. The domains that have already commenced with this process relate to the non-monetary financial institutions, the deposit taking corporations and the external accounts.

Training initiatives

The SARB has also invested in a training programme that aims to bridge the knowledge gap between the accounting balance sheet and flow data and the statistical balance sheet and flow data. This has been done through two training interventions in 2018 where an IFRS expert provided in-depth training to statisticians for a whole week at a time, which also included discussion sessions covering IFRS alignment/divergence from requirements of statistical manuals such as the SNA2008, BPM6, etc. The objective of the training was to raise the level of awareness of the statisticians, and to improve their ability to provide guidance to the respondents in current and revised survey forms.

2.5 Challenges

Even if detailed guidelines are provided, respondents reside in a world where financial reporting dominates, with their thought patterns predominantly formed by IFRS guidelines. As mentioned earlier, one way to assist respondents with the interpretation of and alignment to the macroeconomic statistical guidelines and requirements would be training themes that originate from random statistical audits. This will have to be an ongoing endeavour and should be clearly defined and stipulated in the statistical agency's stakeholder engagement framework. While the first challenge can be addressed at the level of national statistical agency, the second ideally needs coordinated intervention from an international body. The problem resides in the fact that there is no central body that can assess and interpret international financial reporting changes and provide guidance to compilers as to the implication these changes could potentially have for statistical guidelines – an example of this would be the amended treatment of leases in IFRS16 and the resultant divergence between reporting for accounting and statistical purposes. An added complication relates to the implementation of the revised standards - even if such an organising body does assessments and makes proposals, in many cases countries have discretion to implement the guidelines and recommendations which has far-reaching implications for comparability of global statistics. In addition, developments such as financial innovation and globalisation have posed particular challenges for data collection exercises by central banks. These factors have made it more difficult and costly for central banks to collect data through full reporting, not only because there are new financial instruments

to measure but also due to the more complex nature of transactions as well as larger numbers of counterparties involved.

3. Conclusion

The 2007/8 financial crises revealed significant information gaps and highlighted that accurate and timely data are essential to detect vulnerabilities. This provided the overarching impetus for various initiatives that took shape over the past decade. Re-assessing a country's national statistical framework and the activities of different statistical agencies is a cumbersome and time-consuming task. The acknowledgement since the global financial crisis has however forced countries and international organisations to do just this. This has, amongst others, implied two significant realities – firstly that we develop programs to measure identified data gaps and secondly that we re-assess our currently measured data and be honest about the quality and required amendments to improve on less than sufficient quality. This paper focuses on the similarities as well as divergence between two related but different frameworks – that of accounting and that of statistics. It is evident that statistics is related to and reliant upon accounting data – the main emerging thought is that the reliance and interrelationship is not always that clear. This vagueness causes statistical output to diverge from the national and international guidelines. The process of deciding on a proper data sourcing framework and aligning all stakeholders to the framework is no easy task. The author recommends that a micro data sourcing approach would be first prize. However, in the absence thereof the survey approach should, as basic requirement, provide a taxonomical mapping framework linking the accounting and statistical worlds. This should be done through training of both statisticians, as well as respondents. It should also include skills development and guidance to stakeholders. This framework should ideally be coordinated at an international level with national economies contributing to and utilising such a framework.

References

1. European Central Bank. (2014). MFI balance sheet and interest rate statistics, securities holdings statistics and implementing technical standards on supervisory reporting. *Bridging the reporting requirements – Third Edition,* (May).
2. European Central Bank. (2016). Bridging tables between the accounting balance sheet items of the NCBs and the ECB and the items to be reported for statistical purposes, (June).

62nd ISI World Statistics Congress
ISI 2019
Kuala Lumpur
18-23 August 2019

Heading for harmonization of data collection

Arjan Bos, Ruben van der Helm
De Nederlandsche Bank, Amsterdam, The Netherlands

Abstract

Comparing and linking statistics is usually difficult if not problematic. *Loans* in statistics A usually do not equal *loans* or the sum of *loans* in statistics B and C respectively, even when the data are collected from the same reporting agents (e.g. a bank). The difficulty originates mainly from different definitions in guidelines for different statistics, diverting interpretation of the guidelines and, hence a different transformation of the data. This issue usually gets more problematic when data are supplied via various 'nodes' (e.g. institutions), each node having its own interpretation of the guidelines.

Aggregated statistics, for example balance sheet information, allow for some degrees of freedom to hide these issues. In contrast, granular data pose new requirements to the level of harmonization. The more atomic the level of granularity, the less degrees of freedom remain to hide interpretation and transformation issues. A formal language helps to minimize these interpretation and transformation issues across the entire supply chain for data.

This paper demonstrates how the 'separation of concerns' into semantical, logical and technical concerns enables the Netherlands Bank (DNB) to harmonize its frameworks, more in particular those used to collect granular data.

Keywords

Separation of concerns; granular data; data model; AnaCredit

1. Introduction

Whether statistics (or data) are used for increasing the company's revenue or profit, organise the daily operations in a day care, economic research or supervising (financial) companies, data are vitally important for organisations, whatever kind of business or industry you work in. This also applies to national central banks (NCBs), like DNB, where data are becoming –if not already became- the new gold. Statistics lead to insights and, hence, decision making (e.g. in the context of monetary policy and supervision) is becoming increasingly data-driven and, hence, users need more (granular) statistics.

Not only the importance of statistics is changing, also its granularity has increased substantially over the past years. Although most reports and

statistics at DNB are still collected and disseminated on an 'aggregated' level, like balance sheet information and profit and loss statements, over the last decade several granular data collections were introduced. Examples are loan tapes (for supervisory purposes), the Single Customer View (in the context of the deposit guarantee scheme), securities holdings statistics (SHS), AnaCredit, Residential real estate (RRE) and commercial real estate (CRE) (all for statistical purposes). Beside more detailed analysis, granular statistics provides the opportunity to better link and integrate different datasets. Linking for example statistics on residential real estate with income information from another source potentially provides very informative insights into the impact of an increase of the interest rates on the (un)affordability of mortgages in specific geographical locations or for specific groups in society. Granular statistics can also be a useful source to compile aggregated statistics.

Moreover, provided 'countless decisions are made based on the analysis of data [..] and that without reliable data policymakers would be "flying blind"' (Cœuré, 2017), data quality is key to policymakers. Working under the assumption Garbage in – garbage out, means that collecting high quality data is important. One can validate data quality upon (or after) reception of the data, however, accurate and detailed guidelines enable all participants in the reporting chain (e.g. reporting agent) to collect from their internal systems and submit high quality data.

The increased need for (granular) high quality data poses new requirements on reporting agents and NCBs. The main issues that arose were i) how to deal with the increase in data volume of several magnitudes, ii) how to ensure that the reported data are inherently correct and iii) how to ensure the auditability of the process of receiving and reviewing reports. Structuring granular data helps addressing these issues and to that end different kind of models are needed – semantic, logical, and technical data models – that describe the structure of the data concerned.

Note, (semantical) harmonisation is not unique in the context of data collection. NCBs and other (inter)national competent authorities that have integrated reporting frameworks (e.g. the Austrian Central Bank (OeNB) and the Banca d'Italia), and more recently the initiatives undertaken by the European Central Bank (ECB) (e.g. Banks' Integrated Reporting Framework (BIRD) and the ESCB Integrated Reporting Framework (IReF) [1]) aim at harmonisation across datasets. The approach taken, by separating concerns, is however less common.

This paper demonstrates how the 'separation of concerns' into semantical, logical and technical concerns enables the Netherlands Bank (DNB) to

[1]https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/index.en.html

harmonize its frameworks. We use examples from granular statistics to show how the various concerns are addressed. Lastly we discuss the advantages and disadvantages as well as the way forward.

2. Methodology

Tekinerdogan et al. (2007) and Sant'Anna et al. (2007) show separating concerns is very beneficial for a solid software architecture. The same principles of separating concerns are also beneficial to data architecture (Evers, 2018; Damhof, 2018), where the concerns are separated into the levels of representation of the data. Evers (2018) describes eight separate levels – narrative, reference, cognitive, formal linguistic, logical, implementation, technology abstraction and database system. For the purpose of this discussion, these levels are summarized into the semantical, logical and technical level. The semantic level encompasses narrative, reference, cognitive and formal linguistic. The logical level contains the formal linguistic and the logical model level. The technical level entails the technology abstraction and the database system layer.

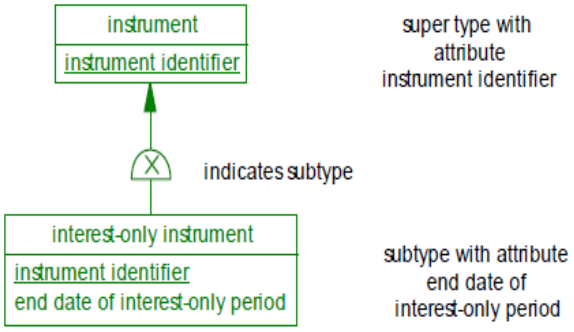
The semantic concerns deal with the narrative, reference and cognitive areas of the data. The aim of the narrative and reference area is to address concerns regarding the understanding of the data, by capturing the vocabulary of the data in a glossary of terms that are common to the business (e.g. in the form of definitions that are stored in a dictionary). These business terms are specific to the department of the organization and not necessarily to the data or other departments. The 'local language' or jargon of each department is registered in local vocabularies, which are ultimately structured into taxonomies and an otology in the cognitive area. In this area the lexical relations – synonyms, homonyms, retronyms, et cetera – are added. More importantly for structuring the vocabulary, two specific relationships are used –'belongs to' and 'is type of' – to create the ontology of business terms. These two relationship are derived directly from the definition of those terms. Together they create the formal structure of the business terms in an ontology.

For example, the AnaCredit regulation (ECB, 2016) contains a glossary of terms, one of these is a business term called 'end date of interest only period'. This term is defined as "The date on which the interest-only period ends. Interest-only is an instrument for which, for a contractually set period, only the interest on the principal balance is paid, with the principal balance remaining unchanged. To be filled in if, at the reporting reference date, the instrument is interest-only". This definition indicates two relationships: 1) interest-only instrument is a type of instrument, and 2) end-date of interest-only period belongs to interest-only instrument. Going through all definitions and guidance documentation of AnaCredit creates a structure between all business

terms, this serves as the basis of the ontology used at DNB (DNB AnaCredit Business Terms v...).[2]

The formalized language structure of the ontology forms the basis for the logical data model (LDM). This phase aims at structuring the data in such a way that the soundness, completeness, consistency and integrity (i.e. the logical concerns) of the reported data are as high as possible. The LDM consists of entity types, attributes and relationship types, which one can derive from all business terms that describe the report and the analysis of their 'belongs to' relationships. Following up on the previous example, let there be four business terms in the ontology to describe the respective report: 'Instrument', 'Instrument identifier', 'End date of interest-only period' and 'interest-only instrument', having the following relationships: 'instrument identifier belongs to instrument', 'end date of interest-only instrument belongs to interest-only instrument' and 'interest-only instrument is type of instrument'. This will result in the logical data model structure showed in figure 1.

Figure 1: example of an LDM



The top of the box shows the entity type name (i.e. '*instrument*' and '*interest only instrument*'). A business term becomes an entity type when it has other business terms that belong to it. When that is not the case it becomes an attribute, in this example that applies to '*instrument identifier*' and '*end date of interest-only period*'. The dependency between entity types determines whether the business term becomes a super-type or subtype. This creates the basis for the LDM and, by using this formal approach, it will be a well-formed and formally correct LDM.

Data modelling techniques are needed to build out the basis into a mathematically correct logical data model. For example data types, domains, primary keys and cardinality of attributes and relationship types have to be determined and described. Only after completing this phase the logical

[2] https://www.dnb.nl/en/statistics/digital-reporting-portal/statistical-reporting/banks/anacredit/index.jsp

structure is ready for transformation into technical "IT solutions". There are many ways to technically implement an LDM (Batini et al., 1992). These range from a semi-structured Excel sheets (via big data solutions with *Schema on read* and document databases) to a highly normalized anchor model (in a relational database management system) (Rönnbäck et al. 2010). The best option depends on the situation, because each option addresses the technical concerns (e.g. performance, storage, persistence, resilience and computation) differently.

In principle each entity type is implemented in a table, or a file, and its attributes become columns or fields. For the reception and submission of a granular data report, the size of the report is a primary concern. The size impacts on the performance, storage and computation. Think about the required speed of the internet connection and the computing power to transfer the data and process it upon reception. One way to limit the size is to include more than one entity type into a table or file, which is called 'flattening of the model'. This model transformation – from multiple files to fewer files – must result in a semantically equivalent model. Also the resilience against change of the data during the transfer is of high importance, to ensure data integrity and avoid assessing and checking incomplete datasets. The resilience is assured by encryption and *hash total calculations*. The latter should be submitted separately to DNB, to check the csv files used for the submission of granular data.

Coming back to our example on interest-only instruments. Reporting agents in the Netherlands should submit 51 files, including 131 entity types. One of the files is called 'interest_only_instrument.csv' and it contains the following columns: "reporting_agent_identifier", "obsrvd_agnt_cd", "reporting_reference_date", "cntrct_id", "instrmnt_id" and "dt_end_intrst_only. The first five columns are the primary key of the file and the last column is the value to be captured. Its heading "dt_end_intrst_only" corresponds directly to the attribute "end date of interest only period" in the LDM. Note, only when there is an instrument that is classified as an interest-only instrument, there can be a value in the column "dt_end_intrst_only" and all records in this file must contain a valid date value. The advantage of this approach is that the reporting agent can only deliver data when applicable, which ensures high data quality.

3. Results

At the moment, DNB implemented LDMs for the collection of data in the context of the deposit guarantee scheme (Single Customer View), AnaCredit and RRE, while the model for CRE is still under development. For the (granular) statistical frameworks one single glossary exists, which means that if a term is used in more than one report, that term has the same name and definition

across the respective reports (e.g. all granular statistical reports include the term '*instrument*'). Table 1 provides some characteristics about the granular reports included in the single glossary. In total, there are 234 distinct attributes in these granular reports of which 135 (57.7%) are shared across them. The share of reused entity types (105 out of 143) is even higher (73.3%). There are only three reporting agents that have to provide all granular reports, while 16 need to submit the reports for AnaCredit and RRE. One reporting agents submits RRE only. The files that the reporting agents have to deliver are not harmonized on the technical level, meaning that the same structure with the same meaning might have additional fields. Hence there are 155 unique files shown in the table.

Table 1: Some characteristics for the granular statistical reports

	AnaCredit	RRE	CRE	Total unique
Attributes	157	131	146	234
Entity types	131	110	114	143
Files	51	53	51	155
Reporting Agents	35	17	3	36
Observed Agents	75	17	3	n.a.

Sharing terms across statistical frameworks yields significant benefits for reporting agents and DNB. A reporting agent can reuse the terms for various reports and hence needs to link it to its internal data storage once. The detailed description of its definition and relations are easily communicated along the data supply chain. Moreover provided the i) description and relationships have been well thought of and ii) the intensive involvement of the business, the business terms and their modelling remain more stable over time. For DNB higher data quality due to better implementation on the side of reporting agents, the possibility to link different datasets and economies of scale in the development of LDMs are the main benefits of sharing terms. The creation of three LDMs on basis of the same harmonized semantics reduced the development time significantly. To give a flavour, the development of the LDM for AnaCredit took about twice the amount of resources compared to the development of both RRE and CRE together.

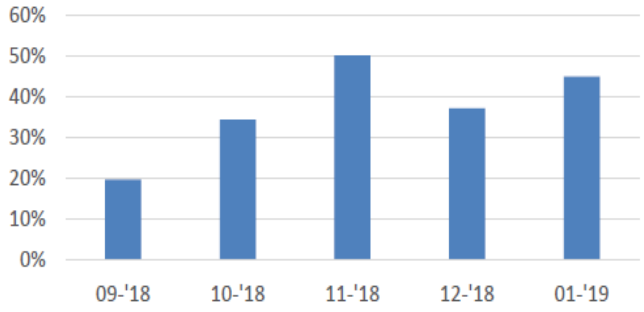
Overall, the approach to use LDM also leads to a significant reduction in imperative – programmed – validations. Imperative validation rules build on requirements documents for each of the validations and use a programming language to implement all the validations. These validation rules are written down in programming language commands, hence the name imperative rules. In contrast, each model element in an LDM restricts the values of the tuples. These restrictions are declared in the model through its underlying predicate logic and set theory (De Haan et al., 2007) and make the validations the LDM

imposes on the data i) mathematically sound and ii) ripe for automated deployment. These are called declarative rules. The former approach is more prone to errors and less rigorous because of the degrees of freedom that programming allows in the implementation. For AnaCredit DNB uses 131 entity types and 150 imperative rules, while the ECB uses 6 entity types and roughly 1640 imperative rules. This shift from imperative to declarative rules leads to fewer errors and less testing efforts on the side of DNB.

Also reporting agents repeatedly stated their appreciation for the exact specifications of the report. The 600+ pages used for the regulation, reporting manuals and validation specification is made visible into one large picture. In contrast, during the development and implementation phase of AnaCredit and RRE, reporting agents faced a lot of issues in fulfilling their reporting obligations. The declarative rules enforce data quality, on a very granular level, 'at the gate' and, hence, requires reporting agents to improve data quality internally. In general for aggregated reports, it is possible to fulfil reporting obligation even if information on a granular level is incomplete or inaccurate. Data quality issues are, as such, hidden in the figures and difficult to detect.

The first data for AnaCredit was collected from (reference period) September 2018 onward. Figure 2 provides an overview of the accepted AnaCredit reports. For the first reporting period only 20% of all reports submitted by reporting agents to DNB were accepted. On average a reporting agent needed to send its report 5 times before it was accepted. The percentage of accepted reports grew gradually to 50% for reference period November 2018, however, after changing some validation rules from signalling to blocking, the number of declined reports grew again. The substantial drop in December can also be partially explained by the fact that some reporting agents submit their reports at quarterly frequency only and hence the December report was their second delivery. Note, their acceptance ratio was higher than for the September report. Looking at the overall data quality, all reports disseminated to the ECB are currently accepted.

Figure 2: Accepted AnaCredit reports (percentage)



4. Discussion and Conclusion

Data-driven decision taking, enhanced techniques/tooling to analyse data and improved hardware that enable the processing of higher data volumes increase the demand for (granular) high quality

statistics. That poses new requirements on reporting agents and competent authorities. Structuring granular data helps addressing issues like how to deal with the significant increase in data volume and correctness of the data submitted. It also helps alleviating the reporting burden on banks and competent authorities. Several initiatives in this field have already been undertaken by national central banks and the European Central Bank. DNB contributes to international initiatives like BIRD and IReF, and uses the 'separation of concerns' into semantical, logical and technical concerns to harmonize its domestic (statistical) reporting frameworks.

The single glossary for various granular reports as well as the LDMs currently implemented yield substantial benefits to reporting agents and DNB. It i) allows for the (partial) reuse of data supply chains by reporting agents, ii) provides clarity regarding the reporting requirements, iii) eases the linking of different statistics, iv) reduces the need for imperative validation rules and v) overall results in higher data quality. Lastly, the direct (and active) involvement of 'the business' in the (semantic) development phase should result in a data collection framework that adds more value to its users.

It should also be noted that such modelling of data requires substantial (specific) resources. First, the active involvement of the business in the modelling phase required significant efforts by experts. Second, DNB hires data modellers, which are relatively scarce (compared to 'regular' programmers) and invested in a tool[3] where the data model itself drives the validation and storage of the data. In contrast, the number of programmers needed for these tasks was greatly reduced.

The current stage of harmonization of data collection could (and should) be further enhanced to improve the usage of the data and derived statistics as well as to alleviate the reporting burden for reporting agents. To that end the harmonization is required for i) the reporting period and frequency, ii) scope and iii) existing thresholds of the reports.

References

1. Batini. C., Ceri. S..& Navathe. S.B. (1992). Conceptual database design: an entity-relationship approach. Redwood City. CA. United States. The Benjamin/Cummings Publishing Company, Inc.
2. Cœuré. B. (2017, March 28). "Setting standards for granular data". Opening at the Third OFR-ECB-Bank of England workshop on "Setting

[3] i-refactory from i-refact. A co-creation between DNB and i-refact. https://www.i-refact.com/

Global Standards for Granular Data: Sharing the Challenge". Frankfurt am Main.

3. Damhof. R. (2018, November 18). "Datascience; more bang for the buck". Retrieved from Data Management & Decision Support: https://prudenza.typepad.com/dwh/2018/11/more-bang-for-the-buck.html
4. ECB. (2016, May 28). Regulation (EU) 2016/867 of the ECB of 18 May 2016 on the collection of granular credit and credit risk data. Retrieved from European Central Bank: https://www.ecb.europa.eu/ecb/legal/date/2016/html/index.en.html?sky=ECB/2016/13
5. Evers. M. (2018). "Data Modeling must DIE! (or the rise of the Data facilitator)". TDWI Europe. Munich: TDWI. Retrieved June 25, 2018, from https://www.tdwi-konferenz.de/tdwi2018/programm/konferenzprogramm/sessiondetails/action/detail/session/mo-61-1/title/data-modeling-must-die-or-the-rise-of-the-data-facilitator.html
6. Sant'Anna, C., Figueiredo, E., Garcia, A., & Lucena, C. (2007). On the Modularity of Software Architectures: A Concern-Driven Measurement Framework. ECSA 2007. Lecture Notes in Computer Science. 4758. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-75132-8_17
7. Tekinerdogan, B., Aksit, M., & Henniger, F. (2007). Impact of Evolution of Concerns in the Model-Driven Architecture Design Approach. Electronic Notes in Theoretical Computer Science, 4564.
8. Rönnbäck, L., Regart, O., Bergholz, M., Johannesson, P., Wohed, P. (2010) Anchor Modeling - Agile Information Modeling in Evolving Data Environments. Preprint submitted to Data and Knowledge Engineering 2010, October 5, 2010
9. De Haan, L., & Koppelaar, T. (2007). Applied Mathematics for Database Professionals. Apress.

Which youths married later than their desired time; Classification tree approach

Arezoo Bagheri[1*]; Mahsa Saadati[2]
[1]Corresponding author: National Population Studies & Comprehensive Management Institute, Tehran, Iran. Email:abagheri_000@yahoo.com, arezoo.bagheri@psri.ac.ir.
[2]National Population Studies & Comprehensive Management Institute, Tehran, Iran. Email: mahsa.saadati@gmail.com, mahsa.saadati@psri.ac.ir.

Abstract

Iranian social norms especially family patterns have changed by increasing modernity in recent years. As a result, the average youths' marriage age has also enlarged which caused conducting more research to study its influential factors during recent years. Thus, this paper aims to measure the gap between the attitudes and behaviours of youths' marriage age and its determinants in Iran by Classification and Regression Trees (CART) algorithm. This method is one of the most applicable classification trees that are applied to model marriage age gap of 12741 females and males selected by multi-stage cluster sampling method from 31 provinces in Iran separately. According to the results, the most influential variables on females' and males' marriage age gap were educational level and the number of siblings, respectively. Females with "university education", "diploma and less education with 5 and more siblings", and "diploma and less education with 3 or 4 siblings and employed" married later than their desired time. Males with "3 and more siblings", "2 and less siblings with 3 and more ideal number of children and employed", "2 and less siblings and 1 or 2 ideal number of children with university education and employed", and "2 and less siblings and 1 or 2 ideal number of children with diploma and less education and negative opinion towards childbearing who are employed" also married later than their desired time. As a conclusion, if the inevitable experience of modernity at the macro, middle, and micro levels doesn't combined with the convenient policy and planning, and, the economic and socio-cultural conditions of the community don't change, negative consequences of such developments would be more than its positive achievements on different social issues specially and more importantly youths' marriage age.

Keywords

Marriage Age Gap; youth; decision Trees; CART algorithm; Iran.

1. Introduction

Since the beginning of the 20th century, the traditional family pattern in Iran, by accelerating the socio-economic changes, has gradually shifted and

these changes have led to changes within the family in the country. One of the areas of this change is youths' marriage age. During 55 years, the average women's age of marriage in Iran increased 5 years from 18.4 to 23.4 years. While the average male's age of marriage in Iran increased gradually 1.7 years from 25 years in the census of 1966 to 26.7 years in the census of 2011. The survey on the marriage age of men and women in Iranian provinces also shows that during 2004-2014, the average men's age of marriage from 26.2 years reached to 28.2 years and the average women's marriage age from 21.8 reached to 23.7 years (Eltejaee and Azizzadeh, 2016).

This trend, after centuries of early marriage experience in Iran, was an important phenomenon in the area of social changes. The increase in the average age of marriage and the development of definite celibacy would disrupt the normal functioning of the family and, as a result, cause disruptions in the community. Moreover, the age of marriage is one of the important indices to assess the physical and mental health of individuals in the society, thus raising its average can affect the health of the community (Murayama, 2000). Some of the consequences of this phenomenon can be increasing anxiety and nervous pressures; the prevalence of depression and behavioral disorders; changing patterns and the norms of marriage and the increasing moral corruptions (Ayatollahi, 2013).

Various studies about the tendency and intendancy of youths to marriage pointed to cultural variables (religious orientation, rate of using media, gender equality, childbearing style, and having bigger siblings), social variables (educational level, gender, self-esteem, pleasure of being single, high expectations, tightening of parents) and economic variables (job status, housing status, income, parents' occupation) (Zarabi and Mostafavi, 2012; Hosseini and Gravnd, 2014; Sadr Al Ashrafi, et al. , 2013; Razadost and Mommunani, 2009).

There are a few studies conducted on ideal marriage age like Asgari Nadushan et al. (2016). They resulted that educational level of respondents and their parents, attitude towards gender equality, individualism and cultural capital has direct meaningful conversely effect, and the variables of adherence to religious values has a reverse and significant relationship with ideal marriage age. Mehrabani (2014) presented an economic model with evidence on decision making for the age of marriage in Tehran. He concluded that the ideal age of marriage for men is higher than women. Some of researches also studied marriage age by gender variable. Aghai and Benchenari (2013). They resulted that there aren't any significant relationship for both genders between household income and attitude towards the marriage age. There are a little studies about the gap in youths' marriage age attitude and their behaviours except Hossaini and Gravand (2013)'s study. Their results indicated that there is a gap in women's behaviour and attitude about marriage age for

women in Kuhdasht city, Iran. The gap is mainly in the negative direction. Preferred average marriage age of more than 80 percentages of women was more than their age at the time of marriage. Based on this study, women's socio-economic status, attitude toward marriage, women's independence and age of respondents had the greatest impact on the gap between the attitudes and behaviour of women's marriage age.

To study youths' marriage age gap, demographic, fertility attitudes and socio-economic characteristics of 12741 pre-married youths from 31 provinces in Iran were collected in 2014. To do so, the following section devoted to introducing briefly CART algorithm and the data set. Section (3) presents the results and conclusion remarks are stated in Section (4).

2. Methodology

Different statistical methods such as logistic regression and multiple regression models were applied for analysing influential factors on marriage age (Hossaini and Geravand, 2014; Habibipour Gatabi and Ghafari, 2011). These traditional methods may counter some problems such as occurring complicated interactions and difficulty of their studying, and handling missing data. To solve the deficiency of these methods, the application of data mining which is a computational process of discovering patterns in large data sets could be applied (Saadati et al., 2017; Bagheri and Saadati, 2016a; Saadati and Bagheri, 2016c; Saadati and Bagheri, 2015; Bagheri and Saadati, 2015; Bagheri et al., 2014).

Amongst data mining methods, Classification and Regression Trees (CART) (Beraiman et al., 1984) algorithm is one of the most applicable classification trees which extract binary splits. CART is a non-parametric statistical methodology developed for analysing classification issues. If the dependent variable is categorical, CART algorithm produces a classification tree (Beraiman et al., 1984). CART methodology is done in three phase of construction or building of maximum tree, selection of right tree size, and classification of new data (Timofeev, 2004).

CART algorithm approaches are applied to "Childbearing Attitudes and Its Social, Economic and Cultural Factors" survey data (Kazemipour, 2014) in this study. Dependent variable rates the gap between women's marriage age behaviour and their appropriate age for marriage of women both genders. The respondent ages which show their marriage age was considered as their behaviour about marriage age. The respondent's attitude about convenient marriage age for youths was also questioned. Finally, the difference in the age of youths in their first marriage and their ideal marriage age was calculated as Marriage Age Gap (MAG).The MAG index consists of two components: the sign and its absolute value. Its sign shows the direction of difference. Its absolute value represents the amount of gap between the youths' attitude and

appropriate age for marriage. This value can be zero, positive, and negative values. The zero value (considered as on-time) indicates that marriage age of youths is their ideal marriage age. Positive values (considered as sooner) indicate that the youths married sooner than their desired marriage age. On the contrary, the negative value (considered as later) indicates that they married later than their ideal marriage age.

Independent variables are also as follows. Opinion towards childbearing is measured according to cultural, social, and economic questions. According to the score of individuals, this variable is categorized to three groups of positive, neutral and negative attitudes. Place of Residence is a place women were living in the study time that could be even urban or rural areas. Job status is respondent's occupation status with two categories of employed and unemployed. Number of siblings has four categories as no-siblings, 1-2 siblings, 3-4 siblings, and 5 and more siblings. Ideal Number of Children (INC) is the ideal number of children respondents desire to have which has three categories of 1-2 children, 3-4 children and 5 and more children (Saadati and Bagheri (2016 a,b); Bagheri and Saadati (2016b); Bagheri and Saadati, 2016). In order to more precisely data analyses of these provinces; they were divided into two provincial classes based on their TFR. According to the values of TFR and replacement level, the province of Iran divided to two categories of TFR≤2, and TFR>2 in this study. Educational level was considered as a categorical variable with three categories of illiterate & primary& middle school, high school and diploma, and university.

3. Results

12741 youths, 49.92 percentages of females and 50.08 percentages of males admitted to public health and treatment centres to receive before marriage consultant in 31 provinces in Iran in 2014 were selected. Ideal average marriage age of males (25.24±3.46) and females (21.79±3.34) were less than the average marriage age of males (26.86±5.05) and females (23±5.57) in this study. The youth's marriage age of almost 80 percent of the sample population is different from their age at the time of marriage (sooner or later). The gaps are mainly in the positive direction which means that youths married later than their desired time. 47.9 percentages of females and 52.8 percentages of males thought that they married later than their desired time. 33.9 percent females and 27.8 percent males thought that they married sooner. Almost 80 percentages of females and males had neutral opinion toward childbearing. The percentages of negative opinion about childbearing for males (14.8 percent) were more than females (12 percent). The most of females (87.4 percent) and males (88.6 percent) lived in urban areas. 24 percentages of females against 86.1 percentages of males were employed. Almost all of youths had at least one sibling (97.7 percentages of females

against 98.3 percentages of males). Only 1.4 percentages of females and 4.1 percentages of males had 5 and more children. More than 40 percentages of females and males were university educated youths. Almost 80 percentages of females and males lived in provinces with TFR less than replacement level. Except opinion toward childbearing and province categories, the other predicted variables had significant effects on MAG for females (p-value<0.01). All the selected variables except place of residence and province category variables were significant on MAG of males (p-value<0.05). Most of youths with different opinions toward childbearing, job status, INC, and province categories married later than their desired time. Most of urban and rural (49 percent and 45.5 percent) females married later and sooner than their desired age, respectively. While, more than 50 percentages of males whether they are living in urban or rural areas married later than their desired time. Most of females with less than 2 comparing to more than 2 siblings married sooner than their desired time. Similar to this group were most of males (39.1 percent) who didn't have siblings. Most of females with secondary school and less and university educational levels comparing to high school and diploma educated ones married later than their desired time. Most of males in different educational levels were similar to this group.

CART algorithm was fitted to classify MAG of youths in Figures (1) and (2) by gender. Classification accuracy of these fitted tress were 0.62 and 0.60 which means that MAG of 62 and 60 percentages of females and males have been correctly classified, respectively (This value indicates that misclassifications of these models are equal to 38 and 40 percentages).

According to the Figure (1), educational level variable is located in the root of CART algorithm of females' MAG. The other influential variables in this figure are number of siblings and job status. In Figure (2), number of sibling variable is located in the root of CART algorithm of males' MAG. Job status, educational levels , INC, and opinion also influences this variable too.

Rules (1) to (3) can be extracted from the classification tree of MAG for youth females in Figure (1): 1-University educated females married later than their desired timer. 2-Diploma and less educated females with 2 and less siblings opposite to 5 and more siblings married sooner than their desired time. 3-Diploma and less educated females with 3 or 4 siblings who are employed opposite to unemployed married later than their desired time.

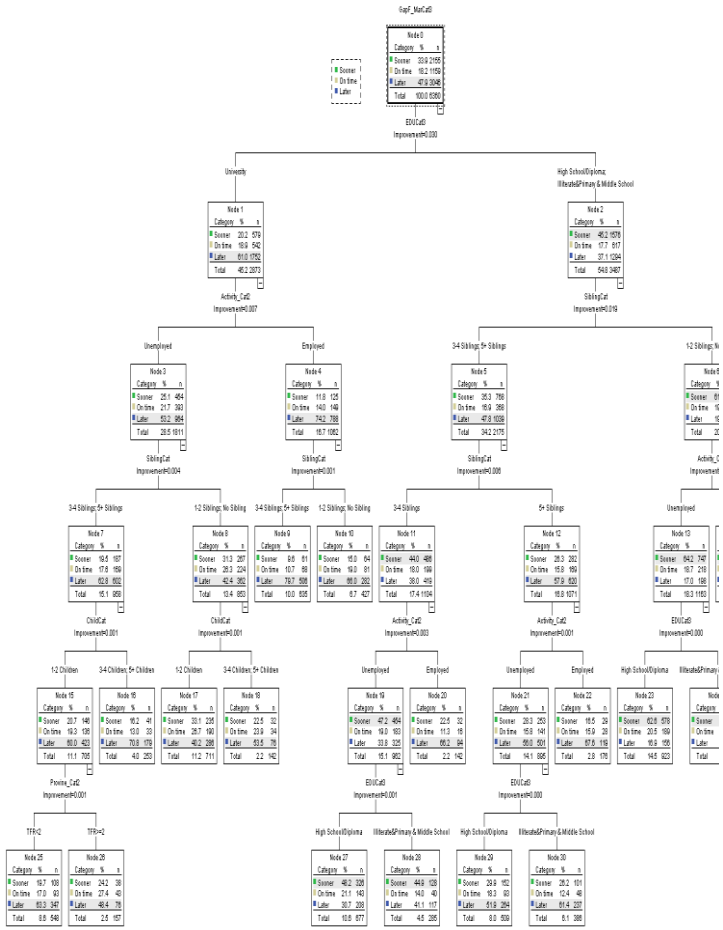


Figure 1. Classification Tree of Female's MAG

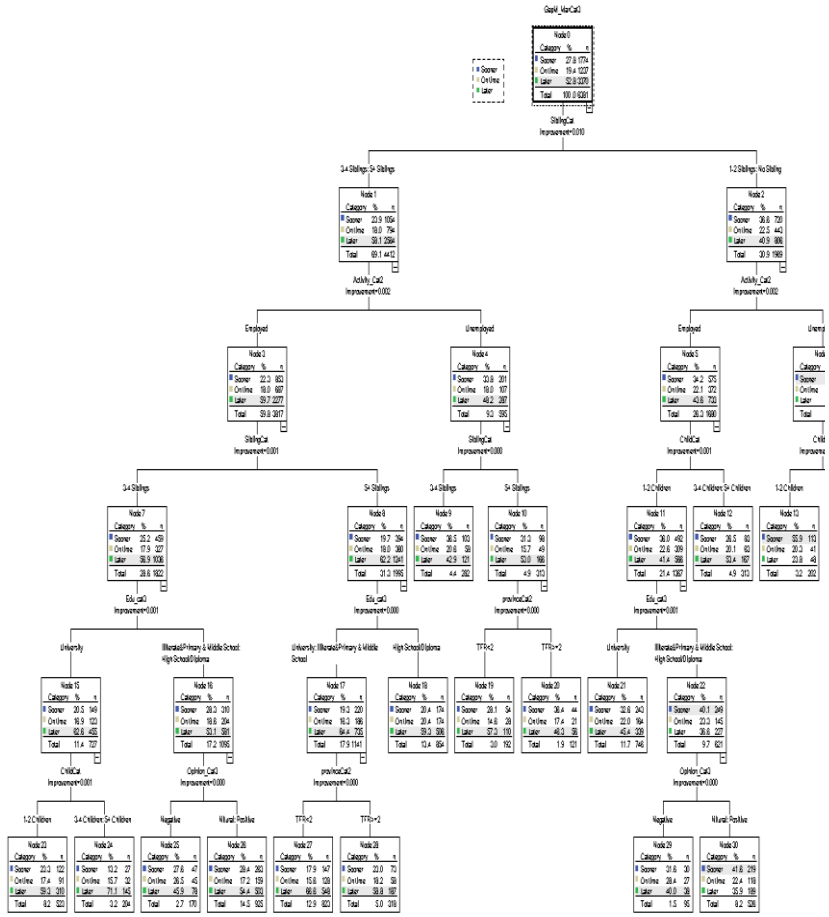


Figure 2. Classification Tree of Male's MA

The followings are the extracted rules from Figure (2): 1-Males with 3 and more siblings married later than their desired time. 2-Unemployed males with 2 and less siblings married sooner than their desired time. 3-Employed males with 2 and less siblings and 3 and more INC married later than their desired time. 4-Employed males with 2 and less siblings and 1 or 2 INC who are university educated married later than their desired time. 5-Employedmales with 2 and less siblings and 1 or 2 INC who are diploma and less educated with negative and positive opinions married sooner than their desired time. 6-Employed males with 2 and less siblings, and 1 or 2 INC who are diploma and less educated with negative opinion married later than their desired time.

4. Discussion and Conclusion

The purpose of this paper was to examine different socio-economic factors on youths' MAG. Most of previous studies were related to the study of the influential factors on behaviours or attitudes of MAG, separately. Moreover, most of them focused on females' MAG. However, this study attempted to investigate the effect of socio-economic variables on the youths' MAG in Iran by gender. According to the results of this study most youths married later than their desired time. Females' educational level was the most important variable in the fitted CART algorithm of females' MAG. However, this variable was also an influential variable in Males' MAG. Youths with higher educational levels married later than their desired marriage age. Some of the other authors are confirmed this results such as Rezaedost and Mommunani (2009); Hossaini and Geravand (2014); Sadrolashrafi et al. (2012); Habibpour Gatabi and Ghafari (2011) and Zarrabi and Mostafavi (2011). They stated that increasing educational levels of youths specially females could cause delay in marriage age and as a result increasing MAG.

Number of siblings was the most important variable on males' MAG though this variable also was one of the influential variables on females' MAG. Youths with more siblings married later than their desired time. This result is in the same line of Tsuya and Kurosu (2000), and Jin et al. (2003). These authors reached this conclusion that people with more siblings or who are themselves among the older siblings, are likely to marry earlier.

Another effective variable on youths' MAG is job status. According to the results of this study, employed youths married later than their desired time. Eltejaee and Azizzadeh (2016); Aghai, and Taheri Benchenari (2012), and Hossaini and Geravand (2014) confirmed the results of this study. Opinion towards childbearing was also influential on males' marriage age. Those with negative opposite to positive and neutral attitudes towards childbearing thought that they married later than their desired time. This variable has been studied in Ardebili (1997) study through measuring males' opinion towards marriage and selecting a wife. Most probably this result indicates that youths

are thinking marriage and childbearing as two separated phenomena. Marrying is not summarized to childbearing for them (Ojaghlo and Sarai, 2014).

Based on the results of this study, in order to solve a social problem called marriage age gap, socio-economic conditions of youth should be improved.

Acknowledgment

This article is extracted from a survey under the title of "Mining Demographic Data by Decision Tree" which is supported by National Population Studies and Comprehensive Management Institute in 2014 by the registered number of 20/15283.

References

1. Eltejaee, E. and Azizzadeh, M. (2016). Investigating the Economic and Cultural Factors Affecting the Age of Marriage in Iran: A provincial study. Cultural Community Studies, Human Sciences Research Institute and Cultural Studies, 7, 1-23.
2. Murayama S. (2000). Regional standardization in the age at marriage: A comparative study of pre- Industrial Germany and Japan. His Fam;6(2): 303-324.
3. Ayatollahi Z. (2013).The population and family planning. Qom: Education Office Publication.
4. Zarabi V, Mostafavi F. (2012).Measuring factors affecting marriage in women of the Iranian. View economic. Journal An economic study. 4: 33-64 (Persian).
5. Hosseini H, Gravnd M. (2014).Measuring factors affecting of behavior and attitudes women to marriage age in the city kohdasht. Women develop and politic Journal; 11(1): 101-118 (Persian).
6. Sadr Al Ashrafi M, Shamkhani A, Yousefi Afrasfteh M. (2013).Investigate factors affecting in the easy marriage from the students women view Payame Noor University Razan. Journal of cultural engineering; 69, 70:86-101 (Persian).
7. Rezaedost, K. and Mommunani, I. (2009). Investigating the relationship between marriage age delay and variables such as income level, educational level, and other variables in employed women, Journal of Applied Consulting, 4(16), , No. 1, 103-120.
8. Asgari Nadushan, A., Abassi Shavazi, M.J., Piri Mohamadi, M. (2016). The ideal age of marriage and Its determinants in Yazd. Quarterly Cultural and Social Council of Women and Family. 19(73), 35-63.
9. Mehrabani, V. (2014). Economic analysis of decision making for marriage age, Women Strategic Studies, 17(65), 69-118.

10. Aghai, S.S., Taheri Benchenari, R. (2012). The attitude of young people to the impact of socio-cultural factors affecting the rising age of marriage (Case study: District 2 and 3 of District 4 of Tehran in 1391). Journal of Sociological Studies of Iran. 3(8), 75-94.
11. Habibipour Gatabi, K., Ghafari, GH. (2011).Causes of increasing girls' marriage age. Woman in Development and Politics (Women's Research). 9(1),7-34.
12. Saadati M, Bagheri A, and Razeghi Nasrabad HBB. (2017).Modeling Children Ever Born and Ideal Number of Children by Classification Tree. Journal of research and heath. In press.
13. Bagheri A. Saadati M, (2016).Comparing classification tree algorithms to forecast sex preferences of women in marriage threshold. 13th statistical conference, Kerman university, 24-26 September, 2016.
14. Saadati M, Bagheri A. (2016).Educated Iranian women in favor of having girls: CART classification approach. Europian Population Conference 2016. Mainz, Germany.
15. Saadati, M; Bagheri, A. (2015). Mining Children Ever Born Data; Classification Tree Approach. *Indian Journal of Science and Technology*. 8(30).
16. Bagheri, A; Saadati, M. (2015). CART Model for Classification Children Ever Born. *Jorjani Journal*. Vol(3), No(2), 63-88.
17. Bagheri A, Saadati M, and Razeghi Nasrabad HBB. (2014). Introduction and application of CART model for classifying 15-49 year old women ideal number of children in Semnan province, Journal of Population Association letters, vol.9, no. 17 summer. 2014.
18. Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) 'Classification and Regression Trees. Belmont', California: Wadsworth, Inc.
19. Timofeev, R. (2004) 'Classification and Regression Trees (CART)Theory and Applications', A Master Thesis CASE - Center of Applied Statistics and Economics Humboldt University, Berlin.
20. Kazemipour, Sh. (2014). Childbearing Attitudes and its social, economical and cultural factors, Statistical Research Center, Tehran, Iran.(Full Text in Persian).
21. Saadati M and Bagheri A. (2016a). Study of ideal marriage interval to childbearing in terms of youth at the threshold of marriage. Payesh Journal, 17(2): 239-250.
22. Saadati M and Bagheri A. (2016b). Comparing Childlessness Ideal Survival Time of Women in the Threshold of Marriage by Job Status. Population change, human resources & employment in Iran. 26-27 October 2016. Yazd university.

23. Tsuya, N. and S. Kurosu. (2000). Economic and household factors of first marriage in early modern Japan: Evidence from two northeastern villages, 1716–1870. Paper prepared for presentation at the International Congress of Historical Sciences, Oslo, August 7. www.oslo2000.uio.no/program/papers/.
24. Jin, X., Li, S., & Feldman, M. W. (2005). Marriage form and age at first marriage: a comparative stud y in three counties in contemporary rural China.Social biology, *52*(1-2), 18-46.
25. Ardebili, I. (1997). Study of the attitudes of single boys aged 20-29 in Mashahd province towards marriage and its determinants. Master thesis, Faculty of Literature and Humanities. Ferdowsi university, Mashahd, Iran.
26. Bagheri A and Saadati M. (2016). Analysis of the childlessness ideal survival time of young's at the threshold of marriage: the parametric log normal model, Pajoohande journal, 21(4), 199-209.
27. Ojaghlo and Sarai. (2014). Studying child time value changes in Iran (Women's study of Zanjan city). Social research studies in Iran, 3(2): 261-283.

Modeling birth intervals by Variance-corrected recurrent models

Mahsa Saadati, Arezoo Bagheri
National Population Studies & Comprehensive Management Institute, Tehran, Iran

Abstract

One of the most important determinants of fertility levels that have an important role on fertility rates and population growth changes, as well as maternal and infant health and mortality is birth intervals. Considering the importance of this issue, the main purpose of this article is to analyze second and third women's birth intervals using variance-corrected recurrent models. Stratified random sampling with probability proportional to size was used for selecting 610, 15-49 year-old married women by a structured questionnaire during the winter and spring of 2017 from different regions of Tehran, Iran. Anderson-Gill (AG), and Prentice-William-Patterson Total and Gap Time (PWPTT-PWPGT) models were fitted to investigate the effects of selected variables on second and third birth intervals. Considering the response variable which is a time interval, the suitable model was PWP-GT. Calendar period, marriage age, and migration status had significant effect on both birth intervals (p-value<0.05) and activity status (p-value<0.01) and residence place (p-value<0.050) influenced on second birth interval. The increase in both birth intervals of recent calendar periods was larger comparing to the last calendar periods. Also, by increasing marriage age, both birth intervals decreased. Both birth intervals of migrant women were 1.298 and 1.404 times shorter than non-migrant women, respectively. Employed women and those living in developed, completely-developed and semi-developed areas comparing to unemployed women and those living in the developing area had longer second birth interval, respectively. As a conclusion, it should be noted that in the analysis of event histories, the range of risk and the hazard set are the most important elements of selecting the best fitted model. Moreover, increasing birth intervals, that influence directly childbearing, may result from economic and social conditions and can be prevented by applying convenient policies.

Keywords

Birth intervals; recurrent event model; Variance-corrected model; Anderson-Gill; PWPTT; PWPGT.

1. Introduction

Fertility is an important component of population dynamic which plays a major role in changing the size and structure of a given population (Yohannes

et al., 2011). Fertility analysis is an important issue for policy makers to develop guidance for population control and also to evaluate family planning programs (Kamal, Pervaiz, 2012). The number of children each woman (or couple) bear during her childbearing years in the population, and the ages at which the woman has given birth to her children are the basic factors which determine population growth. While the former relationship is obvious, the latter (that determines timing or birth spacing), means that for the same number of children born per woman, mothers who give birth during their later childbearing years contribute more towards population control than those who give birth to their children early in their life (Rajaretnam, 1990). Birth interval (spacing) is the length of time between two successive alive births (Central Statistical Agency, 2006). Birth interval analysis is more susceptible technique for measuring fertility than other conservative methods of measuring fertility (Nath et al. 2000). Pattern of birth intervals not only provides pace of child bearing but also chances of transition to higher parity (Pillai, 2010). Since birth spacing has the important role on the health of mothers and children, it also merits special attention in public health. Many researches demonstrated that, shorter birth intervals may not provide enough time for mothers to restore nutritional reserves that are needed for adequate fatal nutrition and growth. Fatal growth retardation can result in low birth weight, which adds to the risk of children premature death. Children born too close together compete for resources and maternal care, including breastfeeding (Siegel, 2011). It is argued that when a new born comes, it is likely that the family will invest more of its limited resources in the form of care to the new born and the other children are more likely to suffer or merely receive inadequate share of the resources distributed among siblings (Hailu, Gulte, 2016; USAID, 2005).

Birth spacing has become a main strategy of the health promotion program for mothers and children over the past two decades in Islamic Republic of Iran (Fallahzadeh, 2013). So many researches were conducted to study determinants of birth intervals, recent years; Hajian et al. (2009) showed that there were significant correlation between birth interval with maternal age, duration of breast feeding, sex of previous child, history of alive children, history of infant mortality of the previous child, type of contraception used, regular attendance at a family planning clinics. Other study by Fallahian et al. (1993) found the duration of breastfeeding and the type of contraceptive used were factors significantly associated with child intervals. Rasekh and Momtaz (2007) stated that the encouraging women for higher education and giving opportunity to them to get employed may be the influential way of extending their birth spacing which result in slowing down fertility in Ahvaz, Iran.

In a statistical point of view, birth intervals must be analysed by recurrent models because of repetitive nature of child birth; women may experience its

several times in their lifetime. The time interval between deliveries can be analyzed using models for recurrent events. In these models, the given event (i.e., childbirth) occurs more than once for each individual does. To investigate recurrent events, selection of the appropriate model depends on the research objective, researcher, and the nature of the data.

In the present study, variance-corrected recurrent event model including Andersen-Gill (AG) and Prentice, Williams and Peterson Total and Gap Times (PWP-TT and PWP-GT) models, were used to evaluate different influence factors on the time interval between second and third married women's deliveries, who lived in Tehran in 2017. To do so, introduction of data and methods displays in Section (2), results and discussion are presented in Section (3) and (4), respectively.

2. Methodology

In this study, data of a cross-sectional survey under the title of "effects of socio-economic rationality dimensions on childbearing behaviour in Tehran" was used (Abdolahi, 2017). Considering the design effect of 2.5 and the rate of non-response (1.25), the sample size of 610 eligible 15-49 year old women from Tehran province in Iran were studied using multi-stage sampling and proportional probability method, in 2017. Using the hierarchical clustering approach, the 22 metropolitan regions of Tehran province were clustered in terms of developmental degree in four levels of development as completely-developed, developed, semi-developed, and developing regions (Rafieian, 2012).

Two important features of recurrent event data are that the events are ordered and that the subject can only be at risk for one such event at a time. Based on the aim of this article, effects of some selected covariates on second and third birth intervals were analysed by three variance- corrected models; These models assume that, conditional on the covariates, the event and censoring times are independent (independent censoring assumption), and these are different in assumptions and the data layout for analysis. Another major difference among them is the way the repeated events are modeled. AG, PWP-TT, and PWP-GT were described as follows:

- **AG model**

The counting process model of Andersen-Gill (AG) generalizes the Cox model, which is formulated in terms of increments in the number of events along the time line (Andersen, Gill; 1982). The outcome of interest is time since randomization for a treatment (or other exposure) until an event occurs, i.e. time since study entry, also known as total time scale. It uses a common baseline hazard function for all events and estimates a global parameter for the factors of interest. The AG model assumes that the correlation between event times for a person can be explained by past events, which implies that

the time increments between events are conditionally uncorrelated, given the covariates. It is a suitable model when correlations among events for each individual are induced by measured covariates (Moulton, Dibley; 1997).Thus, dependence is captured by appropriate specification of time-dependent covariates, such as number of previous events or some function thereof. However, if this assumption does not hold, a remedy is to use a robust sandwich covariance matrix for the resulting regression coefficient estimators, (Cox, 1972) which uses a Jacknife estimate to anticipate correlations among the observations and provides robust standard errors. The AG model is usually indicated for analyzing data when all dependence between subsequent events is mediated through time-varying covariates and the interest is in the overall effect on the intensity of the occurrence of a recurrent event.

- **PWP Model**

PWP model analyses ordered multiple events by stratification, based on the prior number of events during the follow-up period (Prentice, Williams, Peterson; 1981). All participants are at risk for the first stratum, but only those with an event in the previous stratum are at risk for the successive one (Pandeya et al., 2005). The model can incorporate both overall and event-specific effects for each covariate. In practice the data may need to be limited to a specific number of recurrent events if the risk set becomes very small for later strata and event-specific estimates become too unreliable (Kelly PJ, Lim; 2000). Besides using the same outcome (total time: TT) as in the AG model, the PWP model can also be usually defined in terms of gap time (GT), which is the time since the previous event. When using a gap or waiting-time scale, the time index is reset to zero after each recurrence of the event, with assumption of a renewal process. Gaps between events are often useful with infrequent events, when a renewal occurs after an event or when the interest lies on prediction of a next event. Hence, two stratified PWP models can be fitted: PWP-TT, which evaluates the effect of a covariate for the kth event since the entry time in the study; and the PWP-GT, which evaluates the effect of a covariate for the kth event since the time from the previous event. Unlike the AG model, the effect of covariates may vary from event to event in the stratified PWP models.

3. Results

Among 610 married women, 21.2%, 34.7%, 31.3%, and 1.28 of them had 0, 1, 2, and 3 children, respectively. Since in this study, only second and third birth intervals were analyzed, women with one or no children were considered as censored data. Most of women with 2, and 3 children had 30-39 (47.1%) and 40-49 (75.3) years old, and married in their 20-24 (40.1%), and 17-19 (32.5%) ages, respectively. Only 1.3% and 5.5% of women with 3 or 2 children had MS/PhD, and primary and less educational levels, respectively. To

compare results of variance-corrected models for second and third birth intervals, they were fitted to data based on selected covariates and they presented in table (1), and (2), respectively. In this study, calendar period was included in the model as a covariate, which explicitly brings into the analyses the birth risks in four different time periods, including: before May 1987, May 1987–April 1997, May 1997–April 2007 and May 2007–April 2017. This covariate measures the period during which a woman has been exposed to pregnancy of a given-order birth. In fact, the covariate of calendar periods of exposure to birth represents a combined period effect of contextual factors such as population and development policies and people's living conditions that influence women's birth risks (Erfani & McQuillan, 2008; Erfani, 2017a)

Although, AIC is one of the most important indices for determining the best model in many statistical modelling, but selection of the best recurrent models depends on risk intervals and sets, and also the main aim of study. In Table (1), by considering AIC, PWP-TT model must be selected as the final model, but since the aim of this study was investigating the effect of selected covariates on second birth interval, PWP-GT must be considered as the final model. On the other hand, AG model must be selected based on estimated coefficient's standard deviations. This model did not account correlation between two birth intervals for each woman and also all women, with and without any children, were entered in the risk set which lead to invalid results. Based on Table (2), PWP-GT also considered as the final model for third birth interval. Results of two tables showed that calendar period had significant effects on both second and third birth intervals; the largest gap from first to second, and second to third children belonged to women in the last calendar period. By increasing marriage age, both second and third birth intervals were decreased. Second and third birth intervals for migrant women were 1.298 and 1.404 times shorter than non-migrant women, respectively. Job status and region of residence also affected on second birth intervals. Employed women (0.758) delayed second children more than unemployed women. Women who lived in developed (0.576), completely-developed (0.705), and semi-developed (0.819) regions had larger intervals between first and second children compare to women who lived in developing regions.

Table 1. Variance corrected recurrent model for Second Birth Interval

Variables		AG			PWP-TT			PWP-GT		
		β	SE	HR	β	SE	HR	β	SE	HR
Calendar-Period	Before May 1987 (reff)	-	-	-	-	-	-	-	-	-
	May 1987-Appril 1997	-0.711*	0.209	0.491	-1.207*	0.213	0.299	-1.032*	0.212	0.356
	May 1997-Appril 2007	-0.799*	0.209	0.450	-1.459*	0.215	0.233	-1.178*	0.210	0.308
	May 2007-Appril 2017	-0.927*	0.219	0.396	-1.998*	0.228	0.136	-1.252*	0.221	0.286
Marriage Age		0.009	0.011	1.009	0.024**	0.011	1.024	0.012**	0.010	1.012
Educational Level	Primary & Less (reff)	-	-	-	-	-	-	-	-	-
	Secondary & high School	0.051	0.194	1.053	0.399**	0.193	1.490	0.075	0.193	1.078
	Diploma	-0.139	0.183	0.870	0.096	0.184	1.100	-0.156	0.179	0.855
	BS/Associate	-0.136	0.211	0.873	0.088	0.211	1.092	-0.137	0.207	0.872
	MS & PhD	-0.265	0.278	0.767	-0.070	0.274	0.932	-0.271	0.274	0.762
Couple's Educational Level	Primary & Less (reff)	-	-	-	-	-	-	-	-	-
	Secondary & high School	0.034	0.183	1.035	0.108	0.178	1.114	0.072	0.182	1.074
	Diploma	-0.083	0.186	0.920	-0.128	0.181	0.880	-0.110	0.183	0.896
	BS/Associate	-0.088	0.199	0.916	-0.184	0.193	0.832	-0.098	0.195	0.907
	MS & PhD	-0.022	0.246	0.978	-0.134	0.240	0.875	-0.047	0.243	0.955
Job Status	Unemployed (reff)	-	-	-	-	-	-	-	-	-
	Employed	-0.212*	0.104	0.809	-0.247**	0.107	0.781*	-0.277	0.104	0.758
Migration	Non-migrant (reff)	-	-	-	-	-	-	-	-	-
	migrant	0.169*	0.128	1.184	0.314**	0.130	1.369**	0.261	0.129	1.298
Family Expenditure (each months)	Less than 2 million Rials (reff)	-	-	-	-	-	-	-	-	-
	2- 3.5 million Rials	0.005	0.095	1.006	0.020**	0.097	1.020	0.013	0.096	1.013
	More than 3.5 million Rials	0.042	0.156	1.043	0.065	0.161	1.068	0.065	0.157	1.067
Regions of Residance	Developing (reff)	-	-	-	-	-	-	-	-	-
	Semi-developed	-0.153*	0.098	0.858	-0.156*	0.098	0.855**	-0.200	0.098	0.819
	Developed	-0.431*	0.145	0.650	-0.584**	0.148	0.558*	-0.551	0.146	0.576
	Completely-developed	-0.253*	0.144	0.777	-0.289	0.143	0.749**	-0.350	0.143	0.705
AIC		**7674.86**			**6749.26**			**7904.779**		

4. Discussion and Conclusion

Given the relative lack of using appropriate methods for analysing recurrent data using survival analysis, in demographic researches, variance-corrected models were applied to study factors effect on second and third birth intervals among 15-49 year old women lived in Tehran, Iran, in 2017. Analysis based only on the first event time cannot be used to examine the effect of the risk factors on the number of recurrences over time (Pandeya, et al., 2005; Dacourt et al., 2004). Many researchers continue to use logistic regression for such analysis, despite known limitations and the increasing availability of analytical approaches that handle recurrent events (Gill et al., 2009; Purroy et al., 2013). In cohort studies, there is little justification for fitting logistic regression once there are other available approaches for estimating risk (Gill et al., 2009). The count data models, such as Poisson and negative binomial, are the simplest ways to analyse repeated events. However, they consider the total number of events per a fixed period of time, ignoring the time between repeated occurrences. In addition, it is not possible to identify whether the effect of exposures changes the rate of occurrence across the time period (Pandeya, et al., 2005). Thus, survival analysis is preferred when follow-up times are variable among participants, or when there are time-varying covariates or time-varying effects (Gill et al., 2009). Several approaches have been proposed to account for intra-subject correlation that rises from multiple events settings in survival analysis. If it is reasonable to assume that the risk of recurrent events remained constant regardless of the number of previous events, then the AG model is recommended (Therneau TM, Grambsch PM., 2000). The AG model assumes that the time increments between events are conditionally uncorrelated given the covariates. However, omission of an important covariate could induce dependence. In such case, the standard errors would be underestimated, causing inflation of type I error. A possible remedy would be to fit an AG model with a time-dependent covariate for the number of events. Advantages of an AG model include the ability to accommodate time-varying covariates and discontinuous intervals of risk (Castan˜ eda, Gerritse. ;2010).

Table 2. Variance corrected recurrent model for Third Birth Interval

Variables		AG			PWP-TT			PWP-GT		
		β	SE	HR	β	SE	HR	β	SE	HR
Calendar-Period	Before May 1987 (reff)	-	-	-	-	-	-	-	-	-
	May 1987-Appril 1997	-0.121	0.231	0.886	-0.878**	0.236	0.415	-0.737*	0.235	0.478
	May 1997-Appril 2007	-0.384*	0.229	0.681	-1.842**	0.241	0.159	-1.659*	0.239	0.190
	May 2007-Appril 2017	-0.324*	0.228	0.723	-2.040**	0.247	0.130	-1.828*	0.246	0.161
Marriage Age		0.005	0.012	1.005	0.051**	0.012	1.052	0.046*	0.012	1.047
Educational Level	Primary & Less (reff)	-	-	-	-	-	-	-	-	-
	Secondary & high School	-0.067	0.194	0.936	0.287	0.194	1.333	0.194	0.195	1.215
	Diploma	-0.040	0.192	0.961	0.019	0.195	1.019	-0.019	0.194	0.981
	BS/Associate	-0.031	0.240	0.969	0.005	0.244	1.005	0.025	0.243	1.025
	MS & PhD	-0.121	0.337	0.886	0.409	0.331	1.505	0.353	0.331	1.424
Couple's Educational Level	Primary & Less (reff)	-	-	-	-	-	-	-	-	-
	Secondary & high School	0.077	0.187	1.080	0.056	0.189	1.058	0.104	0.188	1.109
	Diploma	-0.039	0.202	0.961	-0.162	0.203	0.851	-0.079	0.202	0.924
	BS/Associate	-0.084	0.224	0.919	-0.236	0.226	0.790	-0.173	0.225	0.841
	MS & PhD	-0.040	0.291	0.961	-0.214	0.287	0.808	-0.126	0.288	0.882
Job Status	Unemployed (reff)	-	-	-	-	-	-	-	-	-
	Employed	-0.048	0.128	0.953	-0.120	0.132	0.887	-0.129	0.131	0.879
Migration	Non-migrant (reff)	-	-	-	-	-	-	-	-	-
	migrant	0.064	0.152	1.066	0.341*	0.159	1.406	0.340**	0.157	1.404
Family Expenditure (each months)	Less than 2 million Rials (reff)	-	-	-	-	-	-	-	-	-
	2- 3.5 million Rials	-0.010	0.111	0.990	0.116	0.113	1.122	0.113	0.113	1.119
	More than 3.5 million Rials	0.072	0.208	1.075	0.031	0.210	1.031	0.083	0.210	1.086
Regions of Residance	Developing (reff)	-	-	-	-	-	-	-	-	-
	Semi-developed	-0.112*	0.104	0.894	-0.084	0.106	0.919	-0.124	0.106	0.883
	Developed	-0.233*	0.180	0.792	-0.192	0.185	0.825	-0.264	0.185	0.768
	Completely-developed	-0.270*	0.191	0.764	-0.195	0.188	0.823	-0.309	0.189	0.734
AIC		**5719.96**			**4687.826**			**4776.780**		

If it is reasonable to assume that the occurrence of the first event increases the likelihood of a recurrence, then PWP is recommended. The PWP models (TT or GT) are also indicated when there is interest in estimating effects for each event separately. The PWP models assume that the subjects can only be

at risk for a given event after he/she experienced the previous event. A limitation for the use of PWP models is that the risk sets for the later events get quite small, making the estimates unstable. Therefore, we usually have to truncate the data.

In this article AG, PWP-TT, and PWP-GT models were fitted to data, and PWP-GT model was selected as the final model for both second and third birth intervals because of few number of children (recurrent events) per woman, the kind of risk set, and the type of interested response which is gap time (interval). Based on PWP-GT model, calendar period, marriage age, and migration status had significant effects on both birth intervals (p-value<0.05). Women in recent calendar period had the largest gap between first to second, and second to third births. Erfani (2013, 2015a, 2015b, 2017a, 2017b), Erfani and McQuillan (2008) reported the same results. Migrant women gave birth to second and third children more rapidly than non-migrant women. By increasing marriage age, birth intervals were decreased. Employed women move on to have their second birth at a slower pace than unemployed women; this is the same as Hajian-Tilaki et al (2009), Rutstein (2011), Fallahzadeh et al. (2013), and Erfani & McQuillan (2014) results. Based on the results of this study, women who lived in more developed regions had larger second birth intervals than not developed regions.

In summary, the choice of the approach for analysis of recurrent event data will be determined by many factors, including: number of the events; relationship between subsequent events; effects varying or not across recurrences; biological process; and dependence structure. Usually the stratified models, as PWP (total or gap times) are used when there are few recurrent events per subject and the risk of recurrence varies between recurrences. On the other hand, AG model are indicated for frequent events with constant hazard between recurrences.

References

1. Yohannes, S., Wondafrash, M., Abera, M., & Girma, E. (2011). Duration and determinants of birth interval among women of child bearing age in Southern Ethiopia. BMC pregnancy and childbirth, 11(1), 38.
2. Kamal, A., Pervaiz, M. K. (2012). Determinants of Higher Order Birth Intervals in Pakistan. Journal of Statistics; 19(1):15-24.
3. Rajaretnam, T.(1990). How delaying marriage and spacing births contributes to population control: an explanation with illustrations. J fam welfare; 34: 3-13.
4. Central Statistical Agency [Ethiopia] and ORC Macro (2006): Ethiopia Demographic and Health Survey 2005. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ORC Macro.

5. Nath, D.C., Leonetti, D. L. and Steele, M. (2000). Analysis of birth intervals in a non-contracepting Indian population: An evolutionary ecological approach. Journal of biosocial Science, 32, 343-354.
6. Pillai, V. K. (2010). Child spacing and contraception among the poor in Zambia.
7. Siegel, J. S. (2011). The demography and epidemiology of human health and aging. Springer Science & Business Media.
8. Hailu, D., & Gulte, T. (2016). Determinants of Short Interbirth Interval among Reproductive Age Mothers in Arba Minch District, Ethiopia. International journal of reproductive medicine.
9. United States Agency for International Development (USAID). (2005).Strengthening Family Planning Policies and Programs in Developing Countries, Washington, DC, USA.
10. Fallahzadeh, H., Farajpour, Z., & Emam, Z. (2013). Duration and determinants of birth interval in Yazd, Iran: a population study. Iranian journal of reproductive medicine, 11(5), 379.
11. Hajian-Tilaki KO, Asnafi N, & Aliakbarnia-Omrani F. (2009), The Patterns and determinants of birth intervals in multiparous women in Babol, Northern Iran. Southeast Asian J Troped Public Health; 40: 852-860.
12. Fallahian, M., Kazemnegat, A., & Ebrahimi, N. (1993). Determinant of short birth interval. J Behboud Kermanshah Med Sci Univ Iran, 18, 35-48.
13. Rasekh, A., & Momtaz, M. (2007). The determinants of birth interval in Ahvaz-Iran: a graphical chain modelling approach. J Data Sci; 5: 555-576.
14. Abdolahi A. Effects of socio-economic rationality dimensions on childbearing behavior in Tehran. National Population Studies & Comprehensive Management Institute; 2017.
15. Rafieian MS, M. (2012). The Spatial Analysis of Tehran's Development Level Based on Metropolitan Areas. the Journal of Spatial Planning; 16(4):25-48.
16. Andersen PK, Gill RD. (1982). Cox's regression model for counting processes: a large sample study. Ann Stat;10:1100–20.
17. Moulton LH, Dibley MJ. (1997). Multivariate time-to-event models for studies of recurrent childhood diseases. Int J Epidemiol;26:1334–39.
18. Cox DR. (1972). Regression models and life-tables (with Discussion). J Royal Stat Soc B;34:182–220.
19. Prentice RL, Williams BJ, Peterson AV. (1981). On the regression analysis of multivariate failure time data. Biometrika;68:373–79.
20. Pandeya N, Purdie DM, le Green A, Williams G. (2005). Repeated occurrence of basal cell carcinoma of the skin and multifailure survival analysis: follow-up data from the Nambour Skin Cancer Prevention Trial. Am J Epidemiol 2005;161:748–54.

21. Kelly PJ, Lim L L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med;19:13–33.
22. Erfani, A. & McQuillan, K. (2008) Rapid fertility decline in Iran: analysis of intermediate variables. Journal of Biosocial Science 40(3), 459–478.
23. Erfani, A. (2017a) Low fertility intention in Iran: the role of attitudes, norms, and perceived behavioural control. Journal of Biosocial Science 49(3), 292–308.
24. Dacourt V, Quantin C, Abrahamowicz M, Blinquet C, Alioum A, Faivre J. (2004). Modelling recurrence in colorectal cancer. J Clin Epidemiol;57:243–51.
25. Gill DP, Zou GY, Jones GR, Speechley M. (2009). Comparison of regression models for the analysis of fall risk factors in older veterans. Ann Epidemiol;19:523–30.
26. Purroy F, Caballero PEJ, Gorospe A. (2013). Recurrent transient ischaemic attack and early risk of stroke: data from the PROMAPA Study. J Neurol Neurosurg Psychiatry;84: 596–603.
27. Therneau TM, Grambsch PM. (2000). Modelling Survival Data: Extending The Cox Model. 2nd edn. New York, NY: Springer.
28. Castan˜ eda J, Gerritse B. (2010). Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death. Rev Colomb Estadı´stica;33:43–61.
29. Erfani, A. (2013) Fertility in Tehran city and Iran: rates, trends and differentials [in Persian]. Population Studies 1(1), 87–107.
30. Erfani, A. (2015a) Family planning and women's educational advancement in Iran. Canadian Studies in Population 42(1–2), 35–52.
31. Erfani, A. (2015b) Tehran Survey of Fertility, 2014: Final Report [in Persian]. National Population Studies and Comprehensive Management Institute, Ministry of Science, Research and Technology, Tehran, Iran.
32. Erfani, A. (2017b) Curbing publicly-funded family planning services in Iran: who is affected? Journal of Family Planning and Reproductive Health Care 43(1), 37–43.
33. Rutstein, S. O. (2011). Trends in birth spacing. DHS Comparative Reports No. 28. ICF Macro, Calverton, MD, USA.
34. Erfani, A. & McQuillan, K. (2014). The changing timing of births in Iran: an explanation on the rise and fall in fertility after the 1979 Islamic Revolution. Biodemography and Social Biology 60, 1–20.

Outlier detection in Poisson Regression Model: Evidence from Bangladesh demographic and health survey data

Sohel Rana[1], Arezoo Bagheri[2]
[1]Department of Mathematical & Physical Sciences, Faculty of Science,
East West University, Dhaka-1212, Bangladesh.
[2]Department of Statistical Methods and Modeling Population,
National Population Studies & Comprehensive Management Institute, Tehran, Iran.

Abstract
The Poisson regression model can be applied to predict a dependent variable of interest in the numerical count type. In demography field of study, births, divorces, and migration could be modelled by poisson regression. In many cases, the assumption of having identical mean and variance for the count response in this model is not fulfilled. So, variability of the data may not adequately be captured. This over dispersion situation may arise due to the effect of outliers, abnormal observations, in the model. Thus, diagnosing these cases is one of the important steps of data analysing specially in the situation of studying national data such as demographic and health survey data. Applying classical diagnostic methods in the presence of outliers may be strongly influenced by them. In this situation, robust outlier detection methods are highly recommended. The main focus of this article is to study different types of poisson regression residuals for identification of outliers and to propose robust outlier detections in modelling children ever born using Bangladesh Demographic and Health Survey (BDHS) data. The merit of the proposed robust approach in detecting outliers in poisson model is also confirmed by simulation results.

Keywords
Poisson Regression Model; outliers; count data.

1. Introduction

Outliers can be defined as observations which deviate significantly from the rest of the data and might be generated by a different mechanism such as experimental abnormalities or errors in the measurements taken (Hawkins, 1980). The anomalies are measures that arouse suspicion due to be much smaller or much larger than the vast majority of the observations. In most of the studies, the presence of a few outliers could be enough to distort the entire results (Fawcett and Provost, 1997). Existence of outliers in real data sets is inevitable and any applied statistician is likely to deal with these observations (Johnson et al., 1998). Thus, the detection of outliers is a valuable process of data analyzing. The easiest solution method to prevent destructive changes of

these observations is to omit them which cause significant modifications in the conclusion drawn from the study. Because of this, knowing how to detect outliers is a vital issue for choosing a suitable statistical analyzing method.

Poisson regression model has received much attention in demographic literature as a model for describing integer values corresponding to the number of birth or death events (Saadati, 2015; Bagheri, 2017). In this model like as any other regression models, to prevent the consequences of outliers, detecting these observations should be considered in the primary analyzing steps (Algamal, 2012). There are some studies devoted to introduce the detection methods in passion regression model (Algamal, 2012). However, to the best of another's knowledge, few studies have been published on applying these methods on demographic data sets.

Comparing to the linear regression model, a limited number of statistics used to detect the outliers in the generalized linear model (Peng et al., 2016). The studies related to diagnostic in Poisson regression models also focus on the identification of outliers, and they mainly study the deviance and Pearson chi-square as diagnostic statistics (Cousineau and Chartier, 2010). Thus, the main purpose of this paper is to define some outlier identification methods in poisson regression model and apply them on a real example of children ever born using Bangladesh Demographic and Health Survey Data. The merit and the robustness of the proposed detection methods are shown by simulation results.

2. Methodology

2.1 Outlier Detection in Poisson Regression by Residual Analysis

Poisson regression is similar to regular multiple regression methods assume that the dependent variable (Y) is an observed count and follows the Poisson distribution (Dobson and Barnett, 2018; McCullagh and Nelder, 1997) with a conditional mean depending on individual characteristics according to the following structural model:

$$\mu_i = E(y_i|x_i) = Exp(x_i^T \beta) \quad (1)$$

Taking the exponential of xβ forces the expected count μ to be positive; this is required for the Poisson distribution. Since, it has been well established to use residuals for the identification of outliers, in Poisson regression model various residuals such as raw residuals, Pearson residuals, Deviance residuals, Studentized Pearson residuals, and Studentized Deviance residuals are available in the literature. Moreover, the Poisson model assumes that the variance is equal to the mean. Though, the variances of the raw residuals are unequal and lead to difficulties in the interpretation of them. The Pearson

residual corrects the unequal variances in the residuals by dividing them to the standard deviations as:

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\emptyset}\,\hat{\mu}_i}}$$

(2)

$$\text{where } \hat{\emptyset} = \frac{1}{n-k}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

The deviance residual is another popular residual due to have the sum of squares of these residuals equal to the deviance statistic. The formula for the deviance residual is:

$$d_i = sign(y_i - \hat{\mu}_i)\sqrt{2\left\{y_i\, ln\left(\frac{y_i}{\hat{\mu_i}}\right) - (y_i - \hat{\mu}_i)\right\}}$$

(3)

The hat values, h_{ii}, are the diagonal entries of the Hat matrix which is calculated by:

$$H = W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{1/2}$$

(4)

where W is a diagonal matrix made up of $\hat{\mu}_i$. They are also used to further standardize Pearson Residuals (p_i) as Studentized Pearson Residuals, $sp_i = \frac{p_i}{\sqrt{1-h_{ii}}}$, and Deviance Residuals (d_i) as Studentized Deviance Residuals, $sd_i = \frac{d_i}{\sqrt{1-h_{ii}}}$. Under the assumption of normality, it is well known that about 99.7% data exist within $\mu \pm 3\sigma$ and observations beyond this limit are considered as outliers. Thus, a data point is flagged as outlier if the corresponding value of the Pearson residuals, Deviance residuals, Studentized Pearson residuals, and Studentized deviance residuals are beyond the range of (-3, 3).

However, it is also known that $\bar{x}$ and and s_d are affected by outliers which leads to violate the rule of $\mu \pm 3\sigma$. Thus, to define a robust measure, $\bar{x}$ and s_d could be replaced by robust location and robust scale, respectively. In this paper, it has been proposed to use two robust locations which are Median and the location of M-estimate using Huber weight function. However, four scales which are Median Absolute Deviation (MAD), S_n and Q_n Estimators, and M-estimate of scale using Huber weight function has been defined for making scale robust to outliers.

The MAD can be defined as:

$$MAD = median\{|x_i - median(x_i)|\}$$

(5)

The estimator of standard deviation for normal population is $\hat{\sigma} = 1.4826 \times \text{MAD}$.

Rousseeuw and Croux (1993) have proposed two robust scale alternative estimators of S_n and Q_n to the MAD as follows:

$$S_n = 1.1926 \times Med_i\{Med_j|x_i - x_j|\} \quad (6)$$

$$Q_n = 2.21914\,\{|x_i - x_j|; i < j\}_{(k)} \quad (7)$$

where $k = \binom{[n/2]+1}{2}$.

The Huber M-estimator for scale is not discussed due to the space limitation. However, one can refer to Huber (1981) and Venables and Ripley (2002).

2.2 Data

In this study, the Bangladesh Demography and Health Survey (BDHS) 2014 data have been used. This data was collected with collaborative effort of the National Institute of Population Research and Training (NIPORT), Macro International, USA, Mitra and Associates. In this study, number of children ever born of women in Bangladesh is regressed to six predictor variables as Divisions (Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Sylhet and Rangpur), age groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49), place of residence (Urban or Rural), educational level (Illiterate, Primary, Secondary and Higher), first birth duration from date of marriage and BMI by poisson regression model.

3. Results and Discussions

3.1 Simulation Study

In this section, the result of a simulation study that is designed to assess the performance of outlier detection methods is reported. Consider the Poisson regression model as:

$$E(Y) = \mu = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) \quad (8)$$

To generate the X values, the sample sizes of 60, 150 and 200 from the uniform distribution in a range of (0, 1.5) with different percentage of outliers (5, 10 and 20 percent) are considered. The model parameters are considered as $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$. The simulation study is drawn based on 1000 bootstrap samples. In this study, Studentized Deviance Residuals have been used as these residuals found to be more suitable residuals than the other classes of residuals considered in this study. However, due to the space limitation, the comparative study of the residuals is not shown. Table 1 presents the Studentized Deviance Residuals and the Studentized Deviance Residuals base different combination of three-sigma limits using M-estimate of Huber location and Median as location parameters, and using MAD, S_n or Q_n as scale parameters.

Table 1: Performance of the different methods of outlier detection

Sample size (*n*)	Methods	Outliers (%)			
		0	5	10	20
60	Studentized Deviance Residuals	0.28	8.82	19.32	63.50
	μ ± 3 σ	0.16	3.33	4.20	1.67
	μ(M-Huber) ± 3.σ(MAD)*	0.60	5.33	10.61	20.07
	μ(M-Huber) ± 3.σ(Sn)*	0.36	5.17	10.34	20.01
	μ(M-Huber) ± 3.σ(Qn)*	0.30	5.10	10.19	19.99
	μ(M-Huber) ± 3.σ(M-Huber)	0.32	5.12	10.24	19.06
	μ(Median) ± 3.σ(MAD)	0.62	5.34	10.57	20.04
150	Studentized Deviance Residuals	0.42	7.85	32.16	81.97
	μ ± 3 σ	0.32	3.33	03.33	0.00
	μ(M-Huber) ± 3.σ(MAD)*	0.56	5.10	10.64	20.44
	μ(M-Huber) ± 3.σ(Sn)*	0.42	5.00	10.50	20.38
	μ(M-Huber) ± 3.σ(Qn)*	0.41	4.94	10.39	20.32
	μ(M-Huber) ± 3.σ(M-Huber)	0.41	4.97	10.37	13.40
	μ(Median) ± 3.σ(MAD)	0.58	5.13	10.73	21.09
200	Studentized Deviance Residuals	0.29	8.59	29.90	82.22
	μ ± 3 σ	0.25	3.97	3.47	0.00
	μ(M-Huber) ± 3.σ(MAD)*	0.36	5.50	10.63	20.88
	μ(M-Huber) ± 3.σ(Sn)*	0.30	5.39	10.52	20.83
	μ(M-Huber) ± 3.σ(Qn)*	0.29	5.33	10.43	20.75
	μ(M-Huber) ± 3.σ(M-Huber)	0.29	5.37	10.42	17.76
	μ(Median) ± 3.σ(MAD)	0.37	5.52	10.69	21.11

It can be noted that for all cases we considered cut-off point is (-3, 3). The performance (percentage of detection to the different nominal levels, 0%, 5%, 10% and 20%) of the methods are then compared in Table 1. The results show that using M-estimate of Huber location and using MAD, S_n or Q_n as scale parameters perform better than the other three-sigma limit to detect the outliers in poisson regression model.

3.2 Application on BDHS 2014 data

3.2.1 Fitting Poisson Model

For BDHS 2014 data set, as the dependent variable (total number of children ever born) is a count variable, the Poisson regression model has been fitted considering six predictor variables as discussed in subsection 2.2. The empirical versus theoretical cumulative distribution function (cdf) have been compared in Figure (1). As a result, Poisson model is fitting well as empirical and theoretical cdf are close to each other.

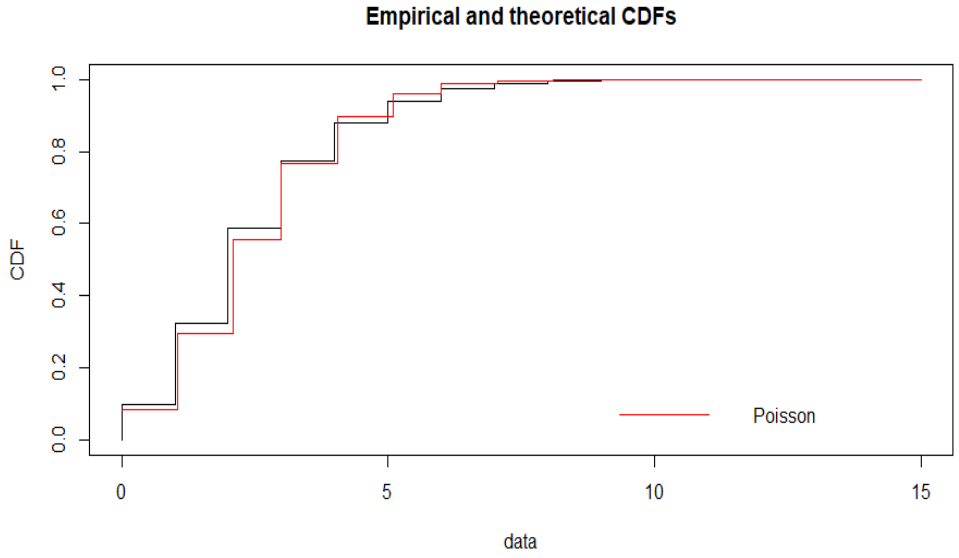
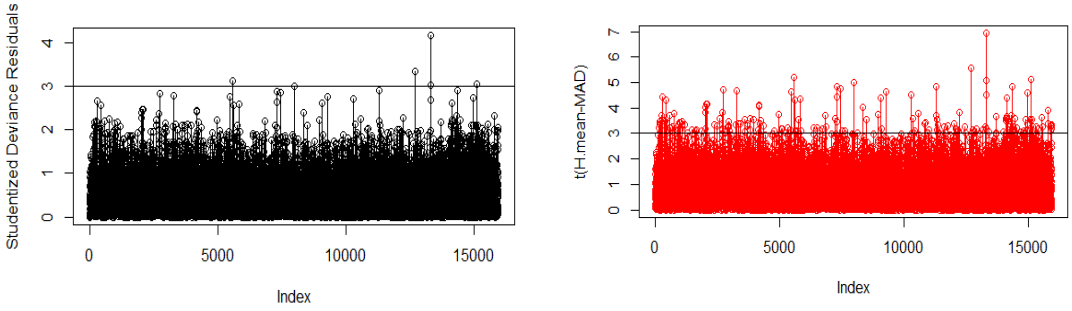


Figure 1: Empirical versus Theoretical Poisson Distribution

3.2.2. Detection of Outliers and their Impact on Poisson Model

In this article, the Studentized Deviance Residuals have been used due to be more reliable than the other types of residuals in outlier detection. Then, the proposed robust three-sigma distances are reused to detect the correct number of outliers in the BDH2014 data based on the Studentized Deviance Residuals. Figure (2) and Table (2) show the detection of outliers by using Studentized deviance residuals and four classes of robust three-sigma distance. From this figure, it is seen that only five outliers are detected by using Studentized deviance residuals as also noted in Table 2. However, the three-sigma rule of using the Huber location and the three classes of robust scales, MAD, Sn and Qn, show different number of outliers (194, 165 and 165, respectively in Table 2). Since the simulation results revile the merit of using the robust classes of three-sigma rules, thus we rely on the robust location and scale base on three sigma rule.



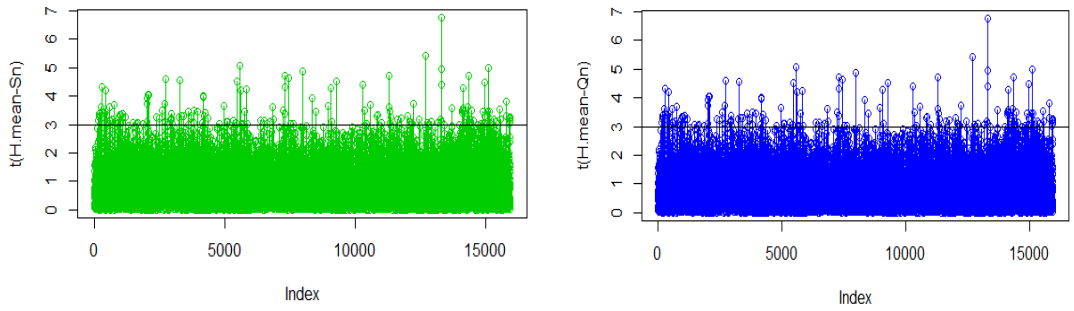


Figure 2: Detection of outliers by using Studentized Deviance Residuals and four classes of robust three-sigma distance.

Table 2. Number of Outliers Detected by each Method

Methods	**No. of outliers**
Studentized Deviance Residuals	5
$\mu \pm 3\sigma$	106
μ(M-Huber) ± 3.σ(MAD)*	194
μ(M-Huber) ± 3.σ(Sn)*	165
μ(M-Huber) ± 3.σ(Qn)*	165
μ(M-Huber) ± 3.σ(M-Huber)	169
μ(Median) ± 3.σ(MAD)	199

A comparative study has been done in Table 3 to see the effect of outliers on Null and Residuals deviance and AIC. As a result, these values become less after removing the outliers. It means that removing the outliers from the data improve the fitted Poisson Model of total number of children ever born.

Table 3. Impact of outliers on Poisson model of total number of children ever born

	With Outliers	**Without Outliers**
Null deviance	14171.5	14040.1
Residual deviance	6977.1	6911.8
AIC	50888	50405

4. Conclusion

In the process of producing, collecting, processing and analyzing data, outliers can be generated from different sources and hidden in many dimensions. The simulation study reveals that the use of robust scale and location in three sigma distance successfully detect the outliers. Thus, before fitting the Poisson model for demographic data where the response variable

is count type, the proposed outlier detection methods can be useful to draw a real prediction and hence helps the researcher to take the right decision of choosing the suitable fitting poission model in presence of the outliers.

References

1. Algamal, Z. Y. (2012). Diagnostic in poisson regression models. Electronic Journal of Applied Statistical Analysis, 5(2), 178-186.
2. Bagheri, A. (2017). Studying the Influential Factors of Children Ever Born of Migrant Women to Tehran, Journal of Ilam Medicine University, 25(6), 118-129.
3. Cousineau, D. and Chartier, S. (2010) Outliers Detection and Treatment: A Review. International Journal of Psychological Research, 3, 58-67.
4. Dobson, A.J, Barnett, A.G. (2018). An Introduction to Generalized Linear Models, CRC Press, 2018, New York.
5. Fawcett T., and Provost F (1997). Adaptive fraud detection, Data-mining and Knowledge Discovery, 1(3), 291–316.
6. Hawkins, D. (1980). Identification of Outliers. Chapman and Hall.
7. Huber, P. J. (1981). Robust Statistics. Wiley: New York.
8. Johnson, T., Kwok, I., Ng, R. (1998). Fast Computation of 2-Dimensional Depth Contours. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 224–228. AAAI Press.
9. McCullagh FRS, P and Nelder FRS, J.A. (1997). Generalized Linear Models, 2nd Edition, Chapman & hall: New York.
10. Peng, L.Y, Midi, H. Rana, S and Fitrianto, A. (2016). Identification of Multiple Outliers in a Generalized Linear Model with Continuous Variables, Mathematical Problems in Engineering, Volume 2016, Article ID 5840523, 9 pages.
11. Rousseeuw, P.J., and Croux, C. (1993). Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association, 88(424), 1273-1283.
12. Saadati, M. (2015), Factors Affecting the Number of Children Ever Born of 15-49 Year-Old Women in Semnan by Poisson Regression Model. Journal of Health System Research, 11(3),627-637.
13. Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. 4th edition. Springer.

A new GDP publication model in the UK

James Scruton, Sumit Dey-Chowdhury
Office for National Statistics

Abstract

The Office for National Statistics (ONS) has recently implemented one of the most notable changes to how it publishes estimates of UK gross domestic product (GDP) in the last 25 years. One of the main purposes was to improve the trade-off between timeliness and accuracy of early estimates, specifically with a view of producing a higher-quality initial estimate that would be less susceptible to revision. These changes have also enabled the UK to become one of a select few countries to produce estimates of monthly GDP. This paper sets out how GDP estimates have historically been compiled in the UK, then describes how this inherent trade-off has been improved under the new publication model. It also includes some analysis of the monthly profile of GDP, explaining how the development of this higher- frequency indicator can provide additional signals to policymakers and users.

Keywords

Gross domestic product; UK economy; economic statistics

1. Introduction

The Independent Review of UK Economic Statistics (Bean, 2016) was commissioned with the aim of identifying the needs that relate to the challenges of measuring the modern economy and assessing the effectiveness of the Office for National Statistics (ONS) in delivering those statistics. One of the challenges explored whether "a slightly later publication of the preliminary estimate [could] lead to a material reduction in the magnitude of subsequent revisions". In 2018, the ONS introduced changes to how it publishes estimates of UK GDP, reflecting some of the most notable developments to the compilation process in the last 25 years. One of the main purposes was to improve the inherent trade- off between timeliness and accuracy of early estimates, producing an initial estimate that would be less susceptible to revision. These changes also enabled the production of official estimates of monthly GDP for the first time in the UK, thus allowing the production of a more coherent and timelier picture of the UK economy and so providing users with a more informed steer of how the economy is evolving.

This paper explains how GDP estimates have been compiled in the UK, setting out the motivation for the change to the publication model and how

this trade-off between timeliness and accuracy has been improved. It includes analysis of the newly published monthly estimates of GDP, highlighting how it can provide additional insights into developments in the UK economy.

2. Methodology

In the UK National Accounts, there are three ways in which GDP is estimated. These are then balanced to produce one single estimate of GDP.

- *Output or Production:* This is the value of the output of goods and services that are produced, less the intermediate inputs used in their production, plus any taxes net of subsidies on those products.
- *Income:* This records the value of income earned by households and businesses in the production of goods and services, plus any taxes net of subsidies on production and products.
- *Expenditure:* This is the value of the final expenditure on goods and services by households, businesses and the government, plus net exports of goods and services.

GDP is first measured in current prices, with the effects of price changes then removed to produce volume estimates of GDP. The international guidance is to confront these estimates using the Supply and Use Tables (SUTs) framework, which reconciles at a detailed level the supply of goods and services with their use, incorporating all available information on output, income and expenditure.

However, the necessary information to allow this more detailed reconciliation is only available after around 18 months, so early estimates of GDP are reconciled only at a headline level. In the UK, a process of single extrapolation is applied to produce these initial estimates. The most recent balanced annual estimate of volume GDP is extrapolated using an output-based indicator. The changes to the publication model in this paper only relate to this process.

Quarterly GDP

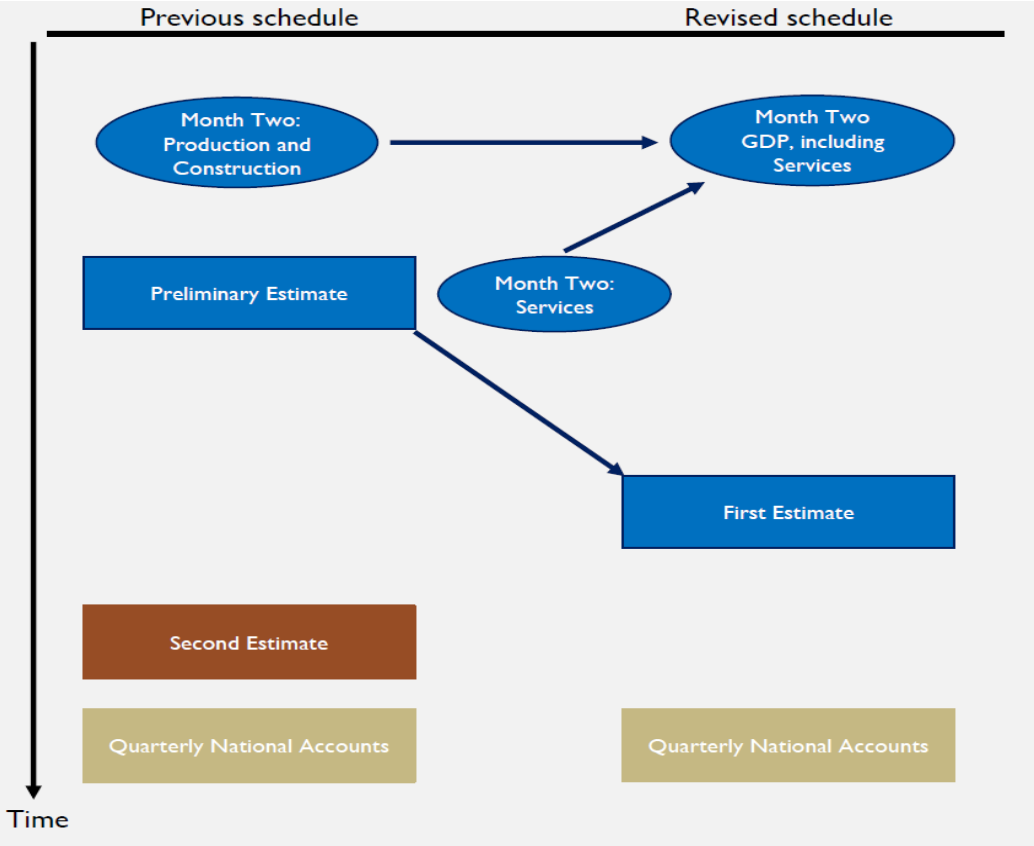
There is an inherent trade-off between the timeliness and accuracy in any estimate of GDP. The user need is typically for timely information, so most economic statistics are published shortly after the quarter to which they refer. However, given the short space of time with which these early estimates are produced, these can only be based on incomplete information. Therefore, each vintage of GDP has varying levels of data content for output, income, and expenditure. As time progresses, more information becomes available and so early estimates are subsequently revised resulting in the production of multiple vintages of data for the same economic variable.

Skipper (2005) provides a detailed overview of the information content that is contained within different estimates of GDP under the previous model.

The ONS produced three initial estimates of UK GDP, published 25, 55 and 85 days respectively after the reference quarter, with the data content increasing with each of these vintages (45%, 65% and 90%). There is a higher data content for the output estimate, so it is considered the most appropriate proxy to determine the short-term movement of quarterly GDP. Under the previous model, only output information was available for the preliminary estimate. As income and expenditure estimates were produced, these were then initially aligned to the output estimate.

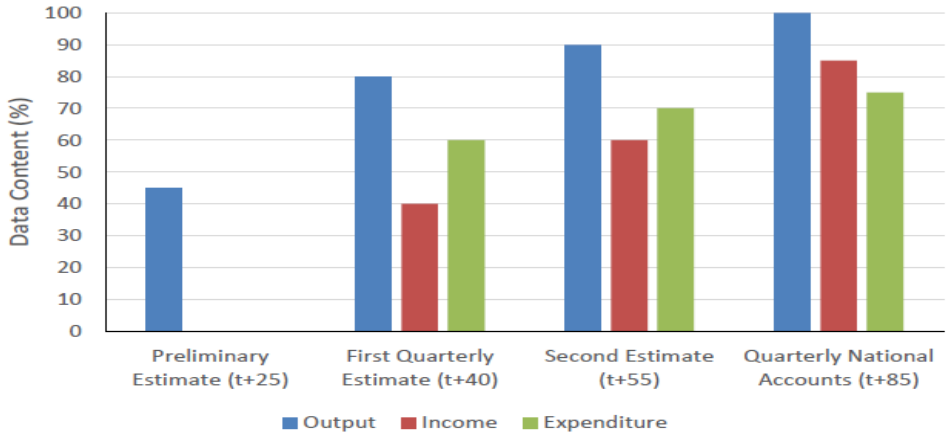
However, The National Statistics Quality Review: National Accounts and Balance of Payments (Barker and Ridgeway, 2014) highlighted that this model made limited use of information relating to expenditure and income in the early estimates of GDP, commenting that these would have very little influence on the path of overall economic activity estimated until these data are confronted in a SUT balancing exercise. The motivation for developing a new GDP publication model was to consider whether the quality of that first estimate could be significantly improved if it were published with a slightly longer lag, enabling it to incorporate richer information from all three approaches and so providing scope for information from the income and expenditure estimates to be incorporated (Figure 1). The new first quarterly estimate of GDP is now published 40 days after the end of the reference quarter, with an increased data content that reflects output estimates now being fully available for the third month of that quarter, as well as initial estimates of income and expenditure.

Figure 1: Change in the GDP Publication Model



The new model reconsiders the balance between the timeliness and accuracy of GDP estimates, thereby looking to reduce the likelihood and frequency of revisions to the first estimate. The publication of a slightly later first estimate allows output estimates to be available for each month in the quarter, as opposed to the previous model where it had largely been forecast for the final month (Figure 2). As such, output will still be considered the best indicator of economic activity initially, and so GDP will still be balanced to output in our short-term estimates. However, it now allows any signals in the income and expenditure estimates to be considered as part of the balancing process itself, as this is now available at this point after the reference quarter (Scruton et al, 2018).

Figure 2: Availability of Data for each GDP Measure from the End of the Reference Quarter



Monthly GDP

There has been much user demand for having high-frequency estimates, leading to the production of unofficial indicators and 'nowcasts' by other UK institutions to complement other official estimates produced. There has been a proliferation of business surveys that aim to provide timelier news on short-term movements in the UK, while financial markets may provide another guide. Higher-frequency estimates tend to be in higher demand when the economy is in less stable times, as timely and reliable signals are necessary for policymakers to respond.

As part of the changes to how quarterly estimates of GDP are produced in the UK, the ONS reviewed the scope for developing monthly estimates of GDP. This advances research in this field to provide a better and timelier understanding of the short-term evolution of the economy, bringing together, estimates of production, construction and services output in a much more coherent manner, which would provide almost complete coverage of the UK economy.

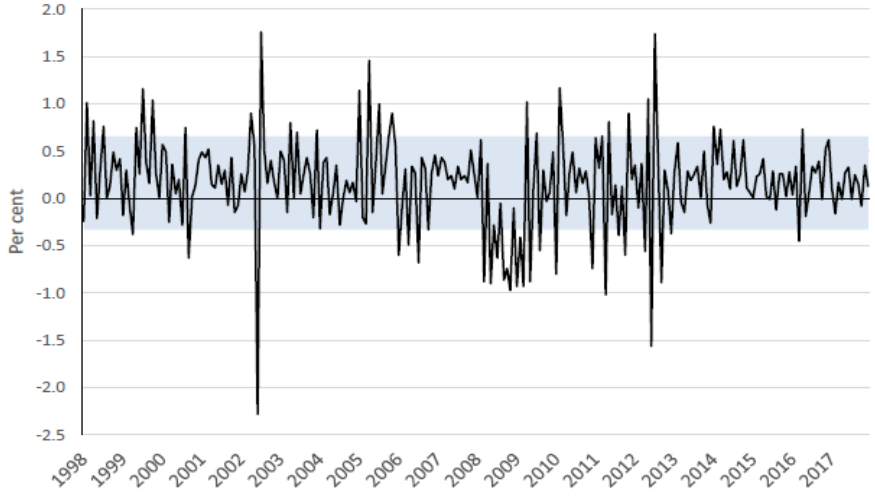
Monthly estimates of GDP will coincide with the new first quarterly estimates, where the latest three- month period reflects the calendar quarter. However, to ensure consistent reporting in the monthly and quarterly estimates, it has been necessary to incorporate some changes to the balancing process. The headline GDP estimates now aligns to output in the current and previous quarter, as monthly estimates are not available on income on expenditure to produce an average GDP estimate for the previous three-month period. That said, there is now scope for the higher content from income and expenditure to form part of the balancing process itself, specifically applying quality and balancing adjustments to output estimates if needed.

The provisional estimate for the first quarter of 2018 showed that GDP had slowed to 0.1%, in part reflecting the effects of the heavy snowfall that affected the UK in March of that year. This would have likely have had a dampening effect on activity. However, the first estimate would have been subject to higher levels of uncertainty, particularly as activity would have largely been forecast for that month of the quarter under the previous model. This may explain why there is some evidence that GDP can be revised more than usual after periods of heavy snow (Ramsden, 2018). Under the new model, there would have been considerably higher output information in the new first estimate, as well as additional insights from the income and expenditure approaches that would have informed the balancing process. The higher-frequency indicator would have also provided a clearer picture on the extent that displacement may have taken place. This would have reduced some of the uncertainty of that initial estimate, providing policymakers with more confidence on whether the weakness reflected the weather or the economic climate.

3. Results

One of the most striking features of monthly GDP is its inherent volatility, which is particularly pronounced around the time of the financial crisis in 2008 and 2009 (Walton and Dey-Chowdhury, 2018). Figure 3 shows the latest estimates of monthly GDP growth over the 20-year period from 1998 to 2017, where further revisions would be expected to reflect largely methodological improvements, as opposed to the incorporation of further survey information. It highlights specific periods where the monthly growth rate falls outside of the band of one standard deviation from this 20-year average.

Figure 3: The Effects of Special Events and the Financial Crisis on Monthly GDP



There are four periods where monthly GDP has had a much more volatile path than might be expected, excluding 2005, where there was increased volatility caused by the challenges of profiling a monthly path through some of our quarterly services estimates.

- Summer 2002: The Queen's Golden Jubilee saw two Bank Holidays in early June. The first one was delayed from May, while the second was an additional one leading to a change in the number of working days in that month. This contributed to a fall in monthly GDP of 2.3% in June, which then then recovered in July as expected, rising by1.8%.
- 2008 to 2009: The recession caused by a financial crisis, with GDP falling 6.3% peak to trough. The monthly movements within this period are not very volatile as all monthly growths fall within quite a narrow range, but the widespread effects of the credit crunch are plain to see.
- Late 2010 to early 2011: A range of special events dominated the monthly movements including heavy snowfall causing disruption in December 2010 and an additional Bank Holiday to commemorate the wedding of the Duke and Duchess of Cambridge in April 2011. This extra Bank Holiday fell soon after Easter and many people took the opportunity for an extended break, as monthly GDP fell by 1.0%.
- Summer 2012: The Queen's Diamond Jubilee took place in June 2012, marked by two Bank Holidays. As in 2002, the first Bank Holiday was delayed from May and the second was an extra one. GDP fell by 1.6% in June, followed by a sharp pickup of 1.7% in the following month. Furthermore, the London 2012 Olympic and Paralympic games took place between July 27th and September 9th, which impacted upon economic activity over this period.

Other than the impacts of the financial crisis, these episodes highlight the effects of lost and/or displaced activity from one-off factors. This has a more pronounced effect on monthly GDP estimates, as there is less scope for activity to be displaced within a month. These temporary effects may be smoothed out within a quarter, but this is likely to be case in the monthly profile so there is an increased likelihood of this being reflected in a more volatile monthly path.

4. Discussion and Conclusion

There is a trade-off between the timeliness and accuracy in producing estimates of GDP. Each vintage has varying levels of data content for output, income, and expenditure. As time progresses, more information becomes available and so early estimates are subsequently revised. There has been a challenge as to whether there is scope for improving early estimates of GDP, by producing a slightly later first estimate. The recent changes to how estimates of UK GDP are produced have not only strived to improve this trade-

off, but it also enabled the production of the official estimates of monthly GDP for the first time in the UK. This has enabled the communication of a more coherent and timelier picture of the UK economy, providing policymakers and forecasters with a more informed steer of how the economy is evolving.

Although understandably inevitable, revisions pose a challenge to economists in assessing the true state of the economy. This is a particularly pertinent area to policymakers as data uncertainty can having a significant impact on policymakers' view of aggregate demand and supply in the UK economy. The scope of future research will look to see if the changes to the publication model have helped improve the quality of initial estimates, as the ONS looks to develop real-time monthly estimates of GDP.

This is also part of a wider transformation of how GDP is compiled in the UK. In 2019, the ONS will introduce a new framework to produce GDP, both in current prices and in volume terms. This will include the introduction of double deflation in the UK National Accounts, which is widely considered as the best approach to producing volume estimates of activity.

References

1. Barker. K and Ridgeway. A (2014), "National Statistics Quality Review: National Accounts and Balance of Payments"
2. Bean. C (2016), "Independent Review of UK Economic Statistics"
3. Ramsden. D (2018), "What's Going On?"
4. Scruton. J, O'Donnell. M and Dey-Chowdhury. S (2018), "Introducing A New Publication Model for GDP"
5. Skipper. H (2005), "Early Estimates of GDP: Information Content and Forecasting Methods"
6. Walton. A and Dey-Chowdhury. S (2018), "A Guide to Interpreting Monthly Gross Domestic Product"

Development in merchandise trade indicators

Valdone Kasperiuniene, Sophie Limpach
European Commission EUROSTAT

Abstract

Merchandise trade statistics is one of the most popular statistical areas among the users. Therefore the development of new trade indicators is very important in order to be able to satisfy growing needs for information. However increasing pressure on the resources and on the reduction of the administrative burden requires that statisticians make better use of already existing administrative and statistical data.

In the first part, the paper provides an overview of traditional European trade indicators focusing on the challenges to produce more with less. It highlights the importance of close administrative cooperation of institutions for the benefit of the information society.

Later, the recent achievements and plans for the development of new statistical products such as statistics of trade by enterprise characteristics and trade statistics by invoicing currency are presented. Trade by enterprise characteristics enriches the traditional output of trade statistics by providing a closer view of traders and their economic characteristics. This statistics is extensively used for the economic analysis of the globalised economy. Statistics by invoicing currency complements trade statistics with additional dimensions, providing information on the currencies traders use for trade outside the EU. This information is valuable for the analysis and monitoring of the role of the EURO and other currencies in international trade.

Finally, the document describes efforts of European trade statisticians in developing new trade indicators. The potential use of the new data available on the EU customs declaration is discussed. Although these data are currently outside of the traditional scope of trade statistics, it could provide users with useful information. Specific insight is provided in recent efforts of the EU Member States to improve the relevance and the quality of the information collected under the nature of transaction. This information is used to identify specific trade flows required for the compilation of balance of payments statistics. Nature of transaction facilitates the conceptual integration of merchandise trade and balance of payments statistics. In this context the indicators needed to measure globalisation are discussed.

Keywords

Globalisation; trade by enterprise characteristics; burden reduction; balance of payments; customs statistics

1. Introduction

Merchandise trade statistics is one of the oldest and the one of the most popular statistical areas among the users. Various users have different needs and expectations towards trade statistics. The analysts use it to measure economic performance of the countries and businesses need them to perform market analysis. The data are required to compile macroeconomic statistics such as balance of payments and national accounts. Some users are looking for traditional trade indicators, whereas the macro economic statistics needs data which conceptually are aligned to their requirements. On the other side, the researches and analysts wish to measure globalisation effects and to be able to see trade as part of the integrated business activities. Statisticians must find solutions for these requirements.

However to satisfy all specific user needs is very challenging for producers of statistics, as growing pressure on the resources and on the reduction of the administrative burden makes it difficult to introduce new statistical surveys which would provide tailored information. For this reason, European statisticians focus on the better use of already existing administrative and statistical data for the development of new trade indicators in order to satisfy growing requests for information.

This paper presents the achievements and challenges of European Statistics in producing new statistical products and in exploring of the new indicators.

2. Methodology

2.1 European trade statistics system and indicators

European trade statistics is based on two data collection systems. The trade data with non-EU countries are collected by customs administrations. Customs declarations are used for statistical purposes as the basic data source, which provides detailed information on exports and imports of goods. The trade between EU Member States is collected via dedicated statistical business survey – Intrastat – where data are collected directly from the traders. The Intrastat system was introduced following introduction of the European Single Market when customs formalities between EU Member States were removed. The subsequent loss of statistical data source required the establishment of a new data collection system for recording trade between EU Member States.

Both data collection systems have their advantages and disadvantages. Customs administrations provide full coverage of trade transactions, therefore compilation of statistical data do not pose any additional administrative burden on the traders. However customs procedures are not always aligned to the statistical needs and any administrative changes in Customs impacts directly statistical data. Consequently compilation of trade statistics requires additional efforts for reconciliation of statistical and customs requirements.

The direct reporting of information via Intrastat system allows compiling statistics which focus only on statistical needs, however it has high costs and puts an important administrative burden on traders. It is very challenging for statisticians to balance contradicting expectations between the data providers, who want data reporting to be easy and simple, and data users, who need detailed information with high quality and require new statistical indicators and new statistical products.

In order to reach the balance between administrative burden and increasing user needs, European trade statisticians are constantly looking for additional administrative and private data sources which while complementing customs and Intrastat declarations can ensure burden reduction needs on the one side and, good coverage, quality and production of European trade statistics, on the other side. The current needs for trade indicators can be fulfilled only in combination of variety of data sources.

In this context, close administrative cooperation and easy access to all administrative data sources is the first prerequisite for the compilation of better and more statistics without increasing of administrative burden. European National Statistical authorities have established good cooperation with tax administrations which provide them with the wide range of value added tax (VAT) data used for complementing statistical data collection via Intrastat system and quality assurance. Thanks to the use of VAT data, only 17% of all intra-European traders are liable to report trade information via Intrastat system, whereas the remaining information is available from administrative forms. The use of international private and administrative shipping and aircraft registers ensures coverage and implementation of the change of economic ownership principle in trade in vessels and aircraft. Statistical business registers provide the structural information about the traders.

With the ultimate goal to decrease statistical (Intrastat) burden on traders, a new data exchange system is currently being built by Eurostat and EU Member States. The major element of the system is the creation of an additional trade data source by exchanging of micro-data on intra-EU exports among EU Member States. The trade statistics transactions by their nature are symmetrical vis-a-vis to the trading partner country and, therefore, they are recorded twice: the exporter records exports in the exporting EU Member State and, consequently, the same transaction is recorded in the importing EU Member State. Implementation of wide scale micro data exchange on a monthly basis has got a high potential to reduce burden on importing enterprises. In fact, the exports statistical information collected by a EU Member State will be transmitted to the concerned importing EU Member States who can then use this information to compile the respective imports statistics.

The reduction of administrative burden is not the only aim of the new micro data exchange system as it also enables a more detailed analytical work on asymmetries resolution as well as creates potential opportunities to develop new statistical indicators.

A key feature of the micro data exchange system is that next to the traditional data elements like product, value, quantity, partner country, etc., two new data elements will be collected and exchanged on exports: country of origin and the ID number of the partner trader in the importing EU Member State.

This information will enable importing EU Member States to improve coverage and the quality of the imports data without putting additional burden on traders. At the same time exports data will benefit from collection of country of origin of goods on exports, which will allow compilation of more precise information on the national exports and re-exports of goods. More over, the link established between exporter and importer in combination with other data sources has a high potential for the development of some globalisation indicators

2.2 Innovative statistical products

Trade by enterprise characteristics indicators

Trade by enterprise characteristics (TEC) is a more recent statistical field of European statistics, which started a decade ago with very first pilot compilation projects. Today this is a well developed statistical area at European level with established definitions, methodology, quality and legal requirements and IT tools. TEC statistics serve as a good example of an innovative statistical product, whose compilation does not put any burden on data providers as the data are compiled from already available data sources. By applying statistical matching and linking techniques on micro data, the compilers are able to create a new integrated product providing the users with enriched content on traders and bringing additional insights on the trade data.

The main objective of the TEC statistics is to bridge two major domains - business and trade in goods statistics - which have traditionally been compiled and used separately. For the business-related information, the statistical business register (SBR) is the main data source, which provides business characteristics such as number of employees, economic activity code, turnover and ownership information. The registers of intra- and extra-EU trade operators and the SBR are linked through a common unit of reference, namely the legal unit. However, in order to achieve better alignment with the business statistics and to enhance the structural comparability of the data among the EU Member States, the TEC statistics is compiled on enterprise concept.

In principle, the TEC statistics describe by size classes, economic activity sectors and exports intensity the businesses which are behind the trade flows.

The TEC indicators can be used to measure the performance of small and medium size enterprises and their exports diversification.

On the other side, the traders are characterised according to their trade patterns, such as number of partner countries and products traded broken down by economic activity sectors.

In addition, the TEC statistics steps ahead in measuring trade globalisation, as it allows identifying performance of multinational enterprises, which have affiliates abroad or the enterprises which are controlled by foreign companies.

Currently the EU Member States compile 10 TEC data sets[1], which provide users with the possibility to calculate various TEC indicators. The compilation of 6 of those data sets is mandatory since 2010, whereas the others are currently still optional.

At the start of the compilation of TEC statistics Eurostat and the EU Member States focused on three major issues: to establish data linking procedures, to assure access to the required SBR information and to develop the necessary IT tools for the production, validation and transmission of the data.

In the meantime, the EU Member States and Eurostat have achieved very important results in the production of TEC indicators. All EU Member States implemented SDMX – ML data transmission format[2] using standard code lists specific for TEC data, which can be reused by other organizations compiling TEC statistics as well. The compilation methodology and the data validation rules were developed and published in the *Compilers Guide on European statistics on international trade in goods by enterprise characteristics.* The quality metadata are provided in the annual quality reports and the data are disseminated to the users via multiple electronic dissemination channels.

Every year the number of TEC users is increasing so the requirements towards quality and availability of the data are growing. For this reason European TEC producers started the next stage in the development of TEC statistics focusing on the quality of the matched data through the improvement of the compilation methods and the enhancement of the comparability of data across EU Member States. Particular attention will be given to better implementation of enterprise concept in the SBR and further harmonisation of the compilation methods.

A new look on TEC data will be provided to the users, which will allow to identify type of traders according to the specific categories (e.g. resident enterprises, private individuals, non-resident traders, etc.). The harmonised attribution of traders to the categories will help to improve conceptual

[1] The published TEC data sets can be found on Eurostat webpage.

[2] The Trade by Enterprise Characteristics DSD is available on Euro SDMX Registry with the following specifications: DSD agency: ESTAT, DSD Name: TEC; DSD Version: 1.2.

alignment of TEC with the business and the balance of payments statistics and, at the same time, will help users in their interpretation of the TEC indicators. The information about the non-resident traders will provide value added for measuring impact of globalisation in the EU.

Trade by invoicing currency

The trade by invoicing currency (TIC) statistics provide information on the currencies the traders use for the trade outside the EU. Goods imported and exported by traders can be invoiced in a range of currencies. The data enable analysts to measure the performance of EURO and other currencies in international trade and to research the markets in relation to the exchange rates. There are many factors which determine the choice of the invoicing currency and the TIC data can help users to ground their assumptions in economic analysis.

The data needed for compilation of the statistics on trade by invoicing currency (TIC) are collected on the customs declarations, which serve as the primary data source. The TIC statistics is compiled on the annual trade data and have been produced since 2010 at least every two years.

Although the TIC data cannot be considered as directly innovative product, it has an important value, in particular, for financial and monetary policy analysis. It is, therefore, worth to be mentioned in this context, as the TIC statistics increase the usability of the trade data without putting additional burden to the respondents.

The invoicing currencies are broken down by major product groups. Further breakdowns are considered to be produced in the future. The TIC data quality is regularly assessed (whenever data are produced) and monitored via the quality report on trade in goods statistics.

2.3 In search of new trade indicators

In the past, the majority of trade transactions were straightforward and concerned two countries only, where one country was selling and another country was buying the goods. The trade transactions, in most of the cases, ended in the change of economic ownership of goods. The trade statistics reflected physical flows of goods that, in most cases coincided with the financial transactions. However during the last few decades with the deepening of globalisation, where the production sites and processing operations can be divided and performed in several countries, the movement of goods between countries do not necessarily imply real trade transaction and change of economic ownership of the goods. The increasing share of triangular trade and complex chain transactions dissociate the financial flow from the physical flow of goods, which consequently impacts statistical data

compilation and changes the content of the trade statistics in comparison to the past.

The users look at the trade statistics from three different perspectives. Some of them need the data which shows the amount of goods which are transported from one country to another, the others need to analyse the trade by country of origin, whereas the compilers of macro economic statistics require the trade data to be compiled based on the change of economic ownership of goods between resident and non-resident.

According to international recommendations, the primary requirement for trade statistics is to record a partner country based on where the goods are physically transported and, in addition, to provide information on the country of origin of those goods. However the partner country with which the trade transaction takes place, i.e. the country where the seller or the buyer is established is not applied in the European trade statistics. In the globalised world, the three types of partner countries may not be the same and, therefore, statistical results would be different for each type of partner country definition as well.

Better use of customs data

Country of purchase

Production of trade data by country of origin, country of consignment, country of destination, country of sales and country of purchase would significantly enlarge the scope of economic analysis and would provide users with the additional tools for measuring the impacts of globalisation. The macro economic statistics in particular would benefit from the trade data compiled on the basis of country of purchase, because the data would be better conceptually aligned to their requirements.

For this reason the efforts are being made to search for the new data sources. Changing environment and the recent developments in the European Customs bring new possibilities for European trade statisticians.

Following the European Commission's e-customs initiative, EU Member States customs administrations are implementing the new Union Customs Code, which harmonises interoperable European IT-systems. The new exports and imports clearance systems will be deployed in the near future and the new data requirements (which are compatible with the World Customs Organization's Data Model) will gradually be introduced. This will allow trade statisticians to analyse the additional data elements available at Customs and to assess their potential use for the trade statistics purposes.

The statisticians are particularly interested in the new information about the seller and the buyer of the goods, which includes ID numbers and the identification of the country. In order to satisfy macro economic statistics needs, the collection of information about the seller and the buyer is highly

desirable. The collection of the country of purchase would help to compile statistics according to the partner country with which the change of economic ownership has occurred.

Improved measurement of e-commerce

The development of Internet and digitalisation has created new business models and new opportunities for consumers, which means that international trade transactions are executed not only between businesses, but increasingly between businesses and the final consumers.

Cross-border e-commerce, where the goods are sold via internet by businesses to private individuals abroad continues to accelerate in the EU and is achieving significant trade amounts, in spite of the fact, that the majority of this trade relate to the very low transaction values in comparison with the traditional trade.

The interest in the e-commerce statistics and the number of the users who are requiring the data is increasing as well. However it was not possible to compile this statistics, because the large part of the low value e-commerce transactions were not recorded on the customs declarations.

The new initiative by the EU Customs to harmonise data collection for the low value transactions will allow trade statisticians to make better use of the customs data from 2021 onwards. The low value imports transactions carried out by private individuals will be systematically recorded in customs IT systems. In addition, a new and specific data requirement for recording imports by postal operators will be introduced.

The changes in customs procedures will enhance the relevance and coverage of imports and will ease estimation of the e-commerce.

Better use of available data elements

Nature of transaction

The Nature of Transaction (NoT) is a data element collected on customs and on Intrastat declarations which serves exceptionally statistical needs. It characterises the different features (purchase/sale, work under contract, etc.) of trade transactions, which are required either for balance of payments and national accounts purposes or for the identification of certain flows of goods, which are excluded from trade statistics.

A two-digit coding system is used to differentiate between types of trade. There are usually three dimensions to be considered when determining the nature of a given transaction, namely physical movement of goods, change of ownership and financial compensation. Over the years, the content of the code list was changing in line with the changing user or compilers needs.

The current NoT coding was introduced in 2010. Since that time, the user needs have evolved and some NoT codes have become obsolete or no longer

fully fit for the purpose, therefore the EU Member States are currently revising the content and the code list of this indicator.

The new emerging needs are significantly driven by globalization process. The users wish to measure which part of international trade is carried out between the branches of the same enterprise (i.e. intra-group trade), which part of trade relate to quasi transit[3], what is the share of national exports and reexports or to know what is the amount of processing trade. For the macro economic statistics, it is very important to identify transactions where change of economic ownership occurs.

The wish list for the desirable codes is quite large, however, it is acknowledged that the NoT coding cannot cater to all needs at the same time without making the codification overly complicated. Too long list, complicated definitions and concepts will not yield good results during the collection process as the provider of information may not supply reliable data.

The discussions among EU Member States are ongoing on how to measure the Intra-group trade, as direct collection of this information imposes a significant burden on the respondents. Moreover, it may not provide with the satisfactory results because of the complexity of concepts. The implementation of the micro data exchange between EU Member States and collection of the ID number of the partner trader will open up new possibilities for pilot studies on measuring intra-group trade flows based on the structural information of the multinational enterprise provided in the EuroGroups Register[4].

3. Discussion and Conclusion

In the rapidly changing world where trade transactions are changing forms, the statisticians must find new ways and new data sources to respond to growing user needs. Searching new data sources, better using administrative data and implementing an integrated approach in business and trade statistics via micro data linking techniques on one side and innovative data collection methods on the other side are priorities for European trade statisticians.

The conference is providing a good opportunity to exchange experiences of the countries in the development of the new trade indicators and innovative data compilation techniques.

[3] Quasi-transit is related to the functioning of the Customs Union and Single Market. It occurs when goods enter/leave an economy and are declared as imports/exports for customs purposes without the transit economy having acquired ownership of the goods. Quasi transit should be excluded from international trade for Balance of Payments and National accounts purposes.

[4] The EuroGroups Register (EGR) is the statistical business register of multinational enterprise groups having at least one legal unit in the territory of the EU or EFTA countries.

Assesing the quality of Indonesian Merchandise Trade Statistics (Mirror Analysis Approached)

Mila Hertinmalyana, Purwaningsih
BPS-Statistics Indonesia, Jakarta, Indonesia

Abstract

International Merchandise Trade Statistics is a strategic data for policy makers, both for making decisions and trade negotiations. For making a good decision and on target, it has to be based on data that fulfilling certain quality. This paper is aim to assess the quality of Indonesian merchandise trade, both exports and imports, by using mirror techniques. Beside single bilateral mirror comparison, this paper also use multiple mirror comparison technique to examine the Indonesia trade statistics. In addition this paper also examine the reason behind the discrepancy data of export and import.

To get deep picture of mirror analysis, the authors do exercise of Indonesia trade to some major partner's countries for aggregate level. The import-export ratio of Indonesia's export with all countries considered in medium level, the range is between 0.13% - 0.31%, which is still acceptable. The import-export ratios between Indonesia and some major trade partners are dominated by medium level. Detail of HS 2 digit and 6 digit give the misclassification information.

Keywords

Indonesia international trade, mirror technique, discrepancy, major partner countries

1. Introduction

International Merchandise Trade Statistics (IMTS) as one of economic indicator of a country is very important to be evaluated continusly to guarantee its quality. UNSD has published the Manual of International Merchandise Trade Statistic as a reference for all member countries in compiling the merchandise trade. Ideally, if country X exports goods "A" to country Y, the exports value of country X should be same with the imports value of country Y. Because in theoretically, export and import is as a "mirror" one another. However, in many cases, the value of exports and imports between two countries are often shows big differences. This is also happen to Indonesia, as a member of UN countries. To evaluate and examine the gaps/discrepancy, some analytical or studies undertaken by some statistician and economist. One of the popular method is " mirror analysis'.

The cause of differences (asymmetry) in international trade statistics are various (Hamanaka, 2011). At first, the differences in methodology of export and import will lead the discrepancies. As IMTS recommendation, exports value is recorded in FOB (Free on Board) but imports value is recorded in CIF (Cost Insurance and Freight). So the imports value will tend to higher than the exports value.

The next possible reason in asymmetry is the different action in recording data by customs office. For example in Indonesia, the accuracy level of data collected by customs office is very different between exports and imports. Indonesian imports data from Customs are more accurate and reliable because it related with revenue (tax) of the country. The Customs office check the documents (import declaration) very carefully and detail. While for exports, the assessment of the document (export declaration) is not so tight and detail as import, except for some commodities.

As mentioned above, International Merchandise Trade Statistics as one of economic indicators is the crucial data for policy makers, both for making decisions and trade negotiations. The right data will drive the right decisions, but the wrong data will drive the wrong decisions. So, in order to be useful data, they must satisfy certain quality standards, including being consistent among countries and also consistent over time.

This paper will assess the quality of Indonesian merchandise trade, both exports and imports, by using mirror techniques. At the first, this study will examine Indonesian trade with all partner countries in the world as general. Then this study will focus on the trade between Indonesia and major partner countries, both in export and import side.

2. Methodology

Methodology applied in this paper is method used by Hamanaka (2011), multiple mirror comparison. According to Hamanaka (2011), this method can tell more about the manner in which miss classification than single bilateral mirror comparison.

2.1 Discrepancy and Asymmetric Commodity Groups

To assess the size of discrepancy between two sides of the mirror, this study will use the import-export ratio as below:

- Import-Export Ratio (aggregate level) = Aggregate Import-side Data/Aggregate Export-side Data
- Import-Export Ratio of 6 digits HS = Import-side data in digit X/Export-side data in digit X

There are two kinds of discrepancies, positive discrepancies and negative discrepancies. The positive discrepancy refers to the case where the import-side data is larger than the export side data more than 10 percent. The

negative discrepancy refers to the case where the import-side data is smaller than export-side data, or the import-side exceeds the exports side data less than 10 percent.

2.2 Assess Direction Misclassifications

The assessment of direction misclassification by multiple mirror technique involves two steps:

1. Compare a test country's bilateral trade data with its major partners bilateral trade statistics at the aggregate level and identify a seeming set of false and actual destination countries whose trade with the test country generates discrepancies in opposing directions (positive and negative discrepancies)
2. Compare a test country's bilateral trade with a possible actual origin or destination and a test country's bilateral trade with possible actual origin or destination at the 6-digit commodity level and examine whether there are common asymmetric commodity groups with discrepancies in different directions.

2.3 Assess Commodity Classifications

The assessment of commodity misclassification by the multiple mirror technique involve two steps:

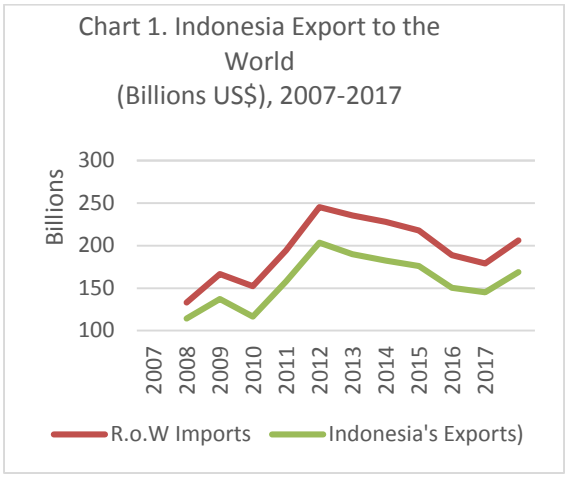
1. Compare the discrepancy between a test country's export or imports to or from the rest of the world and the rest of the world's imports or exports from or to a test country at the commodity level. Then, identify Asymmetric commodity groups with discrepancies in opposing directions (negative vs positive)

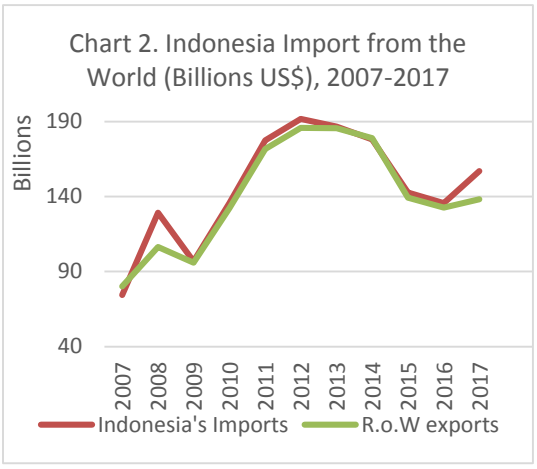
Examine if all major trade partners simultaneously classify the concerned traded goods in different manner than a test country.

3. Results

3.1 Indonesia Trade to the Rest of the World

The discrepancy of export of Indonesia to the world and conversely showed in the Chart 1. The ratio lies between 1.17 to 1.25. This figures depicted that the data of total export of Indonesia is still in "acceptable range" (Hamanaka, 2001).





Source: BPS-Statistic Indonesia and UNComtrade, processed

Furthermore, if we look at the ratio of Indonesia import of goods from the rest of the world to export of the world to Indonesia, it shows much better figure (Chart 2). The ratio lies between 0.93 to 1.21. It proves that the quality of Indonesia import data is much better than export, as mentioned in the introduction.

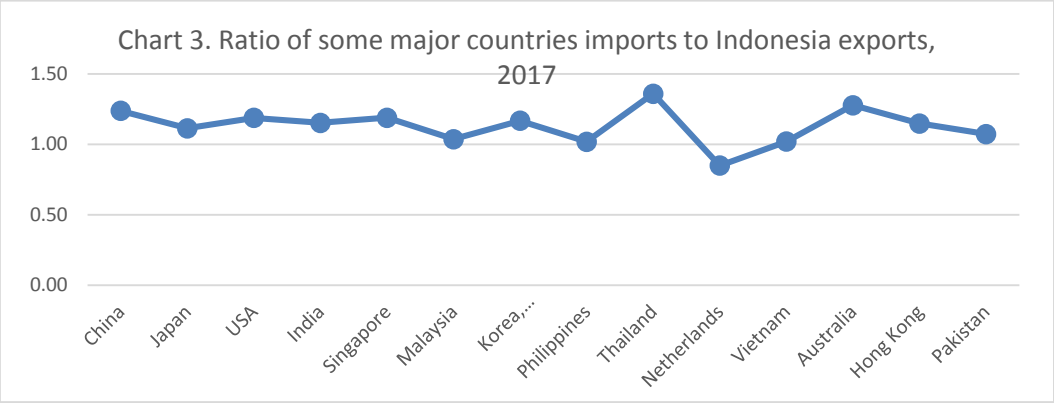
Although in overall the quality of Indonesia's export and import are in acceptable range, misclassification seem to be common within commodity groups. The number of asymmetric commodity group for export data with high discrepancy tend to be less. In 2008 there were 22 commodity groups and in 2017 there were only 13 commodity with high discrepancy.

Some commodities exports were consistently showed smaller number than the rest of the world imports from Indonesia. For example, HS 02 (Meat and Edible Meat Offal) and HS 04 (Dairy Products, Bird's Eggs; Natural Honey, Edible Products of Animal Origin, not elsewhere specified or included). On the other side, HS 21 Miscellaneous Edible Preparations was consistently gave larger number than import of the rest of the world to Indonesia.

As export data, the number of asymmetric commodity groups for Indonesia import data with high discrepancy tend to be less too. In 2008 there were 31 commodity groups and in 2017 there were only 17 commodity. The example of commodity that were consistently higher than export of the rest of the world to Indonesia was HS 52 (Cotton, and commodity), and commodity group that consistently has smaller number than export of the rest of the world was HS 60 (Knitted or Crochetted Fabrics).

3.2 Indonesia Trade to Major Partner Countries

To get a deep picture of Indonesia trade to major partner countries, the mirror analysis was conducted for the year 2017 as shown in Chart 3 and 4.



Source: BPS-Statistic Indonesia and UNComtrade, processed
Note: r.o.w is the rest of the world

From Indonesia export perspective, the significant ratio that show high discrepancy are China, Thailand, Australia, and Netherland (Chart 3). Meanwhile, on the other side (Indonesia import), the high discrepancy are Australia, USA, Hong Kong, France, and Singapore (Chart 4.)

Source: BPS-Statistic Indonesia and UNComtrade, processed

Import of China from Indonesia was consistently higher reported than export Indonesia to China since 2008 until 2017. The possibility of this discrepancy other than the valuation is the availability of consignment country such as Singapore. However we need more examine in this case, and for commodities exported.

Then, for detailed HS two digits, the commodity group that showed high discrepancy (import China is reported higher than export Indonesia) is group 85 (Electrical machinery and equipment and parts thereof; sound recorders/reproducers, television image and sound recorders and reproducers and parts and accessories of such articles), and 84 (Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof). These two group of commodity consist of many kinds of commodities. Misclassification of these

group very much happen. The other commodities is HS 27 (Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes).

The discrepancy of these commodities was also due to the consignment country. When we see the data of Singapore export for three commodity groups (HS 84, 85, and 27) to China, it tended to be reported lower. The trade flow may be as follows. Indonesia recorded export of the three commodities as Indonesia export to Singapore, then Singapore export these commodities to China and recorded as Singapore export to China. China was possible to record these commodities as import from Indonesia.

The misclassification of commodities will be seen more clearly when viewed according to HS in more detail such as 6 digits. Still with the same test country, namely China, it was found that for HS 08 group (Edible fruits and nuts; peel of citrus fruit or melons), there were possible differences in item classification. In 2017, import China from Indonesia for HS 080112 (Nuts, edible; coconuts, in the inner shell (endocarp)) was not significant export Indonesia to China for this HS. Whereas Indonesian exports for HS 080119 (Nuts, edible; coconuts, fresh or dried, other than desiccated or in the inner shell (endocarp)) are quite large with values approaching Chinese imports for HS 080112

As China, import of Thailand and Australia from Indonesia were consistently reported higher than export from Indonesia to Thailand and Australia. Commodities that always reported higher from Thailand is HS 33 (Essential oils and resinoids; perfumery, cosmetic or toilet preparation) and HS 27 (Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes). Meanwhile, for Australia, the commodities are HS 27 (Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes) and 84 (Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof).

Import Netherland from Indonesia was always reported lower than export Indonesia to Netherland. Since 2008 the ratio of import-export consistent under 0.85. It showed the possibility of country of consignment between Indonesia and Netherland. The commodities were unloaded in other countries before brought to Netherland. The possible consignment countries between Indonesia and Netherland were Germany and France, but it need furthermore examination.

On the other hand, Indonesia import was reported higher than United States and Australia export to Indonesia. Whereas. Indonesia import was reported lower than Singapore, Hong Kong, and France export value to Indonesia.

Like other countries, there are some possibilities of commodities misclassification. As an example misclassification possibility between HS 10

(Cereals) and HS 11 (Product of the milling industry, malt, starches, inulin, and wheat gluten) in Australia. HS 10 was consistently had positive discrepancy and HS 11 was consistently had negative discrepancy since 2008.

Singapore, Hong Kong, and France were also possible to be the consignment country for Indonesia. The import-export ratio of these three countries consistent to be negative. Hong Kong is possible to be consignment country between Indonesia and China, then France was possible to be consignment country between Indonesia and European countries.

4. Discussion and Conclusion

Discussion

The quality of Indonesia import data were better than its export data. Since 2008 almost all import-export ratio on import data side below 1.1. Meanwhile, the import-export ratio on export side more than 1.1. The quality of import data in Indonesia is better than export data because it is related to government revenue, so import document more verified than export document. Furthermore Indonesia Government policies is always to encourage exports in order to get more foreign exchange. However this policy has a trade of with minimum verification of export documents by Customs Office due to evading the long process of export.

The discrepancy in international trade is unavoidable. The basic reason for the cause of discrepancy is the valuation. As IMS recommendation, export is valuated in Free on Board (FOB) and import is valuated in Cost of Insurance and Freight (CIF). The discrepancy ratio for export majority is above 1 or tend to have positive discrepancy.

The next reason for discrepancy in Indonesia is misclassification in commodity groups. Although in total discrepancies are low but when detailed by country or commodity group it will be seen that the discrepancies are quite varied. Although, there is a misclassification on both of export and import side but it is difficult to determine which country makes a mistake. At least the result of mirror analysis can be used to improve the quality of Indonesia export-import data.

The third reason for discrepancy in Indonesia is the existence of country of consignment or transhipment country. Indonesia export was not recorded as import in destination country from Indonesia but was recorded as import of destination country from third country (consignment country). Otherwise, Indonesia import was not recorded as export of country of origin to Indonesia but was recorded as export of country of origin to the third country (country of consignment).

Conclusion

1. The discrepancy of export and import data could not be avoided due to some reasons, namely valuation, misclassification, and consignment countries.
2. Indonesia data export and import can be categorized in good quality (the discrepancy in acceptance range)
3. To reduce the discrepancy, all countries should have the same understanding and implementation of IMTS manual, and can do bilateral mirror analysis,
4. Custom office rule is very important in assessing the declaration, so the misclassification can be reduced.

References

1. Carrere, Celine, and Christopher Grigoriou. (2015). Can Mirror Data Help to Capture Informal International Trade? Working Paper. Foundation Pour Les Etudes Et Reseachers Sur Le Developpement International
2. Day, Iris. (2015). Assessing China's Merchandise Trade Data Using Mirror Statistics. Bulletin. Reserve Bank of Australia.
3. Ferrantino, Michael J., and Zhi Wang. (2008). Accounting for Discrepancies in Bilateral Trade: The Case of China, Hong Kong, and The United States. China Economic Review, 19 (3), pp 502-520
4. Hamanaka, Shintaro. (2011). Utilizing the Multiple Technique to Assess the Quality of Cambodian Trade Statistics, ADB Working Paper Series on Regional Economic Integration, No.88
5. Javorsek, Marko. (2016). Asymmetries in International Merchandise Trade Statistics: A case study of selected countries in Asia-Pacific. Working Paper Series. United Nations-ESCAP

Trade imbalances and trade asymmetries: Two sides of a complex relation

Ronald W. Jansen
United Nations Statistics Division/DESA

Abstract

Bilateral trade imbalances are among the main causes for trade disputes. Certain countries will argue that they are taken advantage of, because they are importing much more from their trading partner than vice versa; and that consequently the trading partner would be expected to make efforts to even out that imbalance. Raising tariffs would be one strategy of reducing the imbalance. However, the trade statistics, which form the basis for the trade discussion, may not properly reflect the contributions of the negotiating countries in the production of a final product. With the occurrence of global production networks, the value added of traded goods can actually be decomposed in value added of parts (and services) delivered by many partner countries in the network. Globalization has made trade complex, and merchandise trade statistics do not reflect that complexity. Besides the issue of trade imbalances, many disputes take place also over trade asymmetries. Trade imbalances are not the same as trade asymmetries. Whereas bilateral trade imbalances compare imports of country A from country B with imports of country B from country A, bilateral trade asymmetries compare imports of country A from country B (reported by A) with exports of country B from country A (reported by country B). Implicitly the assumption is often made that imports A → B should equal exports B → A. There are many valid reasons, though, why trade is not symmetrical. A better understanding of the causes of trade asymmetries is necessary to correctly reduce and reconcile bilateral asymmetries. If countries cannot agree on the actual size of their bilateral imports and exports statistics, then negotiations on trade imbalances would have no meaning. Therefore, reconciliation exercises between countries need to be conducted on a regular basis to agree on the bilateral trade statistics. Due to global production networks, the trade statistics need to be decomposed in domestic value added and foreign value added. In the ideal situation, the contributions of multinational enterprises and their foreign affiliates would be properly understood in terms of economic importance in each of the countries involved in the global value chain. This paper will give examples of bilateral trade imbalances and bilateral trade asymmetries, tries to shed some light on their relationship and provide explanation on the complexities of international trade and global production networks.

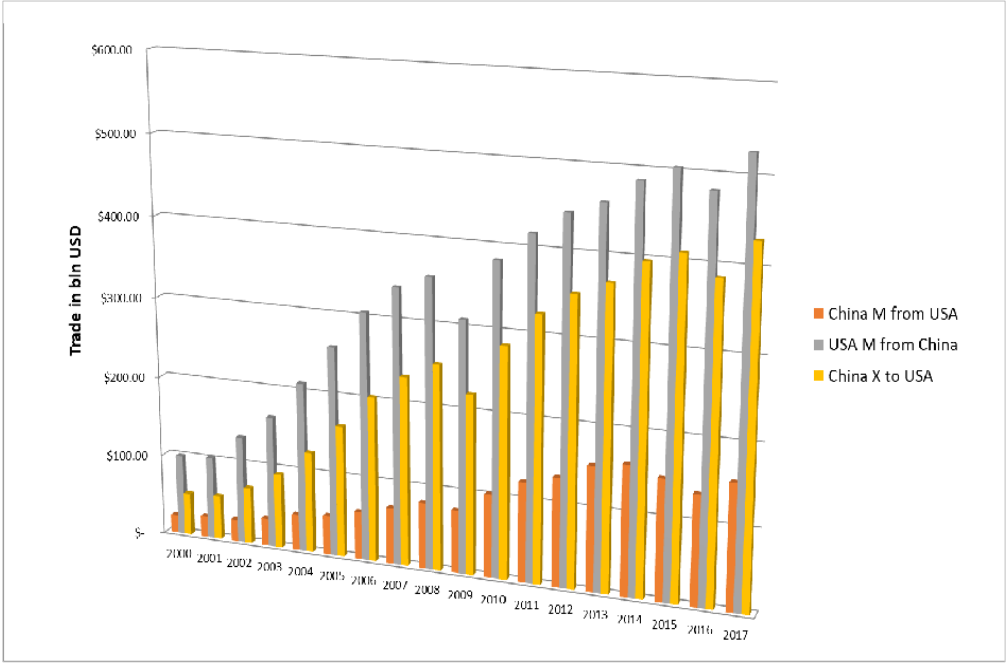
Keywords
Trade imbalances; trade asymmetries; trade in value added; global value chains; foreign affiliates

1. Introduction

Bilateral trade imbalances are among the main causes for trade disputes. For example, the United States (US) has argued for many years that it is unfair that the US is importing far more goods from China than the other way around. It has recently taken action by imposing tariffs on 250 billion (bln) USD worth of Chinese imports and China has retaliated, raising tariffs on US exports[1]. The interpretation of trade imbalances and therefore the justification for imposing tariffs is not as straight forward as it may seem at first glance. This paper will look into trade imbalances based on international merchandise trade statistics, and then 'look under the hood' to inspect what these trade statistics actually represent both in terms of its recording and in terms of its composition.

Bilateral trade imbalances compare imports of country A from country B with imports of country B from country A. Figure 1 shows the US-China bilateral trade statistics from 2000 till 2017.

Figure 1. US-China Bilateral Trade 2000-2017



[1] For a thorough discussion of this matter see a recent paper by Meltzer and Shenai on the US-China economic relationship, https://www.brookings.edu/wp-content/uploads/2019/02/us_china_economic_relationship.pdf

More specifically, Figure 1 shows that for each of these 18 years the US imported many more goods from China, than China imported from the US. In 2017, US imported over 500 bln USD worth of goods from China, whereas China only imported for about 150 bln USD from the US. Do these differences also mean that China is economically benefiting in those amounts from the trade with the US? The answer to that question depends on how much of the production of the traded goods originates (or "sticks") in China. Global production networks have unbundled[2] the production process, where parts and services are now delivered by companies in many countries. Before getting back to the value added of trade, we need a closer look at the reported trade statistics themselves.

In a bilateral discussion on trade, both parties bring their trade statistics to the table. This means that an agreement needs to be reached in cases where these bilateral trade statistics do not coincide. ***Bilateral trade asymmetries*** compare imports of country A from country B (reported by A) with exports of country B from country A (reported by country B). On face value, it could be assumed that imports A → B should equal exports B → A; in other words, that imports and exports are two sides of the same transaction and therefore symmetrical. As logical as that may seem, the reality of trade statistics is different.

The merchandise trade statistics[3] are for a significant part based on Customs regulations as well as on WTO valuation rules. This has been a practical decision by the statistical community, since the recording of international transactions in goods is well regulated and internationally agreed by the Customs administrations. In this context, the Kyoto Convention[4] is the most important internationally agreed set of regulations and contains the Country of Origin as one of its cornerstones, which stipulates that the partner country on the import declaration is the country in which the good was predominantly produced. Adopting the Country of Origin as the partner country of imports has important statistical consequences. Goods may have passed through some intermediate countries (for a variety of reasons and actions) before arriving at the Customs administration, where the imports are recorded. Whereas Country of Origin can be traced by the importing country through the accompanying paperwork, the country of final destination indicated by the country, which originally exported the goods, may not know at the time where the goods would eventually end up.

---

[2] See Baldwin, R. (2016). *The Great Convergence: Information Technology and the New Globalization*. Cambridge, MA: Harvard Press.

[3] See https://unstats.un.org/unsd/trade/imts/methodology.asp

[4] See http://www.wcoomd.org/en/topics/facilitation/instrument-and-tools/conventions/pf_revised_kyoto_conv.aspx

2. The example of reconciling the US-Canada trade statistics

In 1987, the United States and Canada agreed[5] to a unique arrangement with exchange of micro-level trade information. The goal was simple: in order to align the large bilateral trade statistics, Canada would take the imports recorded by the United States (of goods coming from Canada) as its exports to the United States, and vice versa. This should guarantee by definition a full reconciliation of the bilateral trade statistics between the US and Canada. However, if we look at the trade figures of US and Canada in the United Nations Comtrade[6] database, we see some large discrepancies. For example, in 2014 the USA reported total exports to Canada in the amount of 312 billion USD, whereas Canada report imports from the USA for "only" 252 billion USD, which means a 60 billion USD difference. How is this possible?

Table 1. US-Canada reconciliation of trade statistics

Trade Flow	**Total Trade** (in bln USD)
US exports to Canada	**312**
Canada imports from US	252
Origin= US and Shipment <> US	(-2)
Origin <> NAFTA and Shipment = US	(+42)
Origin = Canada and Shipment = US	(+3)
Origin = Mexico and Shipment = US	(+11)
Valuation adjustment	(+7)
Adjusted Canada imports from US	**313**

Table 1 shows how the application of Country Origin leads to bilateral trade asymmetries in the reported imports and corresponding exports. Statistics Canada gave us the exact breakdown of adjustments made. The partner country in those transactions, in which the origin of the goods was not USA (the table makes a distinction between Mexican origin, Canadian origin and non-NAFTA origin), were adjusted to the correct partner country on the basis of the Country of Origin rule. After reconciling all adjustments made to the imports, the totals for imports (reported by Canada) and exports (reported by the US) are virtually identical. What this example clearly shows, is that the Country of Origin rule has a big impact. There are a few other methodological reasons for discrepancies, such as the valuation of the transactions and the trade system used by a country. In a reconciliation exercise, these

[5] See https://www.census.gov/foreign-trade/aip/uscanada.pdf

[6] See https://comtrade.un.org/

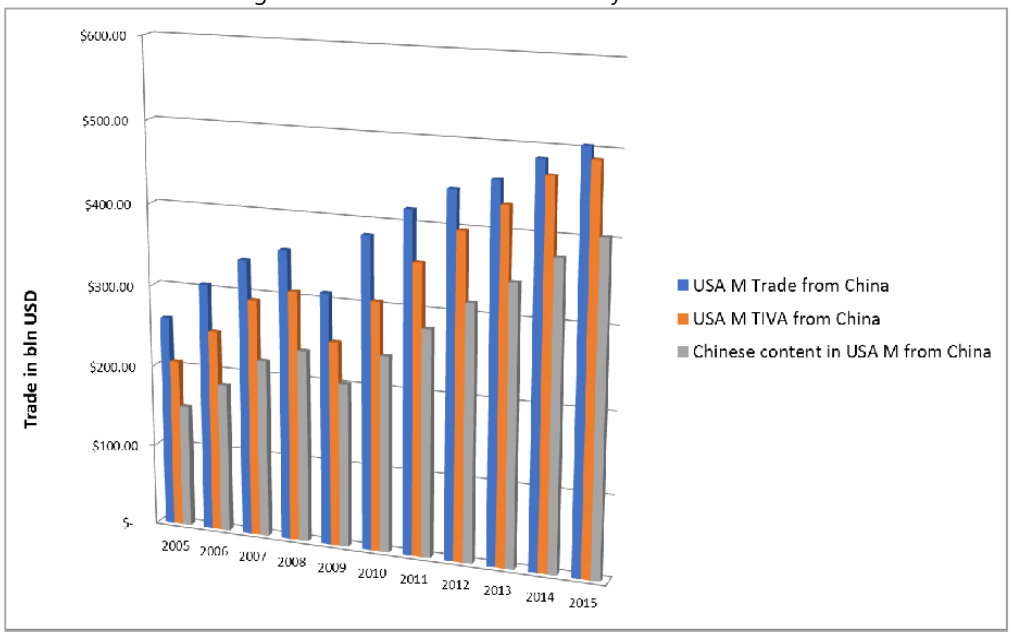
methodological discrepancies (leading to bilateral asymmetries) need to be resolved first before looking at the bilateral trade imbalances.

3. Trade in Value Added

One of the most basic challenges of globalization and the fragmentation of production is an increasing volume of double counting in the real and financial sector. In the real sector, traditional trade measures count gross flows of goods and services as exports and imports each time they cross international borders. As a result, basic raw materials and intermediate products made in one country are counted as exports when they are shipped to a second country to be used as inputs along with inputs from other countries in the assembly of these products into a finished product, which is in turn shipped a third country where it is subject to quality control, repackaging, distribution, and final sales.

Rather than only counting the value added by each country in each stage of the production process, or valuing only the final value of the goods – as is done to avoid double counting in gross domestic product (GDP) – the gross value of the export sales (and the imports) are double-counted, which results in a misleading picture of the economic contribution of countries to trade flows and the contribution of foreign value-added to domestic GDP. For instance, countries that may make only a small value-added contribution to the final value of a product from the final assembly of parts will have the entire value of the gross export counted, rather than the value added of the gross exports less intermediate inputs from other countries. These flows, therefore, do not reflect the value-added of the exporting country in the production of the goods or services.

Figure 2. US-China Bilateral Trade by Value Added



The OECD Trade in Value Added (TiVA) database [7] provides a comprehensive map of international transactions of goods and services in a massive dataset (based – among others – on reconciled trade statistics) that combines the national input-output tables of various countries at a given point of time. Moreover, such input-output analysis covers an entire set of industries that make up an economic system, thus enabling the measurement of cross-border value flows for a country or region, and so they provide scope to track the value-added generation process of every product in every country at every production stage.

Figure 2 shows US imports from China in three ways: (1) US imports from China in terms of gross merchandise trade statistics; (2) US imports from China reconciled (for bilateral trade asymmetries) in the Trade in Value Added (TiVA) database; and (3) the Chinese value added in US imports from China, as estimated in the TiVA database. For 2015, about 80% of the US imports from China was Chinese value added, meaning that 20% was foreign value added, including parts originating from the US itself. Linking back this result to the trade negotiations, the discussion should be held only on the Chinese content of the imported Chinese good by the US otherwise tariffs would affect not only China but also other countries, including US itself.

[7] See http://www.oecd.org/sti/ind/measuring-trade-in-value-added.htm

4. Accounting for Global Value Chains

At the request of the United Nations Statistical Commission, a group of experts (on business and trade statistics as well as on national accounts) prepared a handbook, which provides a measurement framework for international trade and economic globalization. The handbook[8], which was submitted to the Statistical Commission in March 2019, presents a global value chain (GVC) satellite account approach through a system of extended national accounts and integrated business statistics from the perspective of the national statistical system. More specifically, the handbook provides a national perspective on globalization on the basis of a global value chain model that describes the regionally-integrated decomposition of industry-specific global value chains in a multi-country supply chain of goods, services and institutional arrangements. Doing so allows for an integrated presentation of production, income, assets and liabilities by partner country for those GVC industries that play a significant role in the national economy, resulting in GVC-specific, multi-country supply and use tables (SUT) and related institutional sector accounts. This GVC approach can better inform public policies and business decisions on issues related to, for example, growth and productivity, domestic and foreign share of the value-added generated, trade policy, domestic and foreign labour and capital used in the production of goods and services.

A GVC consists of the full range of activities that firms and workers do to bring a product (good or service) from its conception to its end use and beyond. This range includes activities such as research and development, production, transportation and distribution, marketing and sales and after-sales services to the final consumer. Lead firms in GVCs initiate and coordinate the activities of the value-added chain. This first-mover status gives them "power in the chain" because they tender contracts, place orders and select suppliers. For national accounting purposes, as well as for measuring production and trade in GVCs, the lead firm is assigned to a national territory or country. The lead firm should be located where the ultimate decision-making authority is resident. The best proxy for this concept is probably the location where the board of directors and chief operating officer conduct their affairs.

A global enterprise can organize its core production activities (production of goods and services to be sold in the market) in a number of different business lines. Such an enterprise could be a lead firm for various GVCs in different specific industries. Therefore, business, trade and investment data for a GVC satellite account would need to be collected from the business line of

[8] See https://unstats.un.org/unsd/statcom/50th-session/documents/BG-Item3h-GVC%20Handbook-E.pdf

a global enterprise to allow for the correct data specification of the industry-specific GVCs controlled by the lead firm. The list of standardized products explicitly identified in the SUTs reflects the GVC-related products which include the final product of the GVC and the intermediate goods and services that are used to produce the final product. In the multi-country SUTs, the trade of these products between the GVC-partner countries must be explicitly shown and reconciled. In the case of three partner countries in a GVC, the integration of GVC information starts with the compilation in each country of a SUT with a breakdown of industries and products. Once the national SUTs are compiled in each of the GVC partner countries, they are integrated into a multi-partner country SUTs. The analysis of such GVC specific multi-partner SUTs will be a much better basis for negotiations about bilateral trade and other economic relations than only (even reconciled) bilateral trade statistics.

5. Globalization related bilateral trade asymmetries

Besides methodological differences causing bilateral trade asymmetries, there exists a more elusive kind of bilateral trade asymmetries caused by reporting practices of multi-nationals. A case in point was the manufacturing and exports of computer chips by foreign affiliates in Costa Rica. According to Costa Rica its exports of computer chips amounted (in the years before 2014) to about 2 billion USD each year. However, if adding up all the imports of these computer chips by the partner countries of Costa Rica, these corresponding imports amounted to almost 20 billion USD, which is 10 times as much. The most likely reason for this discrepancy was that the exporting companies (affiliates of a foreign multi-national) only added the factory price of the computer chips on the export declaration, whereas the importing country (for example, China) would enter the market value of these chips on the import declaration. The large bilateral difference (about 18 bln USD) is not accounted for in international trade statistics but could be recorded as a trade in services (Charges for the use of intellectual property) between the country of residence of the multi-national and the country importing the computer chips. The GVC-specific (in this case the computer chips manufacturing) multi-country (Costa Rica, China, US) supply and use tables could help disentangling the value chain, and consequently support trade negotiations which more accurately take account of the complex arrangements and revenue generation of multi-national enterprises.

6. Conclusion

Trade negotiations are politically important and sensitive, and they can lead to political actions with serious economic consequences. It is therefore critical that the trade numbers informing the negotiations are adequate. As outlined in this paper, certain adjustments are needed to the bilateral

merchandise trade statistics to make the data more in line with economic reality. First, the trade statistics need to be reconciled to eliminate bilateral asymmetries, which are mostly caused by methodological issues. Subsequently, it is advised to take Trade in Value Added estimates instead of the merchandise trade statistics, because negotiations should deal with domestic content of trade only and not with the foreign value added. Ultimately, it would be best if negotiations could take the GVC-specific multi-country satellite account as the basis for several industry-specific discussions, for example, regarding trade imbalances in the automotive, the pharmaceutical or the electronics industries. In such cases, the reality and economic significance of globalization through production networks could be better understood in its consequences for income generation, jobs, and environmental impact, and this approach would stand a higher chance to find mutual beneficial agreements.

References

1. Baldwin, R. (2016). *The Great Convergence: Information Technology and the New Globalization*. Cambridge, MA: Harvard Press
2. Meltzer, J. and Shenai N. "US-China economic relationship: A comprehensive approach", *Brookings Policy Brief*, February 2019
3. United Nations (2019) *Handbook on Accounting for Global Value Chains*, Forthcoming

The rise of China and the Malaysian electronics and electrical sector (A bilateral trade view)

Kok Onn Ting
Ministry of Economic Affairs

Abstract

After joining the WTO in 2001, China's total exports grew by 19.3% per annum up to 2013 and the country emerged as the world's biggest exporter of manufactured electronics. China's rise has had an impact on developing countries such as Malaysia, a major exporter of electronic and electrical (E&E) goods. Malaysia aims to be a high-income economy by 2020, and upgrading its E&E value chain is critical to this goal. Malaysia is part of the East Asian production network and China imports intermediate inputs from Malaysia's E&E for its final exports while simultaneously expanding in similar product spaces. This means the effect on Malaysia of China's rise is complex.

Contemporary literature divides the impact of China's rise into competitive and complementary effects, and this paper aims to further understand the effect of China's rise on Malaysia's E&E trade channels. As a preliminary analysis, the paper will examine in depth the bilateral E&E trade between Malaysia and China. Subsequently, the paper looks at Trade Complementary Index between Malaysia and China, compares that with Malaysia and other traditional partners, such as the US, EU and China.

The research finds the bilateral trade between Malaysia and China in the E&E sector has been positive for both countries. Malaysia fares better in parts and components, and specifically in the semiconductor industry. On the other hand, its household electrical industry faces the competitive force of China's rise as it increasingly imports consumer goods from China. After 2008 China stepped into product space previously occupied by Malaysian semiconductors, notwithstanding that Malaysia currently still sends more high-value semiconductor to China than it receives from it.

Keywords

Globalisation; global value chain; electronics; electrical; bilateral trade

1. Background

After joining the WTO in 2001, China's total exports grew by 19.3% per annum up to 2013 and the country emerged as the world's biggest exporter of manufactured electronics. China's rise has had an impact on developing countries such as Malaysia, a major exporter of electronic and electrical (E&E) goods (Lall and Albaladejo, 2004). Malaysia aims to be a high-income

economy by 2020, and upgrading its E&E value chain is critical to this goal. Malaysia is part of the East Asian production network (Athukorala and Kophaiboon, 2014) and China imports intermediate inputs from Malaysia's E&E for its final exports while simultaneously expanding in similar product spaces. This means the effect on Malaysia of China's rise is complex. (Winters and Yusuf, 2007)

2. Objective

Contemporary literature divides the impact of China's rise into competitive and complementary effects (Kaplinsky and Messner, 2008) and this paper aims to further understand the effect of China's rise on Malaysia's E&E trade channels. As a preliminary analysis, the paper will examine the bilateral E&E trade between Malaysia and China. Subsequently, the paper looks at Trade Complementary Index between Malaysia and China, compares that with Malaysia and other traditional partners, such as the US, EU and China.

3. Methodology

3.1 The Trade Complementarity Index

The TCI for Malaysia-China bilateral trade gives a sense of the complementarity between the two countries based on the similarity between the structure of one country's exports and the other's imports. TCI is calculated at total trade level and at E&E level for the 1992-2013 period.

The TCI is computed based on:

$$Cij = 1 - \sum (|Mik - Xij| \div 2)$$

where M*ik* is the import share of product *i* in total imports of country *k* and X*ij* is the export share of product *i* from total imports of country *j*
Source: Adapted from Ng and Yeats (2003)

The TCI produces results between 0 and 1, with 0 when the export of a product from one country is not imported at all by the other country and 1 where the import of country k exactly matches the export from country j. The TCI uses China's import and Malaysia's export data.

The TCI is a measure of complementarity (the counterfactual of competitive) between two trade partners, based on their import and export profiles. It is a forward-looking index, specifically looking at the potential of a trade agreement between two countries to see if the imports of one country match the other country's exports. For example if a country exports good *i* and good *i* is imported by a potential trade partner, a trade agreement is deemed potentially beneficial to both trade partners.

Data for E&E sector is based on 338 products at HS Code five digit level and the E&E parts and components level consists of 81 products of

the 338 products sourced from UNComtrade through Worldbank WITS website.

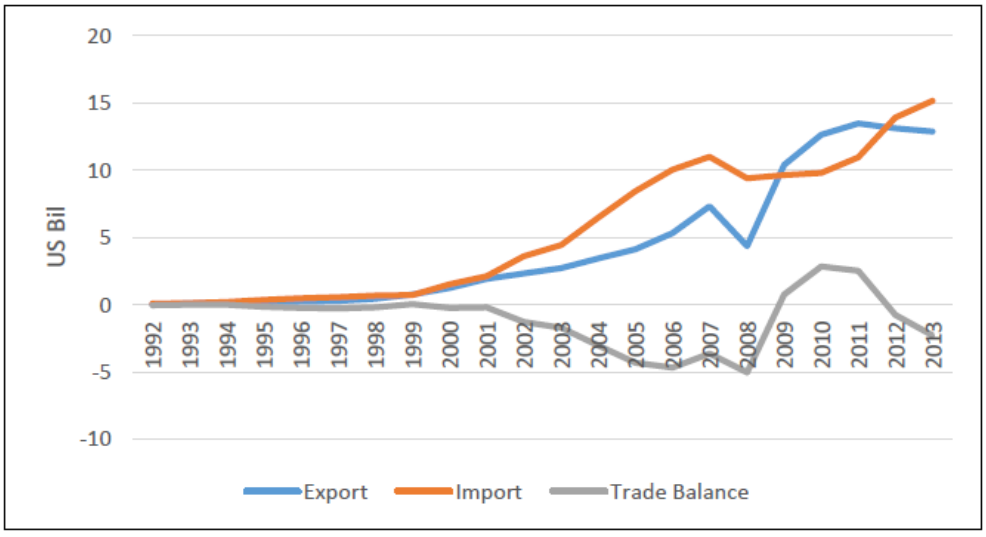
4. Findings

The research finds the bilateral trade between Malaysia and China in the E&E sector has been positive for both countries. Malaysia fares better in parts and components, and specifically in the semiconductor industry. On the other hand, its household electrical industry faces the competitive force of China's rise as it increasingly imports consumer goods from China. After 2008 China stepped into product space previously occupied by Malaysian semiconductors, notwithstanding that Malaysia currently still sends more high-value semiconductor to China than it receives from it. I conclude by discussing how far Malaysia and China's bilateral E&E trade structures support the complementary view of the regional production network.

While Malaysia and China both benefit from the rising volume of E&E trade over time, Malaysia experiences a trade deficit in most years. This deficit notably widens after China's entry into the WTO in 2001 until 2008, before swinging back into a Malaysian trade surplus in 2009-2011, aided by the rise in demand for ICs. Malaysia's E&E trade balance for had slipped back into a deficit by 2012 with E&E parts and components unable to offset imports of E&E goods in other categories.

At the aggregate level, Malaysia's exports and imports from China in their E&E bilateral trade increases over the years from 1992-2013, especially after 2000 (see Figure 1 below). E&E exports from Malaysia to China grew on average by 33.0% per annum from US$0.03 billion in 1992 to US$12.9 billion by 2013. As a share of total Malaysian exports to China, E&E rose from 4.2% in 1992 to 41.9% in 2013, while Malaysia's E&E imports from China rise from US$0.07 billion in 1992 to US$15.14 billion by 2013, with an annual growth rate of 29.2%. The share of total E&E imports from China to Malaysia is only 7.2% in 1992, rising to 45.1% by 2013. The high proportion of Malaysia's E&E exports and imports in its total trade with China confirms E&E as its most important manufacturing sector.

Figure 1 E&E Trade Balance (Malaysia as Reporter)



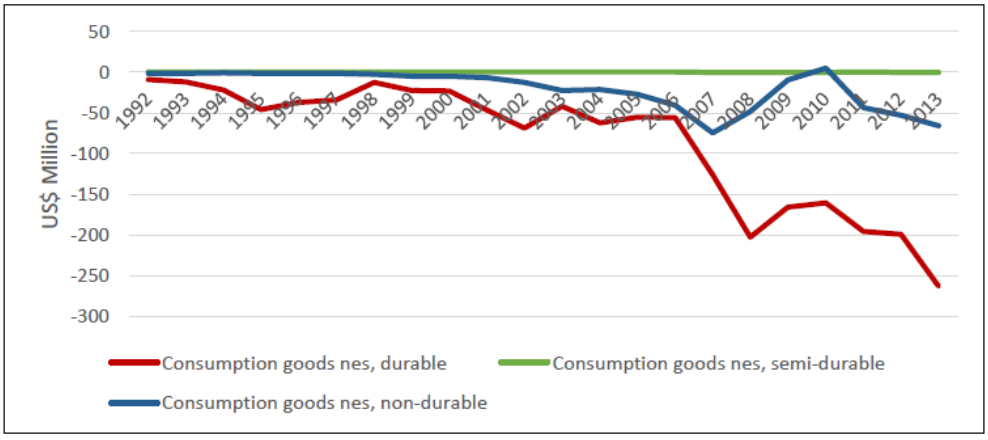
Note: Exports from China to Malaysia are include 327 products, while imports from China to Malaysia include 338 products: Malaysia does not export 11 of the 338 products to China.

Source: UNComtrade

While the regional production network literature, infers that China's neighbour such as Malaysia feeds parts and components into China for the assembly of final goods is broadly reflective of the complementary E&E trade, the analysis above finds that China has also stepped up its exports of parts and components to Malaysia, especially since 2008. However, Malaysia exports more sophisticated E&E (in total and as parts and components) while receiving less sophisticated imports from China.

A final note: although Malaysia's bilateral E&E trade with China is mostly in deficit (except in 1999 and 2009-2011), while it imports more from China, it processes these parts and components and ships them it to third destinations such as in the colour television manufacturers found in Malaysia. The import of Chinese finished goods (durable consumption goods) into Malaysia's E&E sector in Figure 2 below shows the effect on trade balance in the household electrical sector.

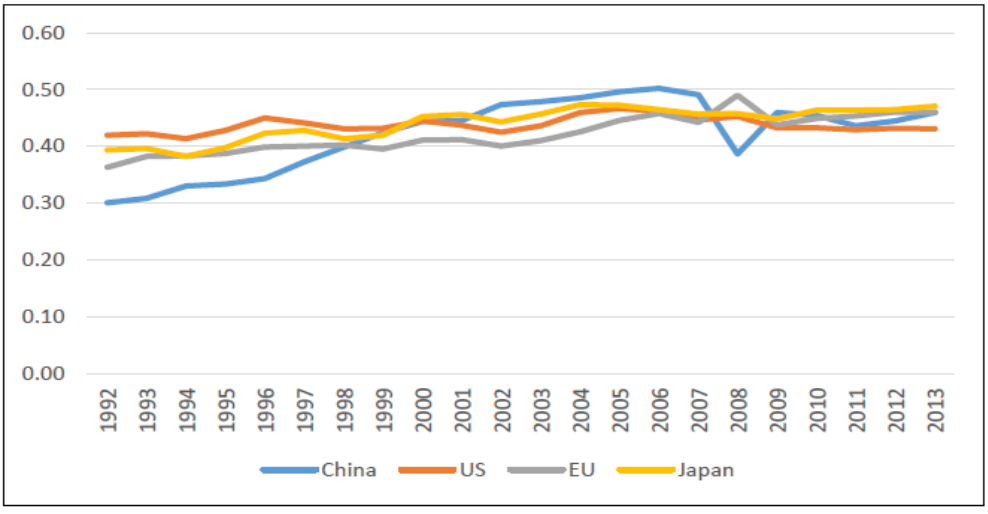
Figure 2 E&E Balance of Trade in Consumption Goods by BEC classification



Source: UNComtrade

As a form of forward-looking measure, this paper measures the extent to which China's imports are integrated with Malaysia's exports compared to other major destination markets. The TCI gauges the complementarity of one country's trade with another relative to other countries by measuring how closely one country's exports match the other's imports. A close match with a high TCI score means that both countries have a better prospect for regional economic grouping and integration compared to where their trade structure is less compatible (Michaely, 1996). The TCI is calculated here for total trade and at the E&E level from 1992-2013, using 4-digit HS 1988/92 nomenclature.

Figure 3 TCI Malaysia's Total Exports to Selected Countries



Source: Based on UNComtrade

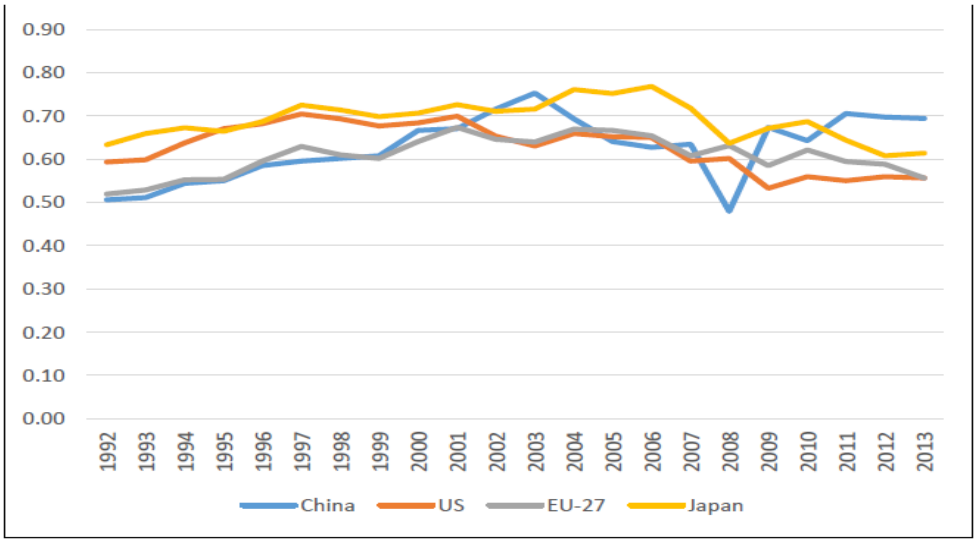


Figure 4 TCI Malaysia's E&E exports with Selected Countries' E&E Imports
Source: Based on UNComtrade

The TCI reveals that Chinese imports are more closely matched with Malaysian exports in 2013 than they were in 1992 at both total export and E&E level. In Figure 3 and Figure 4, the TCI rises from 0.30 in 1992 to 0.46 for total exports in 2012 and from 0.51 in 1992 to 0.69 for E&E exports in 2013. At total exports level in Figure 3, China climbs from least compatible import market among the major markets for Malaysian exports in 1992 to the second most closely-matched import market in 2013 at 0.46, after Japan at 0.47.

For E&E exports, China rises from the least-matched import market in 1992 to the most closely-matched, surpassing all other markets (Figure 4). The TCI for the E&E industries of various countries in 1992-2013 shows that Malaysia-China TCI is increasing overall compared to the US, EU and Japan, which rose during the first decade (1992-2001), The TCI shows that the US and EU peaked in 2001, and Japan peaked in 2006, before declining. For example, for the US, the TCI for E&E was 0.59 in 1992, peaking at 0.70 in 2001 but dropping to 0.56 in 2013. In short, the compatibility of Malaysia's E&E exports with the US, EU, and Japanese imports have decreased, compensated by the overall increase in E&E export compatibility with Chinese imports.

Contrasting Figure 3 and Figure 4, the differences between the trends in the TCI for total exports compared to E&E stands out. This is due to the divergence in 2001-2007. The TCI for China's total exports rose continuously from 2001-2007 before declining in 2008 while the TCI for the E&E sector begin to descend in 2004-2008 before recovering and ascending again from 2009 onwards. The compatibility of China's imports with Malaysia's exports therefore diverges between total exports and the E&E sector.

E&E exports from Malaysia become less compatible with Chinese imports starting from 2004 onwards. This may be because Malaysia is withdrawing, resulting in the share of individual E&E products dropping as a share of total E&E exports, or China is importing less E&E products, or both. A preliminary check shows a decline in the computer goods and electrical sub-sectors in Malaysian exports as a major factor in the decline in the TCI from 2004 onwards. This is partly due to competition in the trade channel between Malaysia and China. Although the TCI for E&E with China recovers from 2009 onwards, its recovery is based on a surge in import demand, mainly for *Monolithic Digital Circuits* (HS8542) from China.

Overall, the TCI findings are consistent with the shifting importance of Malaysia's trading partners over time. The TCI for E&E also shows that Chinese imports are the closest match to Malaysia's exports from 2011 onwards with the US, EU and Japanese match to Malaysia's exports declining. Malaysia would be concern that the Malaysia-China TCI for E&E also notably falls from 2004-2008 but recovers again from 2009 onwards. Therefore while Chinese imports of Malaysia's E&E exports were more integrated in the immediate years China joins WTO in 2001, compatibility between Chinese imports and Malaysian exports in finished goods sub-sectors such as the computer industry and household electrical industry declined during 2004-2008, and this is captured by the TCI.[1]

5. Conclusion

This paper has shown that China has both positive as well as negative effects on Malaysia through their bilateral trade. Their total bilateral trade volume rose rapidly in 1992-2013. China is indeed Malaysia's number-one trading partner, and trade data affirms the E&E sector (in both HS code Chapter 84 and 85) as the top sector traded between Malaysia and China. The preliminary trade analysis lends credence to the observation by Athukorala and Kophaiboon (2014) that the East Asian regional production network is indeed benefiting countries such as Malaysia. While Malaysia's bilateral trade balance with China is reasonably balanced, China is increasingly exporting consumer goods to Malaysia, especially since joining the WTO in 2001.

While the bilateral trade of E&E findings is consistent with East Asian production network literature (Athukorala and Kophaiboon, 2014), some of the results reported in this chapter also diverge. The divergence includes, first, the fact that Malaysia's E&E trade balance widens notably after China joined

[1] Computer manufacturer Dell shifted its desktop PC manufacturing from Penang to Xiamen, China. Top three items measured by decline in TCI points in 2004-2007 are *Automatic data processing machines* (computers) *(HS8471),* P*arts, accessories, except covers, for office machine (HS8473)*, and *Radio and TV transmitters, television cameras (HS8525)*. The period 2004-2007 is taken rather than 2004-2008 because 2008 saw the beginning of the global financial crisis.

the WTO in 2001, with the volume of parts and components exported unable to offset the deficit in total E&E level due to the steady increase in imports of E&E final goods, and the E&E trade balance slips back into deficit from 2012 onwards after a brief interval with a trade surplus in 2009¬2011. These patterns are also reflected in the TCI index at E&E level, where China is currently the most 'compatible' export-import structure with Malaysia. Nevertheless, the deficit in E&E parts and components in the years immediately after China joined the WTO can reflect global production arrangements, with China producing less sophisticated parts and components and Malaysia importing them for production purposes.

References

1. Athukorala, P.-c. & Kophaiboon, A. 2014. Global Production Sharing, Trade Patterns, and Industrialization in Southeast Asia. *In:* Coxhead, I. (ed.) *Routledge Handbook of Southeast Asian Economics.* Hoboken: Taylor and Francis.
2. Kaplinsky, R. & Messner, D. 2008. Introduction: the impact of Asian drivers on the developing world. *World Development,* 36**,** 197-209.
3. Lall, S. & Albaladejo, M. 2004. China's competitive performance: A threat to East Asian manufactured exports? *World Development,* 32**,** 1441-1466.
4. Michaely, M. 1996. Trade preferential agreements in Latin America: an ex ante assessment. *Policy Research Working Paper 1583, Washington D.C.: The World Bank*.
5. Ng, F. & Yeats, A.J. 2003. Major trade trends in East Asia: what are their implications for regional cooperation and growth? *World Bank Policy Research Working Paper*.
6. Winters, L.A. & Yusuf, S. 2007. *Dancing with giants: China, India, and the global economy*, World Bank Publications.

Quantifying China's involvement and participation in global value chains

Kristina Baris[1], Donald Jay Bertulfo[1,2], Paul Neilmer Feliciano [1], Janine Elora Lazatin [1], Mahinthan Joseph Mariasingham [1]
[1]Asian Development Bank
[2]Ateneo de Manila University

Abstract
In the East Asian miracle, the rapid growth of exports is seen as providing the key demand impetus to set in motion a cumulative process of high investment, profits, savings and growth. Leading this growth is the People's Republic of China. In a highly globalized environment, however, the fragmentation of production processes across geographical borders calls for a reevaluation of countries' contribution to global production using the value added perspective. Using recent theoretical frameworks in global value chain (GVC) analysis and utilizing the Asian Development Bank's multiregional input-output tables, we measure PRC's participation in GVCs and its role in cross-border production sharing worldwide. Our analysis covers the period 2000 and 2007-2017, a period characterized by PRC's structural shift towards stronger domestic chains supporting domestic demand, a higher value added share involved in the medium-to-high technology manufacturing global value chains, and a higher value added share contribution to other countries' exports. Our analysis documents PRC's gradual rise towards becoming a formidable force in global value chains, contributing extensively to the production of goods and services finalized outside the PRC.

Keywords
China; input-output; global value chains; TiVA

1. Introduction

In the East Asian Miracle, the rapid growth of exports is seen as providing the key demand impetus to set in motion a cumulative process of high investment, profits, savings, and growth. Leading this growth is the People's Republic of China (PRC). The PRC is an exemplar of a country that has gained from trade liberalization, with its growth rising along with its ascent towards becoming a manufacturing hub and goods exporter. The PRC faced restricted trade before its accession to World Trade Organization in 2001. It experienced rapid growth after its accession into the WTO, partly due to its rapid integration into global value chains. The PRC has made substantial gains by liberalizing its own market, sourcing and supplying goods through GVCs. In

2010, the PRC surpassed Japan as the second largest economy in the world. In 2014, the PRC became the world's largest trading nation.

PRC's economic success buoyed by its exports. However, in a highly globalized world where production processes are fragmented across geographical borders, the status of exports as a reliable indicator of wealth is contested. This argument is supported by recent theoretical and empirical advances in gross trade accounting (Johnson and Noguera 2012; Koopman, Wang and Wei 2014; Wang, Wei and Zhu 2013).

We elaborate on this point by taking a concrete finished product, say an iPod finalized in the United States. Parts assembled to produce this iPod come from different countries in the world. Materials used to produce the parts assembled to produce this finished product is a different story. Inputs used to produce the materials utilized to create these parts also constitute a separate network of production processes, which can occur in multiple parts of the globe. However, when an iPod is exported from the United States, traditional gross trade accounting indicates that its total value should be counted as exports by the United States. This is not necessarily true because, as demonstrated, the value of this finished product is an agglomeration of value added originating from different industries in different countries. Alternatively, in the creation of final products, value added is generated across and between production chains that are dispersed throughout the world. Among the rich topics in global value chain analysis (GVC) concerns the decomposition of gross exports into value added terms. Proponents contend that such an accounting exercise properly attributes wealth creation in a global trading environment characterized by complex networks (Koopman, Wang and Wei 2014; Wang, Wei and Zhu 2013).

This paper contributes to existing literature by utilizing value added-based indicators in order to gain a richer understanding of PRC's role in cross-border production sharing. The multiregional input-output tables compiled by the Asian Development Bank (ADB) are used to generate these indicators. Using a multiregional approach to analyzing GVCs provides two key advantages. First, it allows us to track the flow of value added from source to destination at a remarkable level of granularity[1]. Second, it adheres to the basic principles of national income accounting.

2. Methodology

[1] In particular, it is able to generate flow estimates at the 35-sector level, consistent with the ISIC 3.1 classification, for 62 countries worldwide, plus one capturing the rest of the world.

In this study, we follow Timmer et al. (2014) and define a global value chain (GVC) as a country-industry pair that delivers a good to its final use. Here, production is not entirely localized, but is instead fragmented across countries and industries worldwide. Many GVCs contribute to the production of any given final good in a certain focal GVC. In turn, value is created in each stage of the process. In order to systematically account for value added contributions across GVCs, we utilize decomposition frameworks from the GVC literature. The data underlying these theoretical frameworks are multiregional input-output tables (MRIOTs) which are compiled and balanced nationally and then internationally in a manner that is consistent with international classification standards. For purposes of the analysis, we enumerate the measures generated and utilized based on GVC literature. Interested readers may are encouraged to refer to the theoretical underpinnings of the indicators

(a) Decomposition of gross exports (Wang, Wei and Zhu, 2013): Wang, Wei and Zhu (2013) proposed a framework which traces the flow of trade from one GVC into another. The entire value of gross (i.e., intermediate and final) exports is decomposed into several parts, including the domestic value added embedded in gross exports, value added embedded in gross exports that originate from foreign sources (or foreign value added embedded in gross exports), and double counted terms.[2]

(b) Measures of participation in global value chains (Wang et al., 2017): GVC participation is defined from two interrelated perspectives: (1) forward and (2) backward. Forward GVC participation refers to the fraction of a GVC's value added that is involved in GVC activities while backward GVC participation refers to the fraction of a GVC's final goods production generated from GVC-oriented activities.

(c) GVC income (Timmer et al., 2013): GVC income is defined as the value added contribution of GVCs to the production of final goods. This measure concerns the proportion of the entire value of final production that is dependent on domestic and foreign sources.

3. Results

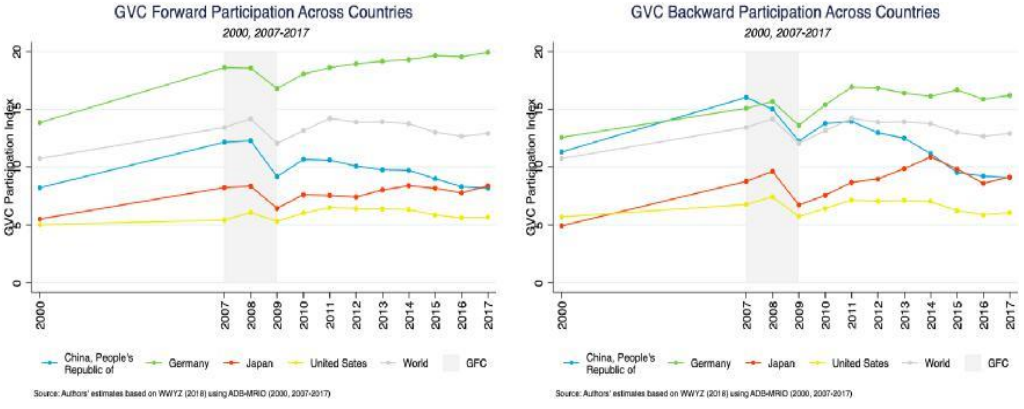
Overall, the PRC participates in GVCs moderately relative to other GVC hubs and the world average. A comparison of its forward and backward GVC participation indices suggests that the PRC is more inclined to receiving GVC-related value added as opposed to contributing value added that ultimately becomes involved in GVC-related activities.

[2] Double counted terms arise due to back and forth trade in intermediate goods.

Looking at forward participation, data shows that 8.2% of the PRC's total value added was generated through GVC-related activities in 2000. This was below Germany's 13.8% and the world average of 10.8%, but above Japan's 5.5% and the US's 5.0%.

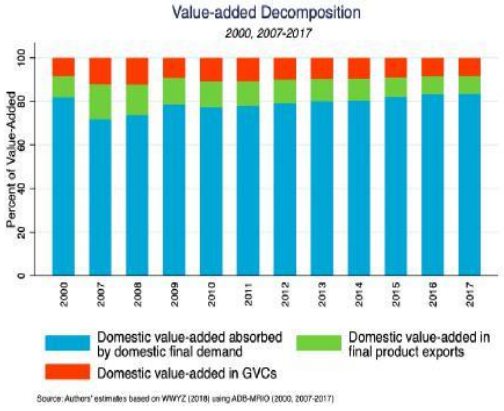
All hubs and the world showed troughs in forward GVC participation amidst the global financial crisis. After reaching peaks in 2008, forward GVC participation across the four hubs and the world dipped below their levels in 2000, before increasing again in 2010. However, the trends for the PRC and the world average are in contrast to other hubs in 2011. Germany, Japan, the US, and the world showed an overall increasing trend in forward GVC participation beginning in 2011, while the PRC's and the world average forward GVC participation indices have since been declining. By 2017, the PRC's forward GVC participation was at 8.2%, slightly below Japan's 8.4% and below its pre-accession level.

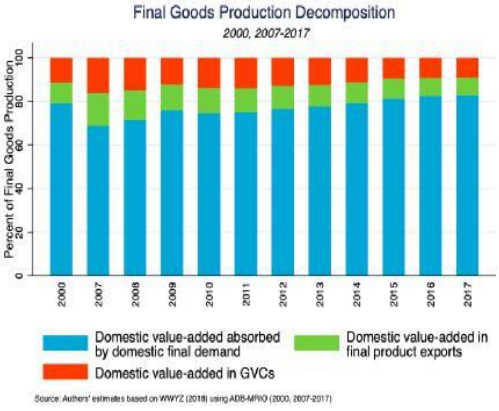
The same trend can be seen when examining backward GVC participation. Latest data shows that the PRC's backward GVC participation declined from its pre-accession level. In 2000, 11.3% of the PRC's production of final goods and services was represented by value added involved in GVC activities. This was below Germany's 12.6%, but above the world average of 10.7%, the US's 5.7%, and Japan's 4.9%. The downward trend in the PRC's backward GVC participation started in 2008, after it reached 16.0% in 2007. By 2017, the PRC's backward GVC participation was down to 9.1% and was surpassed by the world average of 12.9% and Japan's 9.2%



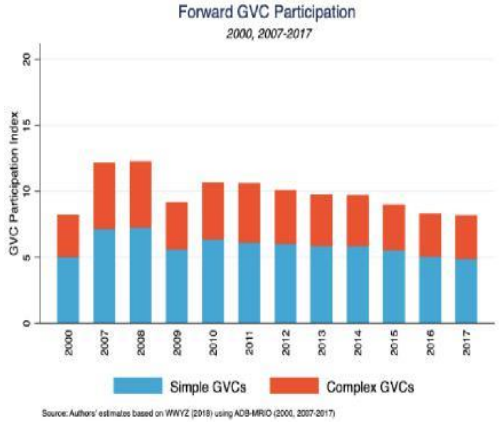
Thus, while domestic value-added has been increasing, the PRC's participation has been declining. This decline in GVC-related activities since 2011 is seen even at the global level. According to the Global Value Chain Development Report (2017), there has been a decline in GVC activities and an increase in the shares of pure domestic and traditional trade value-added creation at the global level since recovery in 2010/2011. Value-added decomposition and the decomposition of final goods production also show

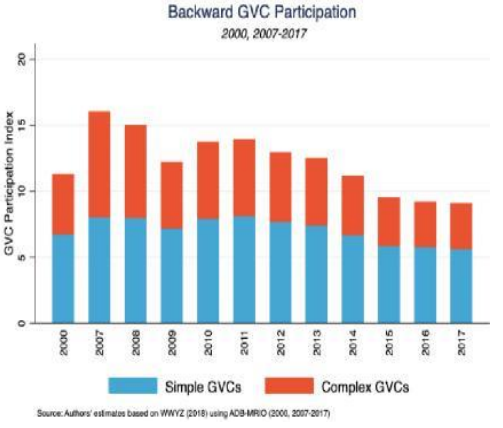
steadily increasing proportions of domestic value-added absorbed by domestic final demand in the PRC. One reason for this structural change in the PRC is its strengthening domestic value chain, which allows the PRC to substitute domestically produced intermediate inputs for imported intermediate inputs.





For the PRC as a whole, simple GVCs dominate GVC participation through both the forward and backward linkages. Thus, most GVC-related value-added that crosses China's borders is used by direct importing partners to produce final products for domestic consumption. Industry upgrading and a decrease in process manufacturing may be reasons for a higher simple GVCs index relative to complex GVC participation index. It can also be that domestic production length within the PRC has increased before and after border crossings as the PRC strengthens its domestic supply chains.

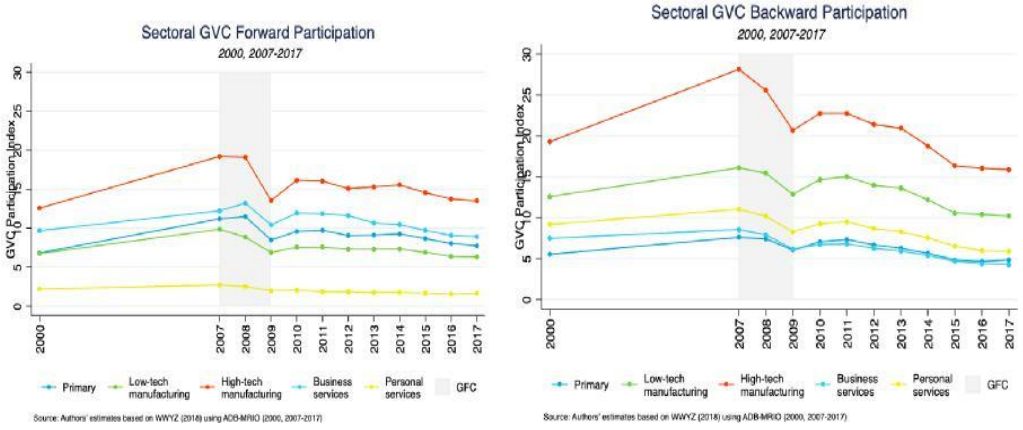




To analyze trends in PRC's GVC participation across major economic sectors, the 35 ADB MRIOT industries were classified into five broad categories: primary, low-technology manufacturing, medium-to high-

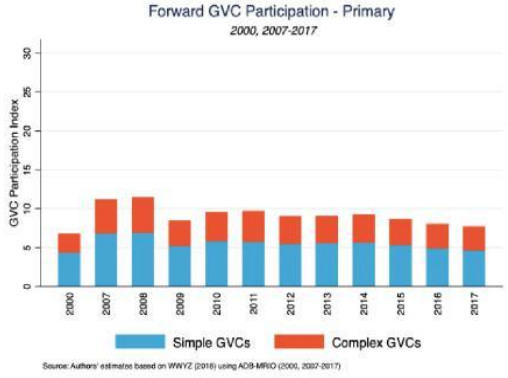
technology manufacturing, business services and personal services. Results show that PRC's backward GVC participation indices for low-technology manufacturing, medium-to-high technology manufacturing, and personal services are higher than their forward GVC participation indices. On the other hand, the PRC's primary and business services sectors have higher forward GVC participation indices compared to their backward GVC participation indices. Thus, these two sectors are more inclided to contributing value-added that ultimately becomes involved in GVC-related activities.

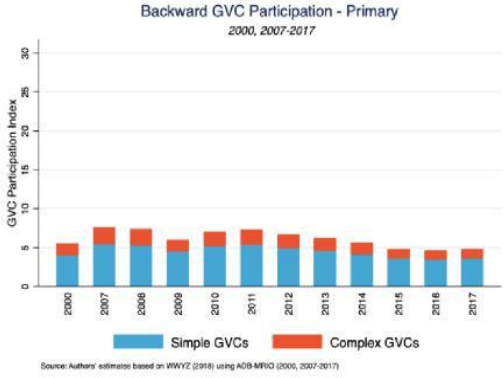
Findings also suggest that among the broad economic sectors considered, medium-to-high technology manufacturing exhibits the highest forward and backward GVC participation. Despite major technological advancements in the past decade, PRC's forward and backward participation in medium-to-high technology manufactures either remained stagnant or showed a declining trend. In 2000, 12.6% of total value-added generated by medium-to-high technology manufacturing industries contributed to GVC-related activities. This forward GVC participation ratio reached 19.2% in 2007 and has since declined. The ratio stood at 13.5% in 2017, reflecting a decline of about 6 percentage points from its value in 2007. Backward GVC participation showed a steeper decline. In 2000, final production of medium-to-high manufacturing industries contained GVC-related value added that represents 19.3% of its total value. After reaching a peak of 28.1% in 2007, the share went down to 15.9% in 2017, over 3 and 12 percentage points below levels in 2000 and 2017, respectively.

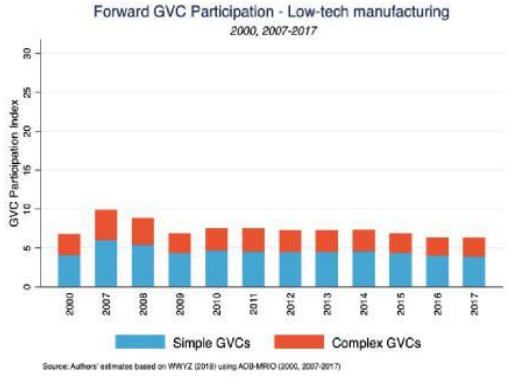


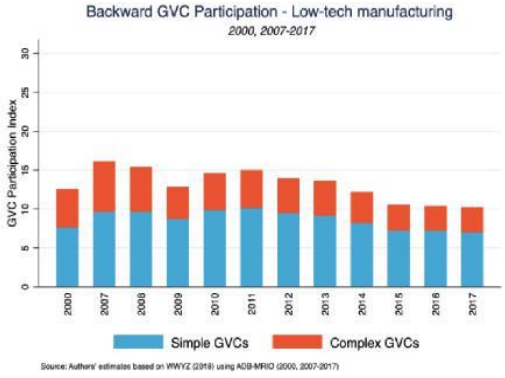
Breaking down GVC-related value-added into simple GVCs and complex GVCs, results show that simple GVCs dominate over complex GVCs in both the forward and backward linkages, except in the backward linkage of the medium- to high-technology manufacturing sector. In this sector, the proportion of value-added coming from complex GVCs is higher than that coming from simple GVCs. This implies that the PRC's medium- to high-technology manufacturing sector uses more intermediate imports that have

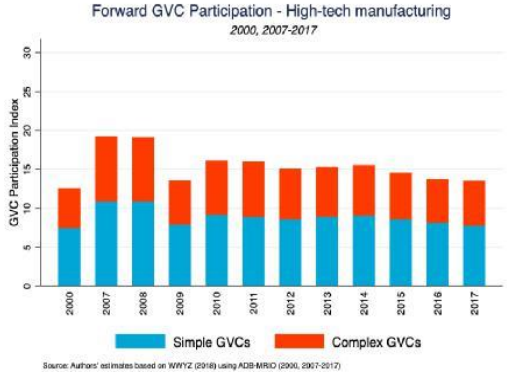
crossed borders at least twice to produce its final products for domestic use or exports. Given the higher share of simple GVCs in the forward linkage, this might imply that the PRC exports near-end medium- to high-technology manufacturing products.

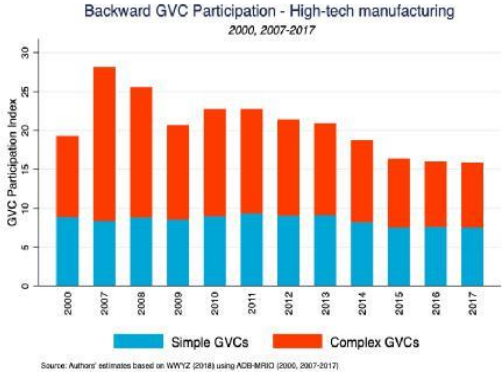


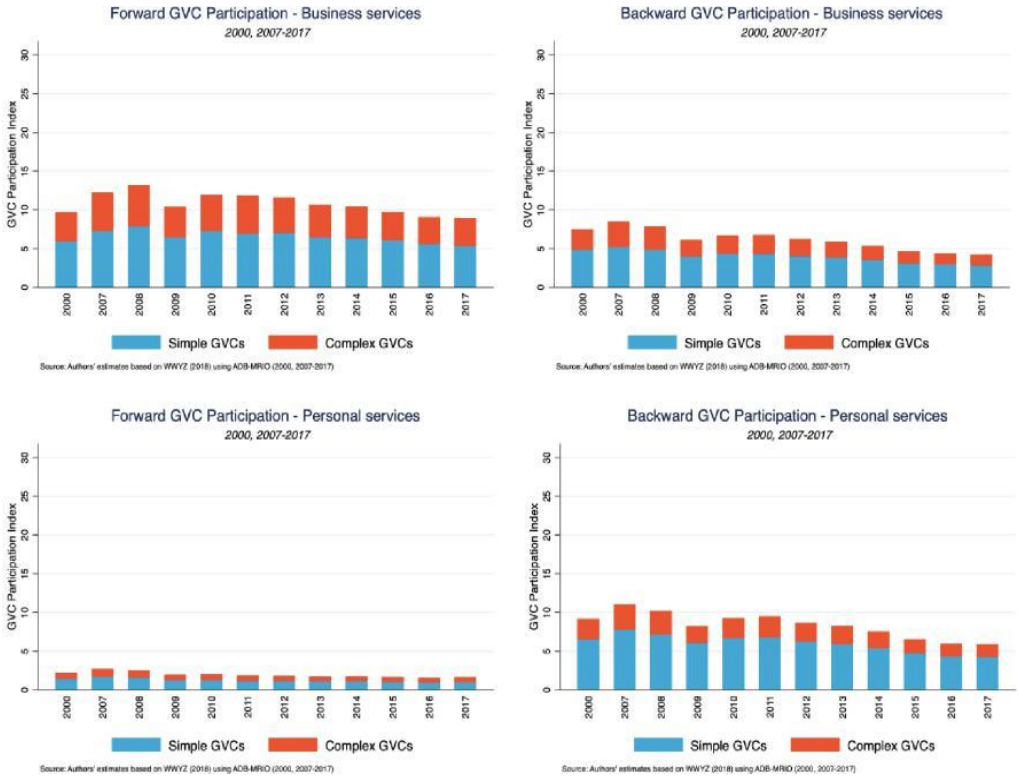












4. Discussion and Conclusion

Overall, we find that in the period covering 2000 and 2007-2017, the PRC has displayed a weak reliance on foreign imports in the production of its exports. Strong forward linkages, especially in exports of goods that require medium-to-high technology inputs, explain PRC's central position in global production networks. The share of domestic value added embodied in PRC's exports of intermediate products displays a rising trend, while its involvement in complex GVCs have strengthened in medium-to-high technology sectors. These suggest that the PRC may be moving up the value chain.

References

1. Johnson, R. & Noguera, G. (2012). Accounting for intermediates: Production sharing and trade in value added. *Journal of International Economics*, **86**(2), 224-36.
2. Koopman, R., Wang, Z. & Wei, S-J. (2014). Tracing value-added and double counting in gross exports. *American Economic Review*, **104**(2), 459-94.
3. Timmer, M., Erumban, A.A., Los, B., Stehrer, R. & de Vries, G.J. (2014). Slicing up global value chains. *The Journal of Economic Perspectives*, **28**(2), 99-118.

4. Timmer, M., Los, B., Stehrer, R. & de Vries, G. (2013). *Fragmentation, incomes and jobs: An analysis of European competitiveness*. Groningen: Groningen Growth and Development Centre.
5. Wang, Z., Wei, S-J. and Zhu, K. (2013). Quantifying international production sharing at the bilateral and sector levels. NBER Working Paper No. w19677. Cambridge: National Bureau of Economic Research.
6. Wang, Z., Wei, S-J., Yu, X. and Zhu, K. (2017). Measures of participation in global value chains and global business cycles. NBER Working Paper No. w23222. Cambridge: National Bureau of Economic Research

Determinants of Afghanistan's exports: A gravity model approach

Mohammad Salim Sadid
Faculty of Economics, South Asian University, New Delhi, India

Abstract

This paper studies the factors affecting Afghanistan's exports flow to its main trading partners. The panel data with the gravity model approach covering yearly data from period 2006 to 2016 is used for investigating the determinants of Afghanistan's export flows. The data includes export flows to 17 main trading partners. The random effect model is chosen as an efficient model for estimation of the gravity model. Results show that the importer's (GDP, openness rate, GDP deflator), and bilateral trade agreement have positive and significant effects on export. The distance is found to influence Afghanistan's export negatively. Furthermore, Afghanistan's GDP affects export positively; however, it's not statistically significant. Policymakers, using these results, will be able to take appropriate policy measure for improvement and expansion of export of the country.

Keywords

Export; panel data, gravity model; random effect model

1. Introduction

Afghanistan's foreign trade sector comprises a crucial part of its economy. According to the world bank (2018), the trade openness ratio decreased from 81.32 percent in 2010 to 47.65 percent in 2016-nevertheless, this sector has been suffering from a chronic deficit over the last years. Afghanistan's export sector accounts for only slightly more than 5 percent of the overall economy. Therefore, it has a relatively low contribution to the economy, and it cannot by itself create high employment or generate enough tax revenue to fix the government's shattered finances.

Afghanistan is in the early stages of export expansion. From 2008 to 2016 Afghanistan's exports increased slightly. In the sense that the decrease has been more in the period between 2008 to 2011 and then there is a slight increase from 2012 to 2016. Total exports in 2008 were around US$ 545 million and with a little increase in value to US$ 596 million by 2016[1]. The average growth rate of export has been approximately 2.1 percent annually. Afghanistan's exports are not widely diversified, however, dominated by a few

[1] According to NSIA (National Statistics and Information Authority) yearbooks of Afghanistan

agriculture products such as dried fruits, fresh fruits, and medical herbs (NSIA, 2017). From 2017 to 2018, dried fruits constituted 36 percent of exports, and that of fresh fruits constituted 15 percent of exports, and medical herbs 13 percent of the total exports of the country.

In terms of the direction of Afghanistan's exports, it mostly has been highly concentrated among a few major trading partners. During 2008-2016, about 70% of the total exports went out to the top two trading partners (Pakistan and India), and 30 percent to the other markets (Table 1).

Table 1: Afghanistan's export by major trading partners (share in export from 2008-2016)

no.	Country	2008	Country	2012	country	2016
1	Pakistan	48.50%	Pakistan	37.79%	Pakistan	47.54%
2	India	25.05%	India	22.18%	India	38.60%
3	Russia	6.74%	Iran	7.04%	Iran	3.16%
4	United Arab Emirate	3.44%	Turkey	6.87%	Turkey	2.04%
5	Germany	3.34%	Russia	4.81%	Iraq	1.88%
6	Turkey	3.28%	United Arab Emirate	4.24%	United Arab Emirate	1.57%

Source: NSIA yearbooks

Based on statistics from the NSIA yearbooks (2017-18), Afghanistan's total Gross Domestic Product excluding opium amounted to $20.2 billion in 2017-18 where GDP per capita amounted only $681. Therefore, exports represent an estimated 4.11% of total Afghanistan GDP, which seems a low share. Furthermore, the country is suffering from a chronic deficit in its balance of payment and needs to have export-led growth policy. Since Afghanistan economy is a transition economy and will be benefited from exporting performance; for instance, relaxation of the balance of payment, upgrading technology, economies of scale and so on, therefore, it is essential to investigate the factors affecting Afghanistan's exports. The aim is to find out the major determining factors of Afghanistan's export using panel data from the period 2006 to 2016, with the application of gravity model approach.

Economists like Tinbergen (1962) was amongst the first who used the gravity model to study trade flows between countries. In the simplest case, based on the gravity model (with assuming other variables constant), trade between countries has a direct (or positive) relation with the income (GDP, GNP) of these countries and inverse relationship with the geographical

distance of the two countries(just like Newton's gravity law in physics) (Pass, 2000). In other words, as much as the two countries size are larger (the two countries have higher GDP, GNP, national income, the wealth of nation and so on) and geographically closer, then it's expected that the trade flow between the two countries would be higher.

Among the new trade theories those by Krugman (1979), Helpman (1984, 1987), Helpman and Krugman (1985) and Deardorff (1984, 1998) special attention is paid on explanation of international trade both empirically and theoretically. Mathur (1999) explains that the factor affecting trades are country size and economies of scale.

According to the gravity model, one would relate the flow of exports with the size of the two countries, the distance between the two countries, and other influential factors (Rahman, 2010). The general form of this model is as follows: $Xij = \alpha k \frac{YiYj}{Dij}$ ... ... 1 Where Xij is the export flow between the two countries i and j, αk is constant, Yi and Yj are respectively the GDP in the country i and j, and D_ij is the distance between the capital of the two countries. The above model expresses the flow of exports between the two countries depending on the size of the two economies and geographical distance. To prepare the above model for the ordinary regression analysis, we take the logarithm of the above model and modifies as follow: $\log(EXPij) = \beta_0 + \beta_1 \log(Yi\ Yj) + \beta_2 \log(Dij) + uij$.....2 One can explain the economic size of a country with GDP, GDP, and population or GDP and per capita income (paas, 2000).

In the present research, GDP and per capita income indicate the economy size of the countries, with bringing of those variables in the model, the gravity model modifies as follow: $\log(EXPij) = \beta_0 + \beta_1 \log(GDPi\ .GDPj) + \beta_2 \log(PCIi\ .PCIj) + \beta_3 Log(Dij) + uij$....3 With the developing of gravity model more variables such as transportation cost, common border, membership in trade blocks (i.e., such as unions and free trade areas (FTAs)), have been added to the basic model as the influential variables (Elshehawy1, Shen1, & Ahmed, 2014). With a few changes and bringing some more independent variables in the equation (3), it is being prepared to explain and describe the factors affecting Afghanistan's export. Therefore, the export flow between Afghanistan and a second country is explained as follows:
$\log(EXPij) = \beta 0 + \beta_1 \log(GDPit) + \beta_2 Log(GDPjt) + \beta_3 \log(PCIit) + \beta_4 Log(PCIjt) + \beta_5 \log(EXRijt) + \beta_6 \log(OPNSijt) + \beta_7 Log(Distij) + \beta_8 lGDPDEFi + \beta_9 lGDPDEFj + \beta_{10} BTA_{ij} + \beta_{11} RTA_{ij} + \beta_{12} Border_{ij} + uijt$.....4. Where all variables (excluding dummies) are represented in natural logarithms. The dependent variable EXP_ij represents the flow of exports into the trading partner j from Afghanistan (i)[2]. Independent variables include Gross Domestic

[2] The (i) subscript indicates Afghanistan and (j) subscript indicates Afghanistan's trading partner

Product(GDP), per capita income(PCI), exchange rate(EXR), openness(OPNS), distance(DIST), and GDP deflator(GDPDEF). Dummies are common border(Border), bilateral trade agreement (BTA) and regional trade agreements(RTA).

2. Methodology

The sampling size in this study is the Islamic Republic of Afghanistan and its 17 major trading partners, which Afghanistan had the highest exports to these countries during the past ten years. The countries included in the study are Pakistan, Iran, China, Japan, India, United Arabia Emirate, Turkey, Tajikistan, Uzbekistan, Turkmenistan, Kazakhstan, Saudi Arabia, United States of America, United Kingdom, Iraq, Germany, and Russia. Panel data is used for the period 2006-2017, which is obtained from the WDI (World Development Indicator) dataset, National Statistic and Information Authority of Afghanistan (NSIA) statistical yearbooks, and UN Comtrade database.

The method for estimation of panel data is pooled OLS, Fixed Effect, and Random Effect. For pooled OLS to produce a consistent estimator of β_1 (see equation 5), we would have to assume that the unobserved effect, a_i, is uncorrelated with x_{it1}, and also u_{it} is uncorrelated with x_{it1}, even if we assume that x_{it1} and u_{it} are uncorrelated, but if ai is correlated with x_{it1} , then the pooled OLS estimate is biased and inconsistent, this is also called heterogeneity biased. And also must remember that even if ais uncorrelated with all explanatory variables in all periods, the pooled OLS standard errors and test statistics are generally invalid: they ignore the often substantial serial correlation in the composite errors, $v_{it}=a_i+u_{it}$, For choosing fixed effect or random effect model suppose the following model: $y_{it}=\beta_0+\beta_1 x_{it1}+\cdots+\beta_k x_{itk}+a_i+u_{it}$ .....5 . According to Wooldridge(2012) The fixed effects model permit for correlation between a_i the used explanatory variables in any time period. The random effects estimator will be applied when the unobserved effect is uncorrelated with all the explanatory variebles. Equation (5) becomes a random effects model when we assume that the unobserved effect ai is uncorrelated with each explanatory variable: $Cov(x_{itk},aj)=0, t=1,2,\ldots T$; $j=1,2,\ldots,k$. … … .6

It's common to see researchers apply both random effects and fixed effects, and then doing Hausman test for statistically significant differences in the coefficients on time-varying explanatory variables. In Hausman test, the null hypothesis is that the random effect model is appropriate, and the alternative is that the fixed effect model is appropriate. The idea is that one uses the random effects estimates unless the Hausman test rejects (6). A rejection of Hausman test means that the key RE assumption, (6), is false, and then the FE estimates are used (Wooldridge, 2012).

3. Results

Table 2. exhibits fixed effects and random effects estimates of the gravity model, the question of whether

Table 2: Estimation results

	Random effect model			Fixed effect model		
	Coef.	Std.Err.	P	Coef.	Std.Err.	P
LGDPi	.8022206	1.332343	0.547	1.104496	1.372124	0.422
LGDPj	.6371107	.2739299	0.020	-1.16232	2.433631	0.634
LPCIj	-.5816107	.354818	0.111	.9442538	2.365327	0.690
LEXRij	-.5816107	.0914405	0.166	1.31815	1.124058	0.243
LOPNSj	1.021061	.6278	0.104	.8739843	1.11427	0.434
LGDPDEFj	1.778355	.6325967	0.005	2.684018	1.180128	0.025
LGDPDEFi	-2.310072	1.753052	0.188	-2.811453	2.090857	0.181
LDistij	-1.523003	.7575798	0.044	-	-	-
Border	-.8508715	.7039504	0.227	-	-	-
BTA	1.935338	.7268324	0.008	-	-	-
RTA	.0216736	.7342148	0.976	-	-	-
Constant	10.5316	8.184623	0.198	-6.205157	13.62918	0.650
Obs.	152			152		
F (p-value)	0.0000			0.0000		
Hausman(P-value)	0.9880					

Source: Authors' finding

A significant correlation exists between the unobserved or country-specific effect, and control variables will be answered by the Hausman test. The Hausman test represents that the null hypothesis can't be rejected, because the p-value is 0.988, which one can't reject the null hypothesis at 5% and 10% significance level. It meant that the country-specific effect is not correlated with the explanatory variables. Therefore, the study selects random effect as the model of choice.

The regression result of the gravity model of export in table 2 is reported based on the result from the fixed effect and random effect method. The result for the export of Afghanistan from the random effect model shows that importers GDP, importer's openness ratio, importer's GDP deflator, the distance between Afghanistan and its trading partners and also bilateral trade agreement are statistically significant. The other variables such as GDP of Afghanistan's, exchange rate, GDP deflator of Afghanistan's, common border, and regional trade agreement are not statistically significant. Note that the per capita income of Afghanistan omitted from the model because of having high correlation with GDP of Afghanistan.

Based on the results from the regression, the coefficient for the GDP of Afghanistan (i) is positive, and it's according to the theory. This variable

determines the economic size of the country, although this is not statistically significant, but economically a large factor, which one cannot ignore its effect on export.

The positive coefficient for the importer GDP implies that as long as the production level of goods and services of importer countries increase, the trade flow increase, which in result the import and export will increase. This result explains that Afghanistan's export increase by .637 percent if the importers GDP increase by 1 percent.

The importer's trade liberalization or openness ratio is statistically significant and has a positive effect on the export of Afghanistan. Whatsoever, the importer reduces its trade barriers; the result is that import will increase. The coefficient value is 1.021061, suggesting that a 1% increase in openness ratio will translate to 1.021061% increase in export of Afghanistan.

An increase in the price of goods and services of importer will translate to a rise in export of Afghanistan, this coefficient (GDPDEFj) is 1.778355, and statistically significant. This coefficient means that a 1% increase in prices of goods in importer country will increase 1.778355% export of Afghanistan. The result implies that if the importer's prices of products increase, then the term of trades would go in favor of Afghanistan and export will increase.

The distance variable is significant at 5% and bears the anticipate negative sign which has an inverse relationship with exports. The distance was factored in as a proxy for the transportation costs. The relationship implies that the further away from Afghanistan the importer is located, the higher the transportation costs and therefore less amount of exports would go out to that particular country. The coefficient value -1.523003 indicates that when the distance between Afghanistan and its trading partner increase by 1%, the value of exports to this destination decrease by approximately 1.523003 %.

The bilateral trade agreement coefficient is statistically significant at the 1% significance level; it indicates a direct relationship between the bilateral trade agreement and export. The coefficient 1.935338 implies that the export of Afghanistan to those countries which has bilateral trade agreement is 1.93% more than the countries which Afghanistan don't have bilateral trade with them.

4. Discussion and Conclusion

Foreign trade is considered as one of the vital fields of the economy and has a crucial role in defining the future of an economy and the process of economic growth. The present research is assessing the factors affecting on Afghanistan's export flow to its 17 major trading partners. The panel data with the gravity model have been used for analyzing the determinants of export from the 2006-2016 period. In addition to necessary variables of gravity model such as GDP and distance, other variables like trade openness, exchange rate,

bilateral trade agreement, GDP deflator, common border, and regional trade agreement also added in the model.

The fixed effect and random effect models have been used to estimate the model, based on the result which revealed from Hausman test, the model with random effects preferred to be used for the exports of Afghanistan. In estimating the gravity model of Afghanistan's export, our result shows that economic size(GDP), trade openness and prices of goods (GDP deflator) of importing countries; furthermore, distance, and bilateral trade agreement have significant impact on Afghanistan's exports.

According to the result of the random effect model, the GDP of Afghanistan doesn't have a statistically significant effect on the export of Afghanistan. The reason could be that the composition of Afghanistan's export is more agriculture products, for instance, dried & fresh fruits, however, the share of agriculture products in GDP is around 20%, which in on the other side, the share of services are more than 50%, and we don't have any remarkable export of services in the past ten years (even the services export is negligible).

Bilateral trade agreements have a significant impact on Afghanistan's exports, in fact, that Afghanistan is a landlocked mountainous country, and get access to sea route by neighbor country (Pakistan), therefore, Afghanistan's trade especially the transit with Pakistan, has always faced severe challenges, and it is sometimes causing considerable losses to the country's economy and traders. Policymakers need to look for alternative ways to reduce its dependence and vulnerability on one or two neighboring countries. In addition, an increase in production of agricultural products, having stable political situation and good relationships with neighbors' will expand the exports more.

References

1. Deadorff, A. V. (1984). Testing Trade Theories and Predicting Trade Flows. In *Handbook of International Economics 1* (Vol. 1, pp. 467-517). Amsterdam, North-Holland.
2. Deardorff, A. (1998). Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World? *The Regionalization of the World Economy,*, 7-32.
3. Elshehawy1, M. A., Shen1, H., & Ahmed, R. A. (2014). The Factors Affecting Egypt's Exports: Evidence from the Gravity Model Analysis. *Open Journal of Social Sciences, 2*, 138-148. doi:10.4236/jss.2014.211020
4. Evenett, S. J., & Keller, W. (1998). On the Theories Explaining the Success of teh Gravity Equation. *NBER Working Paper*.
5. Helpman, E. (1987). Imperfect Competition and International Trade: Evidence from fourteen Industrial Countries. *Journal OF Japanese and International Economics, 1*, 62-81.

6. Helpman, E., & Krugman , P. (1985). *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy* (Vol. 1). MIT Press Book and The MIT Press.
7. Krugman, P. (1979). Increasing Returns, Monopolistic Competition, and International Trade. *Journal of International Economics, 9*, 469-479.
8. Mathur, S. K. (1999). Pattern of International Trade, New Theories and Evidence from Gravity Equation Analysis. *The Indian Economic Journal, 47*, 68-88.
9. *National Statistics and Information Authority*. (2017). Retrieved from https://www.nsia.gov.af/
10. Pass, T. (2000). Gravity Aproach for Modeling Trade Flows Between Estonia and the Main Trading Partners. *Working Paper*.
11. Rahman, M. (2010). The Factors Affecting Bangladesh's Exports: Evidence from the Gravity Model Analysis. *Journal of Developing Areas, 44*, 229-244. doi:http://dx.doi.org/10.1353/jda.0.0075
12. Tinbergen, J. (1962, February). Shaping the World Economy: Suggestion for an International Economic Policy, New York, The Twenthieth Century Fund. *American Journal of Agricultural Economics, 46*(1), 271-273. doi:https://doi.org/10.2307/1236502
13. *UN Comtrade Database*. (2019). Retrieved from https://comtrade.un.org/
14. Wooldridge, J. M. (2012). Introductory Econometrics. SOUTH-WESTERN CENGAGE Learning.*World Development Indicators (WDI)*. (2019). Retrieved from http://datatopics.worldbank.org/world-development-indicators

Index

Index

ISIWSC2019

Organised by :

DEPARTMENT OF STATISTICS MALAYSIA
MINISTRY OF ECONOMIC AFFAIRS

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

Supported by:

MINISTRY OF TOURISM, ARTS & CULTURE MALAYSIA

Malaysia Convention & Exhibition Bureau

Visit Truly Asia Malaysia 2020

ISBN 978-967-2000-63-1

9 789672 000631

#ISIWSC2019