62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

SPECIAL TOPIC SESSION

VOLUME 3

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 3)

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 3, 2019. 454 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 2

The indicator system for Sustainable Development Goals (SDGs) monitoring in the Philippines: Its current status

Wilma A. Guillen, Bernadette B. Balamban, Mechelle M. Viernes
Philippine Statistics Authority

Abstract

The Philippines, like other member countries of the United Nations, has committed to achieve the Sustainable Development Goals (SDG)s by 2030. Progress has been made in the development of the indicator system for the Philippine SDG monitoring based on the global set of indicators and engaging multi-stakeholders that included national government agencies, academe, civil society organizations, private sector, and international agencies. This initiative ensures that the Philippine indicator system for the SDGs is inclusive and works best for the country to meet the targets set. The Philippine SDG indicator system consists of 155 indicators, of which 102 are global SDG indicators, 28 are proxy indicators and 25 are supplemental indicators, spread over 17 goals and monitoring 97 targets. Baseline and latest estimates for most of these indicators are currently available.

To further strengthen the monitoring of the Philippine SDG Indicator Framework and to ensure the achievement of its targets, cascading of the SDG monitoring at the local level is needed. The localization initiative builds ownership of the local government units (LGUs) on the SDGs as these were integrated in the local development plans and programs with corresponding budget allocation and also strengthens the capacities of the LGUs in establishing local monitoring systems for the SDGs. This initiative also increases the LGU's appreciation on their roles not only on the realization of the SDGs but also on the SDG monitoring at the local level.

This paper will discuss the multi-sectoral approach adopted by the Philippines in the development of its SDG indicator system for initial monitoring, as well as the various mechanisms established to facilitate the monitoring and dissemination of SDG indicators. This paper will also highlight initiatives in the localization of the monitoring of SDG indicators that facilitate the link of the monitoring of the Philippine Development Plan, regional development plans and local development plans. The paper will also present the initial results of the mapping exercise of the indicators that are available at the global, national, regional, provincial and municipal level and those that are aligned in the local development plans. This paper will also highlight the non-traditional statistics and various data collection initiatives of the LGUs to address the data gaps at the local level

Keywords
Philippine SDG Indicator Framework, localization initiative, local development plans

1. Introduction

The Sustainable Development Goals (SDGs), or officially coined as Transforming our World: the 2030 Agenda for Sustainable Development, has been a famous in-house term within the Philippine Statistical System (PSS) since 2015. The Goals are "an intergovernmental set of aspiration with 69 targets" that are hoped to transform the world by 2030. A few of its famous goals are eradicating extreme poverty and hunger; achieving universal primary education; promoting gender equality; fighting inequalities; increasing country's productive capacity; increasing social inclusion; curbing climate change; and protecting the environment while ensuring that no one is left behind over the next fifteen years. It was built on the successes of the Millennium Development Goals (MDGs), while including new areas such as climate change, economic inequality, innovation, sustainable consumption, peace and justice, among other priorities.

Relatedly, the Philippine Statistical System has been up and active on the initiatives related to SDG Monitoring. The Philippine Statistical Development Program (PSDP) 2011- 2017 Update, included an additional chapter on the SDG Monitoring, which outlines the statistical development programs that will be implemented to ensure data support for the SDGs, among others. The PSDP was updated to 2018-2023, which still includes chapter on the SDG monitoring.

Further, in line with the mainstreaming of the SDGs in regional, national, and sub-national/local development planning and policy formulation, and to provide information support for the monitoring of the country's progress in attaining the SDGs, it is imperative that the PSS institutionalize the provision of data support for the generation of SDG indicators. The necessary institutional arrangement for the monitoring of the SDGs is supported by PSA Board Resolution No. 4, Series of 2016 "Enjoining government agencies to provide data support to the sustainable development goals". This Resolution also designated the PSA as the official repository of SDG indicators.

Certainly, the SDGs are extremely important because they are powerful platform that sets the tone and direction for the development, aid and partnership until 2030 among and within countries.

Cognizant to the need to implement and monitor the SDGs, it is recognized that countries should develop their country level indicator framework based on the data availability, priorities and relevance in the national context. This paper will build on the efforts and initiatives of the PSS in the development of the SDG Indicator Framework. Specifically, this paper will discuss the

assessment of the global SDG indicators and development of the Philippine SDG Indicators. This paper will also discuss the initiative in the integration of the SDGs in the national and local development plans. Further, this paper will also present the development of core regional SDG indicators and the current SDG assessment initiative at the local level.

2. Methodology

A. Philippine SDG Indicator Framework

The assessment and development of the Philippine SDG Indicator Framework was a collective effort of the PSS and its partners. The country's work on building awareness and ownership for the SDGs began as early as the conception of the Post-2015 Agenda in September 2013. In the Philippines, technical workshops were organized on the SDG assessment in cooperation with the National Economic Development Authority (NEDA) and PSA and funding support from the UN Development Programme (UNDP). Multisectoral and Technical workshops were held to review and discuss the global post-2015 development agenda goals and targets submitted by the high-level panel of eminent persons convened by the UN Secretary General. These workshops served as the best venue to engage different development stakeholders in discussions and dialogues to gather insights and inputs for the crafting of the Post-2015 Development Agenda, and to generate awareness, interest, and ownership.

As part of the commitment of the Philippines to monitor the SDGs, immediately after the approval of the initial Global SDG Indicator Framework in March 2016, the NEDA in collaboration with PSA through the funding assistance from the United Nations Development Programme (UNDP) conducted a multi-sectoral workshop in May 2016, which was participated by national government agencies, academe, private sector, civil society and other non-government organizations. During the said workshop, participants were grouped into different sectoral concerns.

The primary output of the workshop was the SDG Assessment Matrix, which is used to assess and organize indicators for monitoring SDGs in a compact and standard manner. It also served as an input to program implementers/policy-makers/planners in government and private sector, for projects/programs/services aligned towards the achievement of the SDGs. The SDG Assessment Matrix contains the information on the 1) Tier classification at the global and national context; 2) Baseline data information (data availability, frequency of data collection, latest available data); 3)Source of Data; 4) Implementing Organization (agency responsible on the data collection/data sources and agency accountable for the achievement of the target; 5) Relevance; if the indicator is in PDP, not in PDP but relevant, not

applicable in the Philippines; 6) Priority Tier 2 or 3 indicators; and 7) Level of disaggregation need to make the indicator more relevant to the Philippines.
Further, in June 2016, a sector specific workshop was conducted to further assess and review the indicators. A technical workshop was organized by the PSA and the World Health Organization (WHO) to assess the WASH and wastewater indicators in the SDG 6 framework. This workshop became a venue to discuss possible partnerships particularly on the provision of possible technical assistance to be able to monitor the WASH indicators in the Philippines. It was also an opportunity for both the PSS and the WHO to identify how the Joint Monitoring Program of WASH and national priorities can mutually support each other.

Various consultative and bilateral meetings were held to validate and further improve the results of the Multisectoral and technical workshops. These were to ensure that the indicators identified as Tier 1 indicators have clear methodology and definition and possible source/s of data were mapped and this resulted to an expanded SDG Assessment matrix that produced more detailed information about the indicator. These were also a good venue to get the commitment of the data source agencies to generate and monitor these SDG indicators.

This SDG indicators mapping initiative ensured that inclusive participation and wide consultations were made in the development of the Philippine SDG Indicators framework. It involved stakeholders from the national government agencies (NGAs), NGOs, CSOs, academe, media, international organizations and the private sector these stakeholders to ensure responsive, inclusive, participatory and representative decision-making was done at all levels of institutions. This was central to the goal of leaving no one behind.

B. Integration of the SDG in the National and Local Development Plans

It is well-recognized that for the implementation of the SDGs to be successful, it is crucial that these were integrated and mainstreamed in the national and local development plans and strategies.

In the course of the development of the Philippine SDG Indicator Framework, the SDG indicators were matched with the Medium- and Long-Term Vision of the Philippines recognizing that the attainment of the SDGs set for 2030 would pave the way for the achievement of the Philippine Development Plan (PDP) and the country's long-term vision, the AmBisyon Nation 2040 [2].

In the continuing effort to ensure national and local convergence to the international commitments such as the SDGs, in 2018, the Department of Interior and Local Government (DILG) in collaboration with NEDA, PSA and Philippine Statistical Research and Training Institute (PSRTI) conducted regional and provincial workshops to operationalize localization and ensure alignment of the national and local priority thrusts to the SDGs [3]. This effort

contributes to the identification of the provincial and municipal-level indicators that will operationalize and contribute to the attainment of the goals and outcome areas in the PDP and SDGs.

C. Localization of the Philippine SDG Indicator Framework

Given lot of progress that have been achieve in monitoring the SDGs, the Philippines is still faced with challenges particularly on the monitoring of the Philippine SDGIs at the local level.

The localization of an indicator framework is not new to the statistical community however, the process adopted for the localization, which facilitated the linking of the Philippine SDG indicators to the national development plan, regional development plans and local development plans may be considered as revolutionized since local government units were involved in the discussion and the SDGs were integrated in the local development plans.

As a result of the regional and provincial consultations and assessments on the SDGs, the PSA came up with Core Regional SDG indicators (CoRe-SDGIs), which are consistent with the Initial List of the Philippine SDG Indicators. The Core-SDGI was defined as the minimum set of SDG indicators for sub-national compilation and dissemination to facilitate sub-national comparisons to help monitor the achievement of the SDGs.

In order for an indicator to be included in the CoRe-SDGIs, at least one of the criteria must be met, which includes the following: (1) it must be consistently tagged as Tier 1 in all regions, (2) the indicator must be identified available at the regional level during the national multi-sectoral workshops and must have available baseline data in SDG watch, (3) it must be classified as Tier 1 in at least 70% of the regions based on their submitted assessment matrices, and (4) the data source of the indicator is known to have regional disaggregation. These lists underwent review and was endorsed by the Regional Statistics Committees (RSCs)

Further, as a response to the enormous challenge brought about by the localization of the SDGs, the PSA together with the DILG and PSRTI collaborated in the SDG localization initiative, which presented a distinctively new approach in the development of an indicator framework for its monitoring as it reached beyond the usual borders of the PSS and brought in new voices such as the development planners from the local government units and data partners in the discussion, recognizing their roles not only in the monitoring but more importantly in the realization of the SDGs.

The localization mapping resulted to the mapping of the indicators that are available in the Regional Development Plan and Provincial Development Plan Results Matrices. Further, the list of SDG indicators that are common to most of the municipalities were identified.

3. Result

The assessment made on the availability of the SDG indicators in the Philippines showed that out of 232 unique indicators, 93 indicators were classified as Tier 1 indicators, 55 were Tier 2 indicators, 71 were Tier 3 indicators and 13 indicators were not applicable to the country.

Among the Tier 1 indicators, around 32% will be sourced from PSA surveys, censuses and administrative data, 67% will be sourced from the surveys and administrative data of other government agencies and 1% will be sourced from the data of the International Agencies.

The mapping initiative paved a way on the approval of the PSA Board Resolution No 09 Series of 2017- Approving and Adopting the Initial List of Sustainable Development Goals for Initial Monitoring, reinforcing the importance to provide statistical information to monitor achievement of SDGs to the 155 Philippine SDG indicators, which includes 102 global SDG indicators, along with 25 proxy indicators and 28 supplemental indicators.

Similarly, among the Philippine SDG Indicators, 33% will be coming from the PSA, 66% from other government agencies and 1% from other sources. In terms of the sources of the primary data, 66% of the Philippine SDG indicators will be from administrative data, 23% from survey, 5% from the combination of the administrative data and survey and 6% from the combination of the administrative data and census. Looking at the available disaggregation, 66% of the indicators are available at the regional level, 42% are available at the provincial level and 84% can be disaggregated by sex.

The mapping exercise of the Philippine SDG indicators and the PDP Results Matrix (PDP-RM) showed that 68 out of the 155 approved SDG indicators were integrated in the PDP-RM.

In recognition of the need to facilitate sub-national comparisons in monitoring the achievement of the SDGs, there were 14 goals. 42 targets and 68 indicators identified as the initial CoRe-SDGIs, which will be monitored by all the regions.

Linking the Philippine SDG Indicators with the sub-national SDG Indicator Framework, of the 155 Philippine SDG indicators, on the average, there were 68 indicators that were available in Region XI (Davao Region), while on the average, there were 66 provincial SDG indicators. Drilling down to the city and municipal level, on average, there were 57 indicators that were available at the city and municipal level.

Examining the integration of the SDG indicators to the local development plans, of the 68 integrated SDG indicators in the PDP RM, 34 were included in the Regional Development Plan RM, while on the average, there were 28 SDG indicators integrated in the provincial development plan.

In Central Luzon, there were 95 indicators that were available at the regional level. It can be noticed that the Central Luzon had higher number of

indicators than the identified Co-Re SDGIs. Looking at the linkage of these indicators at the provincial and city and municipal level, there were 75 indicators were available at the provincial level while 52 indicators were found to be available or can be monitored at the city and municipal level.

In terms of the integrated SDG indicators in the local development plans in Region III, there were 54 SDG indicators were included in the Regional Development Plan while an average of 45 SDG indicators were included in the provincial development plan RM.

4. Conclusion

It is crucial that in the development of indicators frameworks, multi-stakeholders that includes national government agencies, academe, civil society organizations, private sector, and development partners will be engaged in the consultation to ensure inclusivity and their voices represented. These consultations also serve as a useful mechanism in crafting a well-designed national and sub-national accountability mechanism that is participatory, integrated and transparent to support the SDG monitoring and reporting as well as the SDG implementation.

In the current practice of bringing new voices into the discussion on data, accountability is shared and responsibility on planning, multi-stakeholder partnerships, resource mobilization and monitoring is ensured for the delivery of products and services to the public.

Further, the initiative resulted in an increasing awareness and appreciation of the need to properly plan an evidence-based SDG implementation. It also made the policy- and decision-makers see the need to work together to have a strong and healthy data ecosystem that could ensure that the progress towards achieving the SDGs is properly measured.

It is also worth to emphasized that there should be high appreciation and understanding on the linkage of the national, regional, provincial, city and municipal indicators frameworks to avoid the pitfalls on the achievement of the SDGs and the national and local development strategies. The greater insight of the linkage of these indicator systems will translate into local and regional policy interventions grounded in empirical evidence.

As what Dahl (2012) underscored that having a set of indicators that are linked to the development policies which are adopted, updated and reported by the local government, "can provide clear signals on the success or failure of national policy initiatives and actions."

References

1. Dahl, A.L. Achievements and gaps in indicators for sustainability. Ecological Indicators 2012, 17, 14–19

2. National Economic Development Authority (2016). Voluntary National Review at the 2016 High-Level Political Forum on the Sustainable Development Goals (SDGs)
3. Department of Interior and Local Government (2018). Joint Memorandum Circular No. 1 Series of 2018

Expanding the Philippine Indicator System for Sustainable Development Goals (SDGs) monitoring through research

Jessa S. Lopez, Millete R. Santos
Philippine Statistical Research and Training Institute (PSRTI), Quezon City, Philippines

Abstract

The current indicator system for monitoring the Sustainable Development Goals (SDGs) in the Philippines includes 155 indicators spread over 17 goals and monitoring 97 targets. Of the current indicators, 102 are in accordance with the United Nations (UN) standards, 28 are proxy indicators and 25 are supplemental. The data needed to operationalize the 155 indicators are regularly produced by institutions comprising the country's statistical system.

In 2017, the Philippines, through the Philippine Statistical Research and Training Institute (PSRTI), the research and training arm of the Philippine Statistical System (PSS), engaged the services of some retired government statisticians to conduct researches with the objective of further expanding the existing indictor system for SDG monitoring. These personnel were taken in as researchers by the PSRTI mainly because of their familiarity of the country's data system and vast experience in developing and updating various indicator systems maintained by the statistical system. As the researchers of PSRTI, the main focus of their studies was on SDG indicators currently under Tiers 2 and 3 classifications. Based on the results of the studies, additional indicators were identified for possible inclusion in the current indicator system for SDG monitoring.

The paper will highlight the findings of these recent researches conducted by the PSS to expand its SDG indicator system, the role of PSRTI in the conduct of these researches, the list of indicators proposed for inclusion in the current indicator system as a result of the researches, and the distribution and classification of proposed indicators by goals and targets monitored.

Keywords

SDG Indicators, Philippines Statistical System (PSS), PSRTI

1. Introduction

The Philippine Statistics Authority (PSA) Board through its Resolution No. 4 Series of 2016 enjoins national government agencies (NGAs) to provide data support to the Sustainable Development Goals (SDGs). The Resolution also states that the Philippine Statistical Research and Training Institute (PSRTI), the research and training arm of the Philippine Statistical System (PSS), shall undertake capacity building activities to help PSA and other agencies generate

the indicators to monitor the country's performance vis-à-vis the SDGs, archive data on such, and conduct methodological research to address issues in generating the SDG indicators.

In response to this task, a project entitled "A Five-Year Research and Training Program for the Philippine Statistical System to Measure the Sustainable Development Goals Indicators" was conceptualized by the PSRTI. This project aims to make the PSS ready and able to meet the challenges of meeting and monitoring the SDGs by conducting researches on the indicators and capacity building activities at the local level.

In 2017, the PSRTI engaged the services of some retired government statisticians to conduct researches with the objective of further expanding the existing indictor system for SDG monitoring. These personnel were taken in as researchers by the PSRTI mainly because of their familiarity of the country's data system and vast experience in developing and updating various indicator systems maintained by the statistical system. The paper discusses the findings of the researchers, the list of indicators proposed for inclusion in the current indicator system as a result of the studies, and the recommendations to institutionalize the proposed indicators.

2. Methodology

The studies aimed to expand the current SDG indicator system in the Philippines by proposing indicators for institutionalization. To do this, operationality of the definition of the indicators in the country and at the local level have to be determined and the sources of data have to be identified. However, PSA has included this activity in their work plan. Hence, in agreement with PSA, the PSRTI decided to focus on those indicators which are not in the Philippine Development Plan (PDP) and those in the PDP but the PSS does not have the statistical capacity or concrete plans yet to address data gap. With this, thirty-three (33) indicators were studied.

For these indicators, the researchers reviewed the definition of the indicators at the global level using available metadata and identified the (i) applicability of the definition at the national and subnational level, (ii) variables needed, (iii) measurability and viability of indicators, (iv) data sources and (v) frequency and disaggregation. These activities were done in close coordination with PSA as definitions of terms used for the indicators have to adhere to PSA standards.

The researchers did comprehensive review of the existing data in the country and the methodology of collecting these data. In addition, PSRTI conducted series of consultative meetings with stakeholders and possible data producers to gather their inputs. Results of these meetings were summarized to come up with a set of recommendations to institutionalize the identified indicators. Lastly, validation of the operational definitions proposed at the

subnational level was also conducted in selected highly urbanized cities through consultative meetings with the local government officials.

3. Result and Discussion

As a result of the comprehensive study and consultative meetings conducted, three recommendations have been proposed: (1) adopt the global the definitions; (2) introduce proxy indicators and sub-indicators; and (3) exclude the indicators for the Philippine SDG monitoring.

Out of the 33 indicators, 17 were recommended to adopt the global definitions. That is, these indicators can be operationalized at the Philippine context based on the existing metadata established at the global level. Table 1 lists down the matched indicators with the respective sources proposed by the researchers.

Most of the recommended data sources already produce official statistics which collection and monitoring have long been established in the country. However, for some national surveys identified as possible sources of data, additional modules will be included to obtain the data requirements at the global level. Some of the proposed sources may already be existing, but there is still a need to institutionalize these to ensure regular collection of data.

Table 1. Matched Indicators and their Proposed Sources

Indicator	Proposed Sources
Goal 2: End hunger, achieve food security and improved nutrition and promote sustainable agriculture	
2.1.1 Prevalence of undernourishment	Department of Science and Technology (DOST): National Consumption Survey (NNS) – Food Consumption Survey (FCS) Module
2.5.2 Proportion of local breeds classified as being at risk, not-at-risk or at unknown level of risk extinction	PSA
2.a.1 The agriculture orientation index for government expenditures	1) Commission on Audit (COA): Audited Financial Reports 2) Department of Budget and Management (DBM): Budget of Expenditures and Sources of Financing (BFSE) 3) PSA: National Accounts of the Philippines (NAP)

2.a.2 Total official flows (official development assistance plus other official flows) to the agriculture sector	National Economic and Development Authority (NEDA): Official Development Assistance (ODA) Portfolio Review - Agriculture, Agrarian Reform and Natural Resources (AARNR) Report
2.c.1 Indicator of food price anomalies	PSA: Consumer Price Indices
Goal 3: Ensure healthy lives and promote well-being for all at all ages	
3.1.1 Maternal mortality ratio	PSA: 1) National Demographic and Health Survey (NDHS) 2) Municipal Form 103: Certificate of Death
3.b.2 Total net official development assistance to medical research and basic health sector	NEDA: ODA Portfolio Review - Social Reform and Community Development (SRCD) Report
3.c.1 Health worker density and distribution	Department of Health (DOH): National Database of Selected Human Resources for Health (NDHRH)
Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	
4.5.1 Parity indices for all education indicators	PSA
4.b.1 Volume of official development assistance flows for scholarships by sector and type of study	NEDA: ODA Portfolio Review - Scholarship Assistance Program Report

Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	
8.9.2 Proportion of jobs in sustainable tourism industries out of total tourism jobs	PSA: Philippine Tourism and Satellite Accounts (PTSA)
Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	
9.3.1 Proportion of small-scale industries in total industry value-added.	PSA: 1) Annual Survey of the Philippine Business and Industry (ASPBI) 2) Census of Philippine Business and Industry (CPBI)
9.5.2 Researchers (in full-time equivalent) per million inhabitants.	DOST: Research and Development Survey on Personnel and Expenditures
9.b.1 Proportion of medium and high-tech industry value added in total manufacturing value added	PSA: 1) ASPBI 2) CPBI
Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable	
11.6.1 Proportion of urban solid waste regularly collected and with adequate final discharge out of total urban solid waste generated, by cities	Metropolitan Manila Development Authority (MMDA)
Indicator	Proposed Sources
11.6.2 Annual mean levels of fine particulate matter (e.g. PM2.5 and PM10) in cities (population weighted)	Department of Environment and Natural Resources (DENR): National Air Quality Status Report

Goal 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	
16.10.1 Number of verified cases of killing, kidnapping, enforced disappearance, arbitrary detention and torture of journalists, associated media personnel, trade unionists and human rights advocates in the previous 12 months	1) Center for Media Freedom and Responsibility (CMFR) 2) Commission on Human Rights (CHR)

The researchers then proposed alternative ways to operationally define and measure the other remaining indicators, as summarized below. For some indicators, their global definitions were modified to limit the characterization of key terms and the covered population. For most, proxy indicators were introduced to provide indirect measures for primary data that cannot be obtained due to logistic and budgetary constraints.

Goal 1: End poverty in all its forms everywhere.

1.4.1 Proportion of population living in households with access to basic services. This indicator shall be represented by the proxy indicator, proportion of household population living in communities with access to basic services measured as the proportion of the population living in households within barangays with access to services that support the aspirations of Filipino families for a life that is strongly rooted, comfortable and secure. These basic services focused on housing and urban development, connectivity, education, tourism, agriculture, manufacturing, health and wellness, and financial services. The PSA can provide baseline data through the census of population conducted every five (5) years.

Goal 3: Ensure healthy lives and promote well-being for all at all ages

3.3.1 Number of new HIV infections per 1,000 uninfected population, by sex, age and key populations. At the national context, this indicator shall be defined as the HIV incidence rate (per 1000 population) or the number of new HIV cases per population at risk within a given a period time. The data on HIV incidence shall be sourced from the HIV/AIDS and ART Registry of the Philippines (HARP) of DOH.

Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

4.a.1 Proportion of schools with access to:

(d) adapted infrastructure and materials for students with disabilities; The Department of Education (DepEd) proposed that the number of adapted infrastructure and materials for students with disability be initially measured

in terms of existence of ramp facilities in the school premises. Hence, the indicator shall be measured as the proportion of the total number of schools with ramp facility to total number of schools.

(f) single-sex basic sanitation facilities; In SY 2017-2018, indicators on the number of functional toilet bowls and urinals for males, females and persons with disability (PWDs) were included in DepEd's School Building Inventory Form. These shall provide information to measure this indicator as the proportion of total number of schools with basic functional sanitation facilities separated for males and females to total number of schools, disaggregated by level of education.

(g) basic handwashing facilities (as per the WASH indicator definitions). This indicator is proposed as the proportion of schools with access to basic and functional handwashing facilities, soap or ash and water available to all girls and boys during school hours, disaggregated by level of education. A functional handwashing facility is defined as accessible, with daily water supply, at learner appropriate height and with appropriate drainage. These data were included in DepEd's School Building Inventory Form starting SY 2017-2018.

4.2.1 Proportion of children under 5 years of age who are developmentally on track in health, learning and psychological well-being, by sex. This indicator shall be defined as the proportion of children 5 years of age who have demonstrated readiness for primary education or have demonstrated socio-emotional development, values development, physical health and motor development, aesthetic/creative development, mathematics, understanding of the physical and natural environment, and language, literacy and communication. The total number of kindergarten pupils and their learning assessment shall be sourced from DepEd to obtain an estimate of the number of pupils who are ready for primary education.

4.3.1 Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, by sex. For the Philippines, it is recommended that this indicator be limited to youths, hence, shall be measured as the participation rate of youths in formal and non-formal education or the percentage of youths (aged 15 to 24 years old) who are enrolled in formal and non-formal education or training in the past 12 months. The data on enrolment can be sourced from the administrative records of DepEd, Commission on Higher Education (CHED) and the Technical and Skill Development Authority (TESDA).

4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill. Adopting the global definition for required computer-related activities to measure ICT skills, this shall be measured as the percentage of youth (aged 15-24 years old) and adults (aged 15 years and above) that have undertaken computer-related activities,

whether at home, school or work in a given time period. However, instead of type of skill, it is recommended to disaggregate this indicator by level of skill with Level 0 being the lowest (cannot do any of the computer-related activities) to Level 3 (can do any of 7 to 9 of the activities).

Goal 6: Ensure availability and sustainable management of water and sanitation for all

6.1.1 Proportion of population using safely-managed drinking water. In the Philippines, it is recommended to use family as the unit of measure so that the indicator shall be defined as the percentage of families using safely managed drinking water sourced from piped water into dwelling, yard/plot, neighbor, public tap/stand pipe; tubed well/borehole; protected dug well; protected spring; rainwater; water refilling station and packaged water. Excluded from the category of sources of safely-managed drinking water are unprotected well, unprotected spring, lake/river, tanker-truck, cart with small tank, and surface water. Annual data can be obtained from the Family Income and Expenditure Survey (FIES) conducted every three years and the Annual Poverty and Income Survey (APIS) conducted during the years when there is no FIES.

6.2.1 Proportion of population using safely-managed sanitation facility, including a hand-washing facility. It is also recommended to use family as the unit of measure for this indicator so that it shall be defined as the percentage of families using safely managed toilet facility, that is, the toilet facility classified as any of the following types: flush/pour flush to piped water system, septic tank, pit latrine; ventilated improved pit; pit latrine with slab; and composting toilet. Excluded from this type are open pit latrine, and others that are not considered as safely managed. The data can be sourced from surveys conducted by PSA, FIES and APIS.

Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

9.c.1 Proportion of population covered by mobile network, by technology. There are three (3) proxy indicators recommended to represent this global indicator: (1) proportion of cell sites of all Cellular Mobile Telephone System (CMTS) network operators, broken down by all available technologies (2G, 3G, 4G, LTE); (2) CMTS density or the proportion of CMTS subscribers to 100 population, by technology; and (3) proportion of cities and municipalities with coverage of CMTS facilities, by technology, regardless of whether or not all areas are within the rage of a mobile cellular signal. Regular data collection for these proposed proxy indicators is yet to be established by the Department of Information and Communications Technology (DICT) and the National Telecommunications Commission (NTC).

Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable

11.1.1 Proportion of urban population living in slums, informal settlements, or inadequate housing. There is difficulty to operationalize the three components (shelter conditions) of this indicator given their overlapping criteria. It is then recommended to monitor its attainment through six indicators related to shelter conditions: (1) proportion of households with no sustainable access to an improved water source; (2) proportion of households with no access to improved sanitation; (3) proportion of households whose housing unit is not made of strong construction materials for the outer walls; (4) proportion of households whose housing unit is not made of strong construction materials for the roof; (5) proportion of households with housing units of less than 26 square meters; and (6) proportion of households with no access to secure tenure.

11.4.1 Total expenditure, public and private, per capita spent on the preservation, protection, and conservation of all cultural and natural heritage, by type of heritage (cultural, natural mixed and World Heritage Center designation, by level of government, type of expenditure, and type of private funding. It was recommended to represent this global indicator by two proxy indicators: (1) Per capita consolidated annual budget of concerned cultural NGAs, as provided for in the General Appropriations Act (GAA); and (2) Growth rate of consolidated annual budget of concerned cultural NGAs, as provided for in the GAA. Both proxy indicators cover all NGAs tasked with maintaining, preserving and conserving cultural and natural heritage such as the National Commission of Culture and Arts, National Historical Commission, Cultural Center of the Philippines, National Museum, National Library of the Philippines, National Archives of the Philippines and the Commission on the Filipino Language.

Goal 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels.

16.1.4 Proportion of population that feel safe walking alone around the area they live. In the absence of a global metadata, it is proposed to represent this indicator by a proxy, the ratio of policeman to total population. This shall be measured as the number of active police commissioned and non-commissioned officers of the Philippine National Police (PNP) to the total household and institutional population.

16.3.1 Proportion of victims of violence in the previous 12 months who reported their victimization to competent authorities or other officially recognized conflict resolution. This indicator shall be represented by the proxy indicator, police reporting rate, defined as the percentage of offences of sexual assault and/or physical assault that were reported to the police by the victim

of the offence or someone else, calculated on the basis of the last incident reported by the victim in the last 12 months. At the national level, data shall be sourced from the PNP while the Victims of Crimes module of the Community Based Monitoring System (CBMS) shall produce data at the local level.

16.6.2 Proportion of population satisfied with their last experience of public service. It is proposed to define this indicator by proxy as the proportion of service offices, i.e., government offices with frontline services, that the public were satisfied or very satisfied in accessing public service. That is, the indicator shall be measured by the percentage of service offices which received the rating of acceptable and higher in the Report Card Survey (RCS) of the Civil Service Commission (CSC) in terms of quality, efficiency and adequacy in providing public service.

Among those reviewed, there are two indicators recommended for exclusion in the monitoring of SDG indicators at the Philippine level. Indicator 2.5.1 Number of plant and animal genetic resources for food and agriculture secured in either medium- or long-term conservation facilities is recommended to be excluded given the absence of metadata at the global level while indicator 2.b.1 Agricultural export subsidies is recommended to be dropped from the monitoring system since there are no existing targets for export subsidies cited in the PDP 2017-2022.

4. Conclusion

In summary, through the research conducted by PSRTI in 2017, 33 indicators identified in relation to the PDP and the statistical capacity of the PSS were reviewed in order to identify if these can be added in the current SDG monitoring system of the Philippines. As a result, it was found that the global definition of 17 out of 33 can be adopted in the country while two of these were recommended for exclusion. For the rest of the 14 indicators, proxy and sub-indicators were proposed. Adopting these recommendations will have to undergo the administrative process set by PSA.

Operationalization and institutionalization of Newly Developed Indicators for monitoring the Sustainable Development Goals (SDGs): The Next Step

Sabrina O. Romasoc
Philippine Statistical Research and Training Institute (PSRTI), Quezon City Philippines

Abstract

In 2017, the Philippine Statistical System (PSS), through the Philippine Statistical Research and Training Institute (PSRTI), conducted studies to expand the indicator system for monitoring the Sustainable Development Goals (SDGs) by engaging retired government statisticians as researchers. The studies resulted to the recommendation of additional indicators that may be operationalized for possible inclusion in the current indicator system for SDG monitoring.

This paper discusses the various activities to be done to adopt the results and the recommendations of the different studies on the operationalization and institutionalization of the additional indicators. These activities include the screening of proposed indicators for their appropriateness in measuring the global goals, validation of their measurement methodologies, availability of data to support the regular production of the indicators, adherence to the statistical standards in the production of the needed data, and many others. These activities are in coordination with the Philippine Statistics Authority (PSA) and appropriate inter-agency committees. After these reviews, the indicators will have to be presented to the PSA Board, the policymaking body on statistical matters in the country, for approval. Once approved by the PSA Board, the new indicators will be adopted and included in the existing indicator system for SDG monitoring.

Keywords

Philippine statistical system (PSS), indicator system, PSA, PSRTI

1. Introduction

The Sustainable Development Goals (SDGs) were first introduced at the United Nations Conference on Sustainable Development (UNCSD) in 2012. After much planning of the United Nations Inter-Agency Expert Group, the United Nations Statistical Commission (UNSC) agreed on the 2030 Agenda for Sustainable Development in 2016. The agenda defined 17 Sustainable Development Goals, 169 associated targets, and 230 global indicators. The SDGs and its targets were developed to succeed the Millennium Development Goals (MDGs) and continue to fulfill the gaps that were not achieved by 2015.

Further, the SDGs aim to meet the urgent challenges of the three dimensions of sustainable development namely, economic, social, and environmental.

In the same year of its enforcement, the Philippine Statistics Authority (PSA) Board issued Resolution No. 4 which states that government agencies are enjoined to provide data support in the monitoring of SDGs. The resolution also directs the Philippine Statistical Research and Training Institute (PSRTI) to undertake capacity building activities to help the PSA and other agencies in generating indicators to monitor the SDGs in the country. Moreover, the PSRTI was tasked to conduct methodological researches to address impending issues in generating the SDG indicators.

To fulfill its participation in the PSA Board Resolution No. 4, the PSRTI, in collaboration with PSA, conceptualized the project "A Five-Year Research and Training Program for the Philippine Statistical System to Measure the Sustainable Development Goals Indicators." For this, the institute hired retired government statisticians who are familiar with the country's data system and have vast experience in developing and updating various indicator systems for the research component of the project. This project was geared to review SDG indicators (Tier 2 and Tier 3) that were not regularly produced in the country and recommend possible SDG indicators for possible institutionalization. The research produced a thorough documentation of the review done by the researchers which presented the recommended indicators, its conceptual and operational definitions, its method of computation and recommendations as its way forward.

This paper delves into the results of the Final Technical Report of the Study to Develop Indicators for Sustainable Development Goals (SDG) of the Philippine Statistical Research and Training Institute (PSRTI) that was done in 2017. Particularly, this would focus on the recommended indicators and their way forward. In the list of recommended indicators, some of which were exact matches to their corresponding global indicators; while other recommended indicators were proxy to the indicator or sub-indicators to the global ones. More importantly, the researchers gave their recommendations on whether or not the Philippines is ready to institutionalize the indicator as its way forward. This study would discuss the steps needed to be done in order to institutionalize the indicators that were deemed to be ready for institutionalization on the national level.

2. Methodology

The methodology used for this research can be divided into two parts namely, review of the indicators and the institutionalization process. The review of the indicators that was done by the researchers hired by PSRTI focused on defining selected SDG indicators classified under Tier 2 and Tier 3 that were deemed relevant to the country's increasing efforts on sustainable

development. The second part of the methodology is the institutionalization process. All the recommendations of the researchers were reviewed and summarized to be able to proceed to the possible institutionalization of these selected Tier 2 and Tier 3 indicators. This is done in the hopes of reclassifying these indicators to a better tier classification.

Review of Indicators

The main goal of the project was to make the Philippine Statistical System (PSS) ready and able to meet the challenges of achieving and monitoring the SDGs by providing alternative means to generate disaggregated official statistics using innovative data sources. To achieve this, the PSRTI reviewed several Tier 2 and Tier 3 indicators that were considered viable for institutionalization. The process started with PSA providing a mapping of the possible SDG indicators with their tier classification. From this, 33 indicators were selected to be further explored to see if they can be institutionalized on the national and sub-national level in the country. After which, each indicator was thoroughly researched.

The researchers studied each indicator by mapping the indicator to its corresponding goal and target. Each indicators' global definition was investigated and reviewed to check if it was appropriate or relevant in the Philippine setting. After the review, the researchers recommended one of the four possible options to correspond to the global indicator. The researchers could recommend (i) an exact match to the global indicator, (ii) a proxy indicator to the global indicator, (iii) a sub-indicator to the global indicator, or (iv) exclusion of the global indicator as it is unfit in the country's setting. This review and recommendation brought on the conceptual and operational definitions of the indicators at the national and sub-national level. After these were laid out, the necessary variables in the computation of the indicators were identified together with its possible data sources. The researchers then determined the viability of the indicators/sub-indicators and their method of computation. Finally, the frequency and possible disaggregation of the indicator were identified.

All of these steps were done through various ways. One of which is online research work on the global metadata for the SDG indicators framework. Another way was to research on existing statistical data generated by PSS and other National Government Agencies (NGAs) from websites and publications for information on data sources coverage, frequency of availability, level of disaggregation, definitions and concepts, estimation procedure, data availability, and issues/limitations. It was also done through close communication and verification with the responsible NGA, the corresponding Inter-Agency Committee, or PSA. Finally, the steps were achieved through field validation in different Local Government Units (LGUs) and Highly Urbanized City (HUCs).

Institutionalization of Indicators

After the review of the SDG indicators, the researchers provided recommendations on whether or not the indicator is ready to be institutionalized in the country. These recommendations primarily took into account the readiness of the NGAs, LGUs, and PSA in producing the necessary data and methodology in generating the indicators. There were three possibilities for the recommended indicators: (a) it is ready for institutionalization, (b) it is not ready for institutionalization as it needs further research and coordination, and (c) there is no need to institutionalize the indicator. Global Indicators that were recommended for exclusion (iv) were automatically considered as indicators that need not be institutionalized (c). The remaining indicators that were exact matches (i), proxy (ii), and sub-indicators (iii) were either ready (a) or not (b) for institutionalization.

For indicators that were deemed to be unready for institutionalization (b), there were several suggestions on the necessary steps it must go through before it can be set for institutionalization. One recommendation suggested conducting pilot studies or further methodological research on the indicators before it can be raised to the national level. Similarly, it was suggested to read more literature about some indicators which may be requested from its stakeholders. Another recommendation is to review existing national census/survey questionnaires in order to check if these questionnaires truly contain the variable needed for the indicator or to add particular items in the questionnaire that will accommodate the needed variable. It was also recommended that some indicators be further discussed with the concerned NGAs, LGUs, or PSA. More discussion on the indicator's operational definition, discrepancy with the global indicator, methodology of estimation/computation, availability of data source or the generation of the data is needed before it can be raised for institutionalization. Moreover, discussion and coordination with NGAs, LGUs, and PSA may also include efforts on developing or improving administrative reporting systems whose current shortcomings prove to greatly hinder the generation of needed data.

As for the indicators that were recommended by PSRTI for institutionalization (a), these will be forwarded to PSA. The PSA is set to include these indicators in its Review of the Philippine SDG Indicators. The PSA Review would take into account the methodological developments and available data sources for the indicators at the national level. Part of the review may include deletion, replacements, refinement of indicators, upgrading or downgrading of tier classifications which would be based on the recommendation of several meetings and consultations. The PSA review will also include the refinement of the metadata to ensure the limitations or discrepancies of the national indicator compared with its corresponding global indicator. Further, it may

also include changes in the method of computation, definition or source of data source as deemed necessary by the stakeholders and the reviewers.

The PSA Review of the Philippine SDG Indicators would start with the Interagency Committees for the Sectoral Review of the Indicators. The PSA SDG team will give technical assistance to this event. After which, the Sectoral Technical Consultation will take place. Technical consultations will be done with sectoral stakeholders and representatives from concerned government agencies. This will be followed by the Multi-Sectoral Consultations which is an overall consultation with the concerned stakeholders on the revised draft and refined indicators. After the Multi-Sectoral Consultation, Bilateral Meetings with data source agencies will be held. Data collection methods and revision on the metadata will be discussed to ensure that the data needed for the indicators are truly available. Following this would be the Technical Workshop prepared by PSA and NEDA to finalize the list of indicators. Lastly, the finalized list of indicators will be presented to the focal persons for SDG and then presented to the PSA Board for their approval and definite institutionalization.

3. Result and Discussion

Review of Indicator Results

After reviewing 33 global indicators, PSRTI together with its researchers proposed 38 indicators. The reason the number of proposed indicators (38) exceeded the number of global indicators (33) is that for some global indicators more than one indicator was proposed. An example of this would be global indicator 9.5.2 Researchers (in full-time equivalent) per million inhabitants. For this global indicator, two indicators were proposed namely, Researchers (in full-time equivalent) per million inhabitants and Number of researchers (based on headcount) per million population. There were also multitudes of proposed indicators for global indicators 4.4.4 Proportion of youth and adults with information and communication technology (ICT) skills, by type of skill; 9.c.1 Proportion of population covered by mobile network, by technology; and 11.4.1 Total expenditure, public and private, per capita spent on the preservation, protection, and conservation of all cultural and natural heritage, by type of heritage (cultural, natural mixed and World Heritage Center designation), by level of government, type of expenditure, and type of private funding.

The 38 proposed indicators vis-à-vis their global indicator counterpart can be classified as exact matches (i), proxy (ii), sub-indicators (iii), or recommended for exclusion (iv). Table 1 shows this distribution. It can be seen that based on this classification, almost 50% of the proposed indicators were exact matches to their global counterparts. This is a good indication that there are several global indicators that can be adopted in the Philippine setting. However, the 15 proxy and 5 sub indicators reflect that there is also a good

number of global indicators that cannot be exactly translated in the Philippine setting or it will be difficult to operationalize it in the Philippine setting the same way it is operationalized globally. Furthermore, there are 2 global indicators that were recommended for exclusion as they are not apt for the Philippine setting or is of little relevance to the country. The two global indicators that were excluded are indicators 2.5.1 Number of plant and genetic resources for food and agriculture secured in either medium-or long-term conservation facilities and 2.b.1 Agriculture export subsidies. Indicator 2.5.1 was recommended for exclusion as there is a severe lack of data from most countries and absence of metadata at the global level; while Indicator 2.b.1 was excluded as there are no targets for export subsidies in the current Philippine Development Plan.

Table 1. Count of Proposed Indicators per Type vis-à-vis their Global Indicator Counterpart

Type of Proposed Indicator	Count
Exact match to the Global Indicator	17
Proxy Indicator	14
Sub Indicator	5
No Recommendation/Exclusion	2
TOTAL	**38**

As for the other results of the review of indicators such as the proposed indicators' conceptual and operational definitions, variables needed, possible data sources, methods of computation, frequency, and possible disaggregation, these are discussed in detail in the paper of Lopez and Santos, "Expanding the Philippine Indicator System for Sustainable Development Goals (SDGs) Monitoring through Research."

Institutionalization of Indicators Results

After proposing 38 indicators, each of these indicators were given next step recommendations as its way forward. In summarizing the different recommendations, there were ultimately three types of way forward for the indicators. The proposed indicators may be (a) ready for institutionalization, (b) not be ready for institutionalization, or (c) not needed to be institutionalized.

Table 2. Count of Proposed Indicators Classified by their Way Forward

Next Step Recommendation	Count of Proposed Indicators
Ready for Institutionalization	11
Not Ready for Institutionalization	25
No Need for Institutionalization	2
TOTAL	**38**

Table 2 shows the distribution of the proposed indicators depending on their way forward. Out of the 38 indicators, only 11 of which are ready for institutionalization. This means that these indicators are the only ones ready to be reviewed by the PSA for institutionalization as these are the indicators with sound conceptual and operational definitions, clear method of computation, and available data sources. The indicators that were deemed ready for institutionalization are indicators 2.1.1 Prevalence of undernourishment; 2.a.2 Total official flows (official development assistance plus other official flows) to the agriculture sector; 3.1.1 maternal mortality ratio; 3.b.2 Total net official development assistance to medical research and basic health sectors; 3.c.1 Health worker density and distribution; 3.3.1 HIV Incidence rate ; 16.1.4 Ratio of policemen to total population; 16.3.1 Police Reporting Rate;4.2.1 Proportion of children 5 years of age who are developmentally on track in health, learning and psychological well-being, by sex; 4.3.1 Participation rate of youth in formal and non-formal education and training in the previous 12 months, by sex; and 16.6.2 Proportion of services (government offices with frontline services) that the public were satisfied or very satisfied in accessing public service. Out of these 11 proposed indicators, the first 5 are matched with their global indicators counterparts, the next 3 are proxy indicators, and the remaining 3 are sub indicators.

On the other hand, it can be seen that 25 proposed indicators were deemed not yet ready for institutionalization. These 25 indicators are: 1.4.1 Proportion of population living in communities with access to basic services; 2.5.2 Proportion of local breeds classified as being at risk, not-at-risk or at unknown level of risk extinction; 2.a.1 Agriculture orientation index for government expenditures; 2.c.1 Indicator of food price anomalies; 4.a.1 Proportion of schools with access to: (d) adapted infrastructure and materials for students with disabilities: (f) single-sex basic sanitation facilities, (g) basic handwashing facilities; 4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by level of skill; 4.4.1 Proportion of population with exposure to internet; 4.5.1 Parity indices (rural/urban,

bottom/top wealth quintile and others such as disability status, indigenous peoples and conflict-affected, as data become available) for all education indicators on this list that can be disaggregated; 4.b.1 Volume of official development assistance flows for scholarships by sector and type of study; 6.1.1 Proportion of families using safely managed drinking water; 6.2.1 Proportion of families using safely managed toilet facility; 8.9.2 Proportion of jobs in sustainable tourism industries out of total tourism jobs; 8.9.2 Proportion of small-scale industries in total industry value-added, 9.5.2 Researchers (in FTE) per million inhabitants; 9.5.2 Number of researchers (based on headcount) per million population; 9.b.1 Proportion of medium and high-tech industry value added in total manufacturing value added; 9.c.1 Proportion of number of cell sites, by technology; 9.c.1 Number of CMTS subscribers to 100 population, 9.c.1 Proportion of cities and municipalities with coverage of CMTS facilities, by technology; 11.1.1 Proportion of households with no access to improved sanitation; 11.4.1 Per capita consolidated annual budget of concerned cultural NGAs, as provided for in the GAA; 11.4.1 Growth rate of consolidated annual budget of concerned cultural NGAs, as provided for in the GAA; 11.6.1 Proportion of urban solid waste regularly collected and with adequate final discharge out of total urban solid waste generated, by cities; 11.6.2 Annual mean levels of fine particulate matter (e.g. PM2.5 and PM10) in cities (population weighted); and 16.10.1 Number of verified cases of killing, kidnapping, enforced disappearance, arbitrary detention and torture of journalists, associated media personnel, trade unionists and human rights advocates in the previous 12 months.

Majority of these 25 indicators were not pushed for institutionalization due to the lack of available data for the computation of the indicator. Discrepancies in the operationalized definition and lack of methodological research were also common reasons as to why the indicators weren't considered ready for adoption in the Philippines. As a result, several suggestions were made as the indicators' way forward before it can be institutionalized. It must be noted that the recommendations were given to hopefully benefit the further development of the indicators and for its better monitoring, planning, and policy making.

Out of 25 indicators, 10 of which were recommended to conduct further research. PSRTI recommended that these indicators need further methodological research to improve the way to generate the indicators. Further research may also involve reading more literature about the sector and dimension involved. Some of these indicators are the Indicator of food price anomalies, Proportion of jobs in sustainable tourism industries out of total tourism jobs, and Proportion of urban population living in slums, informal settlements, or inadequate housing. On the other hand, there were 5 indicators which were recommended to be consulted further with the

concerned IAC, NGA or PSA. These indicators still need to be discussed with the concerned stakeholders with regards to their operationalized definitions, its discrepancy with global definitions, and the possible ways to generate the data needed. Discussion on updating existing forms/databases to include the needed variables are suggested for the development of the indicators. Some of these indicators are Proportion of local breeds classified as being at risk, not-at risk or at unknown level of risk extinction; Agriculture orientation index for government expenditure; and Proportion of medium and high-tech industry value added in total manufacturing value added. Finally, there about 4 indicators that were suggested to review existing questionnaires used in institutionalized surveys so that it can include additional questions which will cater to the needed variables. For the indicator Proportion of youth and adults with information and communications technology (ICT) skills, by level of skill, it was suggested that the questionnaire for Functional Literacy, Education and Mass Media Survey (FLEMMS) was reviewed to accommodate questions about ICT skills.

Finally, the two indicators that were recommended for exclusion namely, 2.5.1 Number of plant and animal genetic resources for food and agriculture secured in either medium- or long-term conservation facilities and 2.b.1 Agriculture export subsidies, expectedly fell under the classification that it does not need to be operationalized anymore as it is of no relevance to the Philippine setting.

4. Conclusion

As a developing country, the Philippines is facing a great challenge in monitoring its sustainable development goals due to its still developing data ecosystem. In response to this challenge, the PSS, through the PSRTI, conducted a comprehensive study to identify other indicators which can be included in the current SDG monitoring system in 2017.

In summary, the research of the PSRTI resulted to a list of proposed indicators that corresponded to the global indicators set by the United Nations. Several of the indicators were exact matches to the global indicators while some were proxy or sub indicators. The review of these indicators led to recommendations on whether these are ready for Philippine institutionalization or not. Indicators that were good for institutionalization will be forwarded to the PSA for clearance, while the others are suggested for further research and discussion. Hence, research still has a big role to play in the country in order to fully expand and improve the current SDG indicator system for better monitoring of the sustainable development goal.

Collaborative arrangements in promoting the Sustainable Development Goals (SDGs) at the local level

Maria Praxedes R. Peña
Chief Statistical Specialist, Philippine Statistical Research and Training Institute (PSRTI), Quezon City, Philippines

Abstract

The long term objective of the Philippine government through its statistical system is to support the local government units in integrating the Sustainable Development Goals (SDGs) in local development plans and policies. The end result of this program would be the localization of the global goals with due consideration to the prevailing situation at the local level. Through this, local government units (LGUs) will be encouraged to set up their own indicator system with related data production program laid down to monitor locally the global goals that are existing in present situation and condition. The Philippine Statistical Research and Training Institute (PSRTI), in coordination with the Philippine Statistics Authority (PSA), National Economic and Development Authority (NEDA) and Department of Interior and Local Government (DILG) are expected to actively provide support to the local government units in putting up and maintaining said localized SDG indicator system. With this plan, the global goals will be made more relevant as they become part of the plans and programs at the lowest level of governance and at the same time monitored regularly with the use of local indicators that are developed in accordance with the prescribed statistical standards.

The paper will discuss the proposed collaborative efforts of the PSRTI with the different government agencies such PSA, NEDA and DILG in promoting the SDGs at the local level. The role that will be played by each agency in the endeavour to localize the SDGs will also be specified. In addition, the paper will discuss the activities already undertaken by PSRTI in pursuit of said goal.

Keywords

Localization of SDG, indicator system, Philippine statistical system (PSS), PSRTI

1. Introduction

In September 25, 2015, the United Nations (UN) adopted a global development agenda, referred to as the 2030 Agenda for Sustainable Development, which aims to promote holistic development, human rights, and equality for all regardless of distinctions of any kind. The agenda includes 17 Sustainable Development Goals (SDGs), which has 169 targets and adopted by 193 world leaders as universal goals for all countries to end poverty, protect the planet and ensure prosperity for all.

In the implementation of the SDG by countries, it primarily started at the national level and limited attention has been given to subnational SDG implementation. However, recognition of the need to localize the SDGs grew and witnessed by countries with a number of initiatives and discussions giving attention to the need to accelerate SDG implementation through increased efforts at the local level. There is a realization that SDG achievement at the national level strongly depends on progress made at the local level. This is done through the effective integration of the SDGs into the mandates of institutions and promoting cross-sector collaboration at all levels which makes for a multilevel governance and coherent SDG implementation.

Integrating the SDGs into the mandates of Local Government Units (LGUs) means SDG localization, which is defined as "the process of defining, implementing and monitoring strategies at the local level for achieving global, national, and subnational sustainable development goals." It is where local authorities and stakeholders will adapt and implement these targets within cities and human settlements. Subnational governments or LGUs are more than just implementers of the Agenda, but also policy makers and catalysts of change best placed to link the global goals with local communities. SDG localization is critical for achieving sustainable development by 2030 because SDG agenda may not fully be achieved without the involvement of urban and local actors.

2. Methodology: Collaborative Arrangements/Strong partnership for the Localization of SDG

Recognizing the importance of SDG localization, the Philippine government through its different agencies, namely the Department of Interior and Local Government (DILG), National Economic and Development Authority (NEDA), Philippine Statistics Authority (PSA), Philippine Statistical Research and Training Institute (PSRTI), Philippine Statistical System (PSS), all LGUs, and other institutions (policy and decision-makers), Civil Society Organizations (CSOs), media, development partners, academe and private sector/organizations, made collaborative arrangements in promoting the SDGs at the local level. PSS is composed of national government agencies that deal with statistics/statistical data, either as a data producer or data user. Its main task is to deliver quality statistical information to the public.

Roles of these agencies in implementing SDG are described in the Guidelines on the Localization of the Philippine Development Plan (PDP) 2017-2022 Results Matrices and the Sustainable Development Goals (SDGs), a Joint Memorandum Circular No. 01 Series of 2018 dated November 26, 2018 of NEDA and DILG. These efforts also integrated the SDG to the local plan - Philippine Development Plan (PDP) 2017-2022, which is geared towards the AmBisyon

Natin 2040 and anchored on the 0-10 Point Socioeconomic Agenda. The responsibilities of each agency are summarized as follows:

National Economic and Development Authority (NEDA)

- Spearheads the formulation of the Regional Development Plans (RDPs), with corresponding Results Matrices (RMs). RMs are documents which contain statements of the results to be achieved (goals, outcomes, and outputs) with corresponding indicators, baseline information, annual and end-of-plan targets, and responsible agencies. RMs includes identified performance indicators and sources of data/information that consider the SDGs.
- Assists LGUs in the integration of the SDG framework in regional and provincial planning & programming through the preparation of regional and provincial RMs and in accordance with their respective development plans and investment programs.
- Links the attainment of SDGs at the subnational level (region, province, city/municipality) to public policy
- Part of the advocacy program in raising awareness on SDGs and in taking an active role in the localization of the SDGs among LGUs

Philippine Statistics Authority (PSA)

- Develops a core regional indicator system composed of indicators that support development planning, implementation, and monitoring of local programs and projects
- Serves as official repository of SDG indicators in the country by establishing a team to facilitate coordination of monitoring of SDG-related activities.
- Develops and maintains a webpage on SDGs, SDG Indicator database and SDG Watch
- Part of the advocacy program in raising awareness on SDGs and in taking an active role in the localization of the SDGs among LGUs

Department of Interior and Local Government (DILG)

- Assists LGUs, together with NEDA, in the integration of the SDG framework in regional and provincial planning & programming through the preparation of regional and provincial RMs and in accordance with their respective development plans and investment programs.
- Presents the Community-Based Monitoring System (CBMS) indicators that provide a source of data to supplement indicators in the provincial RMs
- Part of the advocacy program in raising awareness on SDGs and taking an active role in the localization of the SDGs among LGUs
- Monitors performance of provinces vis-a-vis their set targets thru its Bureau of Local Government Development (DILG-BLGD).

LGUs in the provinces, cities and municipalities (which are under the administration of DILG)

- Integrates the SDG framework in planning & programming through the preparation of their respective development plans and investment programs.
- Translates respective development plans to regional and provincial RMs and SDG indicators
- Establishes targets using available provincial/local data
- Validates the baselines and set targets
- Provides data at the city and municipal levels to supplement data gaps identified by PSA. Local administrative data together with the Community-Based Monitoring System (CBMS) can be used in planning & target setting.
- Monitors contributions/performance vis-a-vis commitments within their geographic boundaries based on their respective cities/municipalities and provincial RMs

PSRTI

- Part of the advocacy program in raising awareness among LGUs to take an active role in the localization of the SDGs
- Undertakes methodological researches and capacity building activities to help PSA and other agencies generate the indicators needed in the planning, implementation, and monitoring of local programs and projects, including SDGs

Policy- and decision-makers (Sangguniang Panlungsod/Bayan representatives, City/Municipal Sectoral Representatives, City/Municipal Mayors, Provincial Governors, etc.)

- Links the relevance of SDG indicators to public policy, planning & programming
- Monitors and links the attainment of SDGs at the subnational level (region, province, city/municipality) to public policy

Data producers and other agencies that produces data (CBMS, PSS, etc.)

- Compiles and generates data at subnational levels to supplement data gaps and for LGUs to use in their planning and programming, target setting, monitoring and evaluation of SDGs

Civil Society Organizations (CSOs)

- Compiles and generates data at subnational levels to supplement data gaps and for LGUs to use in their planning and programming, target setting, monitoring and evaluation of SDGs
- Part of the advocacy program in raising awareness on SDGs and take an active role in the localization of the SDGs

Media

- Part of the advocacy program in raising awareness on SDGs and take an active role in the localization of the SDGs

Development partners (ADB, World Bank, UN Organizations, PARIS 21, etc.)

- Provides funding support and technical assistance

Academe

- Undertakes methodological researches to help PSA and other agencies generate indicators needed in the planning, implementation, and monitoring of local programs and projects, including SDGs

Private Sector/Organizations

- Conducts activities and provide investments to support achievement of the SDGs

3. Conduct of the Training Program through Collaborative Arrangements/Strong Partnership

In pursuit of its mandate and fulfilment of its responsibility in the localization of SDG, PSRTI proposed a training program for the LGUs and selected employees of the PSS. In conducting the training program, PSRTI also adopt the concept of strong partnership of agencies gained from the collaborative arrangements among DILG, NEDA, PSA, and all LGUs.

The purpose of the training program is to teach concepts and methods of basic statistical measures used in computing SDG indicators and collection of local administrative data as means of generating disaggregated data for computing various SDG indicators at subnational level. Specifically, the objectives of the training program are: (i) raise awareness on SDG with focus on localization; (ii) appreciation of basic statistical measures used in assessing attainment of SDGs; (iii) computation of indicators needed in the planning, implementation, and monitoring of SDGs; and (iv) collection of local administrative data, including CBMS data to provide data at subnational level that can be used to measure performance/implementation, and monitoring of SDGs.

The training program is composed of two (2) training courses describe as follows:

A. **SDG Seminar 1.** A one-day Appreciation Course on SDG Indicators, Data Ecosystem, Basic Statistics to be used in assessing attainment of SDGs and survey on SDG indicators being collected by the different local government units. The target training participants are 1,634 municipal and city planning officers, and DILG, PSA and NEDA selected employees from 17 regions of the country. This Appreciation Course will be conducted from 8:00 a.m. to

5:00 p.m. (8 hours). This training course will be conducted for the 17 regions of the country from 2018-2019.

The course outline for this covers the following:

I. Overview of the SDG

II. Presentation of the Data Ecosystem

III. Lecture on the Basic Concepts in Statistics

1. Definition and Examples of Official Statistics
2. Importance of Statistics
3. Basic concepts about variable, indicator and data
4. Some Summary Measures used in Computing SDG Indicators
5. Methods of Data Collection

IV. Workshop on SDG Matrix

1. Mechanics of the Workshop and computation of some SDG Indicators
2. Open Forum

B. **SDG Seminar 2.** A three-day training course on descriptive statistics, data collection and the Community Based Monitoring System (CBMS), and monitoring and evaluating the different SDG indicators using statistics. The target training participants are 3,268 city planning officers, municipal planning officers, K to 12 principals and teachers from the 17 regions. The trainings will be for 3 consecutive days, targeting four batches per region with a class size of 50 participants. This training course will be conducted from 8:00 a.m. to 5:00 p.m. (24 hours). Knowledge on basic computer operations is required and data management using MS Excel® is a pre-requisite to this course.

The course outline for this covers the following:

I. Introduction and Review on basic concepts in Statistics

II. Methods of Collecting Data and CBMS

1. Survey
2. Use of Documented Data
3. Registration
4. Questionnaire Construction
5. CBMS and its instrument/questionnaire

III. Summary Measures that will be used in planning, implementing, monitoring and evaluating the SDG indicators

1. Measures of Central Tendency
2. Measures of Location
3. Measures of Dispersion
4. Proportions, Ratios, Rates, Percent Change

IV. Monitoring and Evaluation (M & E) of performance vis-a-vis targets 1. Processing and analyzing M&E Data

V. Methods of Presenting Data
1. Textual, tabular and graphical

4. Discussion and Conclusion: Training Evaluation and Next Steps

For 2018 SDG Seminar 1 were conducted for eight (8) regions, namely: Region III, Region IVA, Region VII, Region XI, Cordillera Administrative Region (CAR), Autonomous Region in Muslim Mindanao (ARMM), Region VI, Region V and National Capital Region (NCR). These were conducted successfully through collaborative arrangements/strong partnership among DILG, NEDA, PSA, PSRTI and all LGUs despite the absence of PSRTI regional and provincial offices. DILG was tap to intensify the invitation to planning development officers of various LGUs at provincial, city and municipal levels and for these participants to actively participate in the seminar. PSA focal persons were also invited to serve as resource persons on SDG overview and Data Ecosystem while NEDA took a significant role in setting the tone of the event by reminding the LGUs of their commitments in achieving the SDGs that are embodied in their respective Regional Development Plans.

Each **SDG Seminar 1** started with an opening ceremonies which includes introduction of PSRTI as the lead agency in conducting the SDG Seminar 1, its partner agencies and participants; and an overview of the seminar. Participants were provided with flash drives containing pdf copies of all lecture materials, excel file of the SDG matrix for city/municipalities as well as the SDG CBMS indicators. For the workshop, the participants were grouped per city/municipality to accomplish SDG Matrix per city/municipality, which inquired about: data availability, frequency and data source per SDG indicator, whether it is or is not monitored in their area, and where it is used for. All LGUs were asked to submit the accomplished SDG Matrix per city/municipality. Summary on SDG matrix submission per cities/municipalities by region is presented in Table 1.

Table 1. Summary on the number of cities/municipalities with SDG matrix submission

Region	No. of cities and municipalities	No. of cities/ municipalities with SDG matrix Submission	% of submission
NCR	17	7	41.18
CAR	77	3	3.90
Region I	125	85	68.00
Region III	93	62	66.67
Region IVA	130	67	51.54
Region V	114	45	39.47
Region VI	133	70	52.63
Region VII	132	39	29.55
Region XI	49	38	77.55
ARMM	118	3	2.54

Significantly low SDG matrix submission rate were noted in CAR and ARMM. Reasons identified for low SDG matrix submission rates are: (1) affected by typhoon during the conduct of the seminar (CAR) and (2) independence/autonomy of the region which greatly affected the collaborative arrangements in the conduct of the seminar (ARMM).

Evaluation about the seminar was also administered. This evaluation inquired about the level of satisfaction in the training program, its content, duration, relevance to work, materials/handouts, presentation technique and time allotted for questions/discussion. The level of satisfaction for these items has a rating ranging from satisfactory to very satisfactory. There were recommendations on lengthening the time allocation for workshops as most participants find the exchange of opinions and ideas relevant to their own situations and challenges they face in planning, programming and target setting.

For 2019, SDG Seminar 1 will be conducted for the remaining eight (8) regions, namely: Region II, Region IVB, Region VIII, Region X, Region XII, Region IX and Region XIII. The conduct of these seminars will adopt the same collaborative arrangements with DILG, NEDA, PSA and LGUs.

For SDG Seminar 2, the training course will be conducted for the 17 regions of the country. Given the constraint in the available staff of PSRTI, the capacity building program will be conducted for four years (2020-2023). For 2020, training courses will be conducted in NCR, CAR, Regions V, VI, and XII; with a total of 20 trainings. Continuation of the training for the rest of the regions not covered in 2020 will be done in 2021 for Regions IVA, IVB, VII, and VIII with a

total of 16. For 2022 Regions I, II, IX, and XI will be covered with a total of 16 trainings, while for 2023, Regions III, X, XIII, and ARMM will be covered, with a total of 16 trainings. These training courses will also be conducted through strong partnership with DILG, NEDA, PSA, and LGUs.

PSRTI intends to actively participate in the achievement of the global goals through strong partnership with agencies through the conduct of capability building initiatives. The focus on statistics will be helpful in the monitoring and evaluation stages of attaining SDG indicators in the subnational level.

References

1. Global Taskforce of Local and Regional Governments, UNDP and UN Habitat (2016). Roadmap for Localizing the SDGs: Implementation and Monitoring at Subnational Level.
2. Pytrik Dieuwke Oosterhof (2018). Localizing the Sustainable Development Goals to Accelerate Implementation of the 2030 Agenda for Sustainable Development. The Governance Brief. Asian Development Bank.

Evolving statistics education for a data science world

Alison L. Gibbs, Sotirios Damouras
University of Toronto, Canada

Abstract

Statistics is undergoing a period of profound transformation, disrupting its practice, research, and teaching. The ubiquity of data and the data-driven scientific paradigm have fundamentally changed the way we analyse data and have placed new demands on statistics education. There is a recognised urgency for undergraduate statistics curricula to include the development of software and programming skills for working with data, and systematic efforts are already under way to address this. Beyond ensuring the acquisition of practical skills, this transition period offers the opportunity to realign the principles and focus of statistics education for the future. In particular, students need to develop new traits and attitudes that will support their ongoing academic and professional development in the age of data science. We propose a set of qualities and higher-order skills that we believe are essential for our graduating students to remain relevant during the evolution of Statistics, and we describe practical strategies for fostering these in the context of introductory courses in statistical reasoning and data science at our institution.

Keywords

statistics education; introductory course; adaptive expertise; lifelong learning

1. Introduction

The ongoing "Data Revolution," fuelled by the increasing availability of both data and computation, is transforming Statistics. This revolution has created a new land of opportunity, what Cobb (2015) aptly describes as "the valuable territory [of] the science of data." Statistics used to have almost exclusive rights on areas concerned with extracting knowledge from data, but now finds itself contending with new fields such as Machine Learning, Data Mining, and Analytics. While realizing that this territory is too expansive to claim sole ownership, our discipline is trying to identify its place and purview in this new environment. There is a commonly expressed understanding that the best way to ensure Statistics does not become marginalized is to engage with other disciplines. The American Statistical Association (ASA) issued a statement on the role of Statistics in Data Science (van Dyk et al., 2015) calling for a "sustained and substantial collaborative effort" in which "statisticians

must engage, learn from, teach, and work with [researchers with expertise in data management and computation]." Embracing computational thinking and applications, while upholding our traditional strengths, is an important prerequisite in this effort. In their epilogue, Efron & Hastie (2016) give a historical summary of the shifting focus in Statistics between applications, mathematics, and currently computation, all the while maintaining a principled approach. Although excessive preoccupation with mathematical underpinnings has led our discipline to periods of introversion, Efron & Hastie also identify important recent developments fuelled by applications and computation, the type of developments upon which the sustained success of Statistics will ultimately be judged.

Parallel to asserting Statistics' place in this new world of Data Science, there is another effort to attract and educate the future generations of statisticians who will populate it. The current hype around the Data Scientist profession, with Glassdoor ranking it "Best Job in America" for the 4th year in a row (https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm), is fuelling unprecedented growth in Statistics programs' enrolments (Pierson, 2017). But the rise of Data Science has also placed new demands on undergraduate statistics education and provided new impetus for its reform. Horton & Hardin (2015) recognize that "the traditional statistics curriculum with mathematical foundations has not kept up with pressing demands for students who can make sense of data" and call for curricula that "prepare students to engage in the entire data analysis process." As envisioned in the most recent ASA guidelines (2014), this curriculum should provide a balance between mathematics, computing, and applications, while offering many opportunities for practice and skill-building. But even with such bold reforms, the rapidly changing landscape of Data Science suggests our graduates will face conditions they have not seen before. We claim that alongside teaching new content and skills, our programs should also cultivate a new mindset, one that will prepare students to learn on their own and apply their knowledge flexibly. We propose a set of three traits and attitudes, namely inquisitiveness, extroversion, and statistical thinking, which comprise this "adaptive statistical mindset."

2. Teaching Statistics in View of Data Science

Accompanying the enrolment growth in Statistics programs and the proliferation of programs of study in Data Science (see, for example, NASEM, 2018), new and modified guidelines for both types of programs have been created. These guidelines reflect the evolving understanding of the knowledge and skills needed to learn from data in our current context. While statistical thinking remains a core feature, the guidelines include calls for enhancing computational skills to deal with larger and more complex data, greater

attention to algorithmic approaches, and increased emphasis on skills needed for professional practice, including communication and collaboration. Some key features of the guidelines are as follows:

Revisions to the Guidelines for Assessment and Instruction in Statistics Education (GAISE College Report ASA Revision Committee, 2016) for the first course in statistics have increased emphasis on reasoning with multivariate data and on engagement in a complete investigative problem-solving process. In 2014, the ASA endorsed new guidelines for the curriculum of undergraduate programs in statistics (ASA Undergraduate Guidelines Workgroup, 2014), aiming to ensure continued relevance of graduates from such programs. In comparison with previous ASA curriculum guidelines, the 2014 guidelines have increased emphasis on analyzing complex data, modelling for prediction, and computing, including the skills required to access and process complex and large datasets. They also underscore the importance of developing the skills needed to engage in statistical problem-solving applied to questions from other domains.

More recently, guidelines for undergraduate programs in Data Science (De Veaux et al., 2017 and NASEM, 2018) emphasize the integration of skills from statistics, computer science, and mathematics, focused on the aspects of these fields that are important for learning from data.

All three guidelines promote the importance of involvement in the complete statistical process, including formulating questions, acquiring suitable data, analyses, communication of results, and critical assessment of each step that may lead to iterations of the process. Many of the emphases in the guidelines are not new. For example, there has been a long-standing conversation about the need for our students to develop a broad set of non-technical skills to enable them to become effective contributors to the solutions of problems in a variety of applications. Utts (2015) described four themes that regularly appear in initiatives related to statistics education in the ASA over its 175-year history. Among these is the need for statisticians to develop "soft skills," including the ability to communicate results to non-technical audiences and function effectively as part of a team. Building on these calls for reform and considering the rapid changes in the field, in the next section we offer another perspective on teaching Statistics for an evolving world.

3. Teaching Statistics for an Evolving World

The demands imposed by Data Science are transforming the way we teach statistics at the undergraduate level, with several initiatives highlighted in the previous section. But as Data Science continues to evolve, and its practice continues to change so quickly, no program of study in Statistics or Data Science can hope to teach all the knowledge and skills that students will need

in the future (Zheng, 2017). New fields of application are constantly emerging, dealing with increasingly complex data such as high-dimensional sequence or network data, and generating new types of problems that statisticians are called to address. This shifting landscape makes it difficult for statistics educators to foresee students' needs, let alone cater to them through curricular reform, a typically slow and gradual process. Instead of trying to anticipate every development, we believe our programs should prepare students to be able to adapt to these changes on their own. We borrow the term "adaptive expertise" from the fields of cognitive development and pedagogy to refer to such desired behaviour.

The concept of adaptive expertise was introduced by Hatano & Inagaki (1986) to describe the ability to aptly address new types of problems, as opposed to routine expertise which is focused on efficiently solving familiar problems. It is closely related to the concept of transfer of learning, i.e. "the ability to extend what has been learned in one context to new contexts" (NRC, 2000), but we prefer the term adaptive expertise as it encompasses both knowledge and practice. We first look at a general framework for attaining adaptive expertise, one that also offers an interesting lens on Statistics Education, before going on to propose specific strategies.

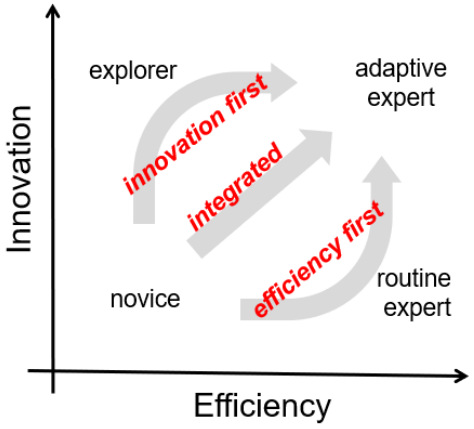


Figure 1. *Trajectories to Adaptive Expertise through the Innovation-Efficiency space*.

Adaptive expertise calls for superior performance in novel situations, something that requires domain-specific proficiency as well as general inventiveness. Schwartz, Bransford, & Sears (2005) discuss the trade-offs involved in developing such skills with the help of a two-dimensional space of innovation and efficiency, similar to the one presented in Figure 1. They consider three possible trajectories to adaptive expertise: one which prioritizes efficiency, one which prioritizes innovation, and one which balances the two; they call the latter the "optimal adaptability corridor," but we use the simpler term "integrated" trajectory. They observe that formal education has typically

been geared towards efficiency, i.e. the ability to readily solve routine problems. This certainly rings true for the traditional undergraduate statistics curriculum, with its emphasis on set-piece methodological solutions. The typical undergraduate student first acquires procedural skills and knowledge, and then tries to develop adaptive expertise through projects, capstones, internships, or, eventually, work experience. On the other hand, Schwartz, Bransford, & Sears note that content-free activities in critical thinking and problem-solving (i.e. innovation) help develop the relevant skills, but these might be insufficient for tackling larger, more complex problems. They advocate for a balanced approach, one that supports simultaneous development along both dimensions. Moreover, they conjecture it is not enough to have parallel but separate content and thinking courses, but that the two should be interlaced.

Adopting an integrated approach to developing adaptive expertise, we look at specific ways to attain it by progressing along the innovation dimension. Hatano (1988) lists three conditions that support the development of adaptive vs routine expertise: a) students should continuously encounter novel types of problems, b) they should be encouraged to seek comprehension, and c) they should be allowed to experiment without a pressing need for rewards. Other researchers have built upon these ideas, especially in the context of professional education such as medicine or engineering. Carbonell et al. (2014) provide a comprehensive review of characteristics and learning environments for adaptive expertise, while Mylopoulos et al. (2018) offer 12 practical recommendations for designing curricula that support it. Alongside Hatano's three general principles and related suggestions, we propose three traits and attitudes that are distinctive to Statistics; these are:

1. *Inquisitiveness*: We should instil into our students an intellectual curiosity and spirit of inquiry. Statistics is the science of extracting knowledge from data, and to do so students must be inclined to ask meaningful and interesting questions and be willing to pursue their answers outside of conventional beliefs and practices.
2. *Statistical Thinking*: We should provide our students with a general conceptual framework through which they approach problems. We use this familiar term to describe a way of thinking, rather than just a set of procedures and tools. Although a formal definition is elusive, we follow Chance's (2002) description as "the ability to see the process as a whole (with iteration), including 'why,' to understand the relationship and meaning of variation in this process, to have the ability to explore data in ways beyond what has been prescribed in texts."
3. *Extroversion*: Statistics is by its nature an outward-facing discipline, one that draws energy and stimulation from its interactions with other

disciplines. We believe this fact should be reflected in the way we teach Statistics. We should prepare and encourage our students to engage in interactions, by cultivating, among other things, their collaboration and communication skills.

We believe that, collectively, these three traits and attitudes comprise an *adaptive statistical mindset*, i.e. a disposition and way of thinking that is essential for the independent and continued development of learners and practitioners in the field. These proposed traits and attitudes parallel the entire statistical process, from planning, to analysis, to result dissemination. The call for statistical thinking is well-documented and goes back to at least Cobb (1992). The other two characteristics are complementary to the statistical process, in that they incite and advance its development. All three include higher-order thinking skills which, although difficult to teach, should nevertheless be integrated and cultivated throughout the curriculum. In the next section, we present our approach for doing so in the context of introductory Data Science courses, showcasing learning activities and environments that promote such a mindset.

4. Teaching with an Adaptive Statistical Mindset

To illustrate our approach at cultivating an adaptive statistical mindset, we describe some elements of introductory courses in Data Science at two campuses of the University of Toronto. Introductory courses in a discipline can serve many purposes, including acquainting students to its ways of thinking, attracting them to future study in the discipline, and preparing them for more advanced work through teaching foundational knowledge. Our courses are very much in the spirt of Wild's (2015) call for a "further, faster, wider" approach to the introductory course, showcasing a wide range of exciting problems that can be tackled with statistical reasoning.

Table 1 indicates some of the strategies we use in our courses to cultivate specific components that comprise our proposed adaptive statistical mindset. In order to start students on a trajectory to developing adaptive expertise, we have designed our courses to begin the development of innovation and efficiency in an integrated fashion, while cultivating inquisitiveness, statistical thinking, and extroversion. Below the table we give some details of corresponding class activities.

Table 1. *Strategies for cultivating an Adaptive Statistical Mindset*

Strategies \ Component	*Inquisitiveness*	*Statistical Thinking*	*Extroversion*
Open-ended investigations	X	X	X
Complex, real-world data	X	X	
Collaborative work			X
Facilitated problem solving		X	X
Authentic assessment	X	X	

- *Open-ended statistical investigations*: Students in our courses carried out team projects. Key features of these projects include purposely vague, ambiguous questions and socially significant contexts. As a recent example, students were provided with counts of incidents of harsh braking, accidents, and near misses in motor vehicles, and the corresponding traffic flow and location. Students were tasked with comparing hazardous driving among locations. To do this they first needed to consider how they might define "hazardous driving." The marking scheme rewarded exploration, including careful consideration of competing approaches, and innovation, particularly extending the course concepts. The data were provided by a local company (www.geotab.com) and company data scientists introduced the project and visited the poster fair where students presented their findings.
- *Complex, real-world data*: Concerted effort has been made throughout all aspects of our courses, lectures, practice problems and assessments, to expose students to a variety of data collected for a variety of purposes with rich contexts. To incite students' curiosity, we tried to use data that were as close to their experiences and interests as possible. For a course project we gave students the question "Is university education worth it?", which they were free to tackle from any angle they chose. Students were pointed to survey microdata (e.g., Labour Force Survey, National Graduates Survey, Canadian Income Survey, Census) collected from Statistics Canada, the national statistical office, and accessed through our institution's data library. They had to explore these rich data sources, potentially combining them with other ones, and extract relevant information for addressing their questions.
- *Collaborative work*: Collaborative work was an integral expectation of projects and problem solving in our courses. Students were given time to work in teams in informal environments, in which the teaching team supported student investigation and experimentation and encouraged reflection on the process. Teaching assistants arranged students in

different teams each week, to promote the acquisition of strategies to effectively work with others.

- *Facilitated problem solving*: Throughout our courses, students solved regular practice problems in a facilitated manner. They were given worksheets in the form of electronic notebook documents, which they completed with the help of teaching staff and their peers. These worksheets were designed as low-stakes formative assessments, geared towards building skills and understanding concepts. This format is particularly amenable to what-if type questions; e.g., it allowed us to use simulation to explore important concepts such as p-hacking and overfitting. An additional benefit was dynamic two-way feedback, with the instructor experiencing first-hand where students had difficulties and providing immediate help.
- *Authentic assessment*: Summative assessments for one course were performed on computers, in the same software environment used throughout the course. We tried to assess students in a way that was as close as possible to how they would analyse data in real life. Students were given a set of yet unseen data and asked to perform specific analytic tasks, as well as answer conceptual, interpretation, and open-ended questions. At the end of the exam, students submitted individual reports combining their code, results, and answers in a reproducible manner. This format afforded us the flexibility to make assessments that are aligned with what we value.

More details on these strategies, together with material from our respective courses, can be found at sta130.utstat.utoronto.ca and utsc.utoronto.ca/~sdamouras/staa57. These courses have been designed to start students on the trajectory to developing an adaptive statistical mindset. Continued progression on the trajectory requires integration of novel problems, learning that emphasizes understanding, and exploration and discovery throughout our programs of study. For an overview of a complete program of study designed with such considerations, see Gibbs (2018).

5. Conclusion

Statistics curricula have consistently ensured our graduates are highly skilled and efficient at solving standard problems in familiar settings. The requisite procedural knowledge and skills for doing this are sufficient in stable environments. But our discipline has transformed rapidly with the advent of Data Science. In light of this transformation, statistics educators have put a great deal of thought and effort in updating guidelines, programs, and courses. Along with teaching new knowledge and skills, we need to prepare students to be able to respond to ongoing change. For this purpose, we have proposed a focus on an adaptive statistical mindset, one characterised by

inquisitiveness, statistical thinking, and extroversion. People with this mindset engage in problems with intellectual curiosity, approach solutions with mature statistical thinking, and are inclined to contribute to other domains. For illustration, we gave examples of how we are initiating the development of such a mindset in introductory courses at our institution.

References

1. American Statistical Association (ASA) Undergraduate Guidelines Workgroup. (2014). 2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science. Alexandria, VA: American Statistical Association.
2. Carbonell, K.B., Stalmeijer, R.E., Könings, K.D., Segers, M. & Van Merrienboer, J.J.G. (2014). How Experts Deal with Novel Situations: A Review of Adaptive Expertise. *Educational Research Review*, **12**.
3. Chance, B.L. (2002). Components of Statistical Thinking and Implications for Instruction and Assessment, *Journal of Statistics Education*, **10**(3).
4. Cobb, G. (2015). Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up. *The American Statistician*, **69**(4), 266-282.
5. Cobb, G. (1992). Teaching Statistics. In Heading the Call for Change: Suggestions for Curricular Action, Ed. Steen, L. A. MAA Notes, No. 22, 3-34. Washington, DC: Mathematical Association of America.
6. De Veaux, R.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.X., Washington, T.M., Wesley, C.L., White, D., & Ye, P. (2017).
7. Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and its Applications*, **4**, 15-30.
8. Efron, B. & Hastie, T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. New York: Cambridge University Press.
9. GAISE College Report ASA Revision Committee. (2016). Guidelines for Assessment and Instruction in Statistics Education, College Report 2016.
10. Gibbs, A.L. (2018). Building a Foundation in Statistics in the Era of Data Science. In Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July 2018), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.
11. Hatano, G. & Inagaki, K. (1986). Two courses of expertise. In Child development and education in Japan, Eds. Stevenson, H. W., Azuma, H. & Hakuta, K., 262-272. New York, NY: W H Freeman.

12. Hatano, G. (1988). Social and motivational bases for mathematical understanding. *New Directions for Child and Adolescent Development*, 1988 (41), 55–70.
13. Horton, N.J. & Hardin, J.S. (2015). Teaching the Next Generation of Statistics Students to "Think With Data": Special Issue on Statistics and the Undergraduate Curriculum, *The American Statistician*, **69**(4), 259-265.
14. Mylopoulos, M., Steenhof, N., Kaushal, A. & Woods, N. (2018). Twelve tips for designing curricula that support the development of adaptive expertise. *Medical Teacher*, **40**, 1-5.
15. National Academies of Sciences, Engineering, and Medicine (NASEM) (2018). Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press.
16. National Research Council (NRC). (2000). How People Learn: Brain, Mind, Experience, and School. Washington, DC: The National Academies Press.
17. Pierson, S. (2017). Bachelor's, Master's Statistics and Biostatistics Degree Growth Strong Through 2016. *AmStat News*, October 2017.
18. Schwartz, D. L., Bransford, J. D. & Sears, D. (2005). Efficiency and Innovation in Transfer. In Transfer of Learning from a Modern Multidisciplinary Perspective, Ed. Mestre, J. Greenwich, CT: Information Age Publishing.
19. Utts, J. (2015). The Many Facets of Statistics Education: 175 Years of Common Themes. *The American Statistician*, **69**(2), 100-107.
20. van Dyk, D., Fuentes, M., Jordan, M., Newton, M., Ray, B.K., Temple Lang, D. & Wickham, H. (2015). ASA Statement on the Role of Statistics in Data Science. *AmStat News*, October 2015.
21. Wild, C. (2015). Further, Faster, Wider. Online discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum From the Ground Up," by George Cobb. *The American Statistician*, **69**(4).
22. Zheng, T. (2017). Teaching Data Science in a Statistical Curriculum: Can We Teach More by Teaching Less? *Journal of Computational and Graphical Statistics*, **26**(4), 772-774.

Looking back – looking forward; statistics and the data science tsunami

Jim Ridgway, James Nicholson, Rosie Ridgway
School of Education, University of Durham, UK

Abstract

The discipline of statistics arose from pressing needs to address a variety of social and scientific problems. The founders of the Royal Statistical Society in the UK, and the American Statistical Association were very diverse in their backgrounds and interests, but shared a common purpose – namely, to address difficult and interesting challenges. They also acted in similar ways, by working across disciplines, and inventing mathematics and models suited to new problems. Computer scientists have also addressed real-world problems, have pioneered interesting and exciting approaches to handling new sorts of data (e.g. from sensors and social media) and have developed new analytic tools (notably, tools based on machine learning); their work is having dramatic (and sometimes unexpected) impacts on society. Early encounters between statisticians and computer scientists often resembled 'turf wars' – with claims that statistics was fast becoming redundant, and that computer scientists' ignorance of core statistical concepts such as sample bias would prove fatal to their entire enterprise. The problems that beset the start of the twentieth century have not gone away; modern societies face a wide range of existential threats such as global warming and nuclear war. As before, collaboration across disciplines, and the creation of new modelling tools are needed to address these problems. Here we begin by drawing lessons from the development of computer science in its earliest days, focussing on Babbage's Analytical Engine. We then highlight key epistemological differences between traditional statistics and traditional computer science, such as the role of theory and the use of 'black-box' models. We argue the case for the development of the Epistemological Engine – a tool for analysing and improving the processes of knowledge creation and utilisation that will require the skills of both statisticians and data scientists. We conclude by identifying competences and dispositions relevant to students of statistics and data science, drawing on both contemporary developments and the earliest days of computing.

Keywords

Modelling; Turf wars; Epistemology; Black-box; engineering

Those who view mathematical science not merely as a vast body of abstract and immutable truths, whose intrinsic beauty, symmetry and logical completeness... entitle them to a prominent place in the interest of all profound and logical minds, but as possessing a yet deeper interest... this science constitutes the language alone through which we can adequately express the great facts of the natural world... will regard with especial interest all that can tend to facilitate the translation of its principles into explicit practical forms. Lovelace (1843, p2).

1. Lessons for young minds from histories

Let us visit Victorian England for some lessons for young minds – these include lessons that go beyond sciences and technologies themselves. We derive our lessons from the lives of Charles Babbage (CB) and Ada Augusta King, Countess of Lovelace (AL) in the period 1833-1852. CB is often credited as the designer of the first computer; AL as the first programmer. CB's Difference Engine was designed to create tables of numbers relevant to astronomy, navigation and mathematics, and was based on calculating successive terms in a given series using the method of differences which was built into a mechanical system of cogs and levers. His even more brilliant insight was the idea of a general purpose device with a store, a mill (CPU), a printer, and operation cards (for the program and numeric input) that would not need to be reset mechanically for each new table. Both the Difference Engine and the Analytical Engine were to be driven by steam. CB received huge amounts of state funding for the Difference Engine (more than the cost of building 2 battleships (a measure of research funding no longer used in the UK)). He realised that a successful Analytical Engine would do everything and more than the Difference Machine was capable of, and so devoted his energies to the Analytical Engine. A fully-functioning version of the Difference Engine was never built. The development of the Analytical Engine was never properly funded.

- Technologies change – and contemporary technologies can present barriers to brilliant ideas
- Funding streams depend on delivering what you (more or less) promised

If the state of technology was a limitation, what of the state of mathematics? Hot topics of the day included early explorations of non-Euclidean geometry (imagine that! – but why bother?), and imaginary numbers (again – surely impossible?). William Frend was a university mathematician and sometime tutor of AL who did not believe in negative numbers; he ridiculed the idea of zero. Along with some other mathematicians, he rejected the idea of using undefined symbols in algebra. Boolean algebra was first set out in Boole's (1854) The Laws of Thought – two years after the death of AL. By way

of balance, CB held the most prestigious chair in mathematics in England; AL was a talented mathematician in her own right. These provide the conflicted context for AL's extraordinary insight that a machine could manipulate arbitrary symbols, and that these arbitrary symbols could refer to anything. "The distinctive characteristic of the Analytical Engine... the executive right-hand of abstract algebra... is in this that... [it]... weaves algebraical patterns like the Jacquard loom weaves flowers and leaves" (Lovelace, 1843, p3). "Supposing... the science of harmony and musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent." (Lovelace, 1843, p2).

- *Don't always believe what your tutors tell you cannot or should not be done*
- *Become an excellent mathematician*
- *Explore the art of the possible*

Women's scope for action in Victorian England was severely limited. Concerns were raised about the serious dangers to the mental health of both Mary Somerville (the mathematician after whom Oxford University's first college for women was named) – friend and correspondent on things mathematical with AL – and for AL herself, from studying mathematics to a high level. AL published just one paper. This paper began with a translation of Menabrea's (1842) paper on a talk by (his friend) Babbage in 1840 about the Analytical Engine. At the time, few women wrote original articles, but DID occasionally translate and summarise men's work. Babbage suggested she write notes on the article (and says she was responsible for "the algebraic working out of the different problems" (Babbage, 1864, para 136). Babbage himself published extensively on a wide variety of subjects. He discussed the Analytical Engine ad *nauseam* with friends and colleagues, but published little about it (despite huge volumes of notes and diaries).

- *Don't accept cultural norms that restrict your thoughts and actions*
- *Publish!*

What of the social dimension? CB (and therefore AL) was friendly with Michael Faraday, Charles Darwin, The Herschels, Mary Somerville, Augustus De Morgan, Florence Nightingale, Elizabeth Gaskell, Tennison, and the Duke of Wellington (and sundry other politicians).

- *Cultivate smart people from a wide range of disciplines*
- *Travel and learn*
- *Talk about your ideas*

So much for computer science. What of the history of statistics? The logo of the Royal Statistical Society (RSS) is a wheatsheaf, carried over from its predecessor the Statistical Society of London (SSL). The SSL originally adopted the motto *aliis exterendum* – to be threshed out by others. This conveys a clear

world view from the world's first professional association of statisticians – that the primary function of statistics is to gather and organise resources that others will transform into something useful. Pullinger (2013) paints a very different picture. The motto was dropped after a year. He points to the diversity of the founders of the RSS (which included CB – mathematician, mechanical engineer, astronomer, and philosopher) and to their commitment to study practical problems and to find (and implement) solutions with direct social benefit – inventing new mathematics when needed. This tension between gatherers and analysts, and between theoreticians and practitioners, articulated by Lovelace in the introductory paragraph, mirrored in both mathematics and statistics, is alive and well.

It is captured in some critiques of statistics curricula. Cobb (2015) and Ridgway (2015) argue that introductory courses over-value tractable statistical models, resist algorithmic thinking, and devote far too little time to realistic problems. This critique begs two questions: 'whose realistic problems?'; 'what models are missing'? In the early days of the RSS, the answer to the question about 'whose problems' might well have been 'everyone's' – illustrated via pioneering work in meteorology, health, genetics, agriculture and economics, and often associated with the invention of new mathematics. The extent to which this tradition of conducting pioneering work with practical applications, and inventing appropriate supporting mathematical structures, has continued can be judged by inspecting the list of past RSS presidents (see RSS, 2019).

The question of 'missing models' raises bigger issues. All models are simplifications of some reality, and the choice and applicability of any model depends on the phenomenon to be modelled, and the purpose to which the model will be put. "All models are wrong, but some are useful" (Box and Draper, 1987, p424). A problem with introductory statistics courses has been an over-emphasis on standard models (e.g. using the Normal distribution) developed to solve problems in a pre-computer age, and a focus on generalising from samples to populations. This is appropriate where data is expensive to collect, where small samples can represent populations (often the case in agriculture and medical trials - but not in situations where disaggregated data show different patterns), and where phenomena are stable over time (again, agriculture and some medical trials, but not social phenomena over time), and where there is little computational power. Even in favourable circumstances, models can be applied badly – see Ioannidis (2005) on *why most published research findings are false* and the Open Science Collaboration (2015) on failures to replicate 'well-known' results in psychology. These failures constitute a serious threat to the business of creating new and useful knowledge, and advancing progress in a number of academic disciplines. The failures themselves can be traced to poor practices of data collection, analysis and interpretation, which can be recognised, and remedied.

2. Conflicting epistemologies

Breiman (2001) described two approaches to analysing data. He argued that most statisticians typically apply transparent models where a small collection of well-defined inputs are used to predict outputs - so models are used primarily to explain and also to predict (he argues that this leads to irrelevant theory and questionable conclusions). In contrast, a small proportion use algorithmic modelling; techniques such as neural nets and random forests are used to map inputs and outputs. The focus is primarily on prediction with little attempt to explain. This can be viewed as a 'data science' stance.

Ridgway et al (2018) map out some challenges for algorithmic models – notably that what you get out is determined by what you put in. So algorithmic models are strong on 'what is' but weak on 'what ought to be' and can have undesirable consequences when used for (for example) job selection or predictive policing. Perez (2019) provides further examples. These problems are exacerbated when the data set itself does not represent the population as a whole – for example drawing conclusions from (conventional) medical research that is based almost exclusively on Caucasians. This is a particularly problematic challenge for data science, where decisions about analysis are often based on pragmatism; a variety of models are applied to a data set, and the final choice of model is based on fit and the ability of the model to predict future events.

Statistics has been characterised by engagement with real-world problems; what of data science? Consider these examples of computer uses, software and devices:

- Google, Amazon, Facebook, Skype;
- nrecognition of individuals via face, fingerprint, voice, gait, patterns of key presses;
- tracking (via fitness trackers, credit card use, data from transport networks);
- speech recognition and language translation;
- medical diagnosis;
- detection of disease outbreaks via analysis of google search data;
- the Internet of Things – smart refrigerators, TVs, cars, and domestic robots;
- 'deep fake' videos;
- predicting crime and recommending custodial sentences;
- satnav; autonomous vehicles and weapons systems;
- mapping dwellings from aerial images, in remote settings;
- emotion detectors for classrooms and cars.

A striking feature of data science has been the variety of problems addressed, the kinds of data analysed and used, the range of novel models developed, and its direct effects (intended and unintended) on people's lives.

Most of these developments can be described as 'engineering' – a useful product emerges from an analysis of an interesting challenge. The relationship between statistics and data science is analogous to the relationship between mathematics and engineering. Engineers don't do 'applied mathematics' they do 'engineering', and used mathematics where appropriate. Similarly, data scientists don't do 'applied statistics' they build things, and use statistics when they (think they) need to.

It is worth reflecting on the extent to which analytic models, p-values and effect sizes have contributed to the developments in computer science that have radically reshaped the modern world. For the practical examples listed above, the designers' ambitions are for 100% success, not for theoretical nicety, nor for performance that is 'significantly better than chance'.

3. Designing the epistemological engine

We are living in interesting times; new phenomena are emerging (associated with billions of people having internet access, much greater wealth and better health, worldwide). New sorts of data are available; there are new sorts of analytic tools; there are new creators of knowledge (notably technology companies) and new distributors, consumers and users of knowledge. The problems that beset the start of the twentieth century have not gone away; modern societies now also face existential threats such as global warming and nuclear war. There is a need for knowledge-generators to engage with problems that can be characterised as 'messy', 'complex', or 'wicked'. These problems are characterised as being ill-defined in terms of specifying relevant variables or measuring progress; they often involve interacting systems at different levels. For example, climate change is influenced by the actions of individuals (e.g. car choice and use), local structures (e.g. support for recycling), national structures (e.g. policies on house insulation and domestic solar power), and international initiatives (e.g. consensus on restricting carbon emissions). There is no 'right' level to work at; there are multiple ways to measure system states and the results of different initiatives.

Addressing 'wicked problems' is likely to involve working with multiple sources of messy data, and using a variety of analytic tools (see Ridgway *et al* 2018). Inter-disciplinary action is almost certain to be essential to success. However, scientists working with even relatively simple problems can make a mess of things. There are serious challenges to current methods of knowledge acquisition, illustrated by the very poor quality of much of the research funded at great expense in universities (see Ioannidis (2005); Open Science Collaboration (2015)). These cannot be explained away as the result of poor practice by a few individuals; they reflect systemic failure by some academic communities. There is an urgent need to analyse and improve the whole

system associated with the creation and use of knowledge – in short, designing an Epistemological Engine (EE) has become a priority. The prime candidates for creating and building the EE are statisticians and data scientists.

Early encounters between statisticians and data scientists were often acrimonious; 'statistics' would be a casualty in 'the death of theory', and data scientists' ignorance of core statistical concepts such as sample bias and overfitting would prove fatal to their entire enterprise. The EE should be founded on techniques and skills used in both data science and statistics. Data scientists create open data repositories (e.g. https://registry.opendata.aws/), and have adopted a culture of sharing code – especially Workflows (e.g. https://github.com/) to facilitate a comparison of different analytical techniques and modelling assumptions. They use Common Task Frameworks wherein success is judged on terms of actual performance in analysis, not theoretical niceties. Statisticians bring sophistication about data acquisition (including synthesising and triangulating data sources), preparation, and exploration. They can contribute to analyses, data representation and communication, and can comment on issues such as the likely generalisability of findings. They bring considerable sophistication about modelling. Identifying the style of modelling being used by different researchers (explicitly or implicitly) should be automated in the EE.. Ridgway (1998) classifies styles of modelling, and describes analytic models (such as those found in school physics), systems models (such as those found in school biology) and macrosystemic models – these are systems models where the system itself undergoes change. Macrosystemic models can be divided into two groups – models where the changes in the system are relatively predictable (e.g. ecological restoration; the life cycle of the butterfly) or unpredictable (Brexit; climate change and global political stability in the Trump era).

The EE should comprise a large tool collection. Sample tools include:

- Critical evaluation of specific studies, using criteria for evaluation such as those identified by Ioannidis (2005) and the Open Science Collaboration (2015), e.g. identifying weak effects using small samples, and testing multiple hypotheses until a 'significant' result is found;
- Identification of academic areas where there is insufficient sharing of data, code and workflows;
- Identification of academic areas that are paradigm-bound (i.e. characterised by analyses of rather few classes of data, and by the use of a small set of analytic tools);
- Tools for automated testing of code and workflows;
- Identification of results that are important for some theoretical claims, where the evidence base is weak (e.g. where there has been little replication across relevant populations);

- Automation of literature searches, and the conduct of meta-analyses;
- Creating semantic nets of academic papers in terms of both content and authorship in order to document the flow of discovery processes;
- Methodology classification systems, that support automated classification;
- Analogy generators, to suggest developments in fields other that the one in which a method or tool was developed;
- Methods for analysing large corpora of research in different fields to examine the epistemological assumptions made (including pragmatism).

Knowledge gaps

There are some glaring gaps in our knowledge that need to be remedied, we need: more formal theories of data analysis; more work on the cognitive psychology of data visualisation and interpretation; and more and better modelling of emotion, social behaviour, and cognition; better understanding of the processes of knowledge generation, distribution and use, and more tools for working with very large data sets.

4. Competences for students of data wrangling

So what do students need to know in order to work in this brave new world? Here, we offer some more lessons for young minds.

- Be aware of the politics of technology: technologies are never neutral (e.g. cars cannot be driven by the very young or old, or the poor)
- Attend to unintended consequences (e.g. cyberbullying via social media) via 'what if' games
- Engage with moral issues (e.g. the dangers of the Panopticon)
- Be aware of epistemological issues: the nature of knowledge as conceived in different academic disciplines - how it is created, shared, learned, and used (and by whom, and for what purposes)
- Understand modelling and the limits of modelling, and the principles of model validation;
- Explore the reasons for the existence of data sets – adopt a hermeneutical approach
- Create a conceptual web to link between seemingly different methods
- Understand the principles underpinning different techniques (e.g. neural nets)
- Learn to represent the same problem in a variety of ways
- Become fluent in the use of major data repositories
- Share your code and workflows
- Invent and modify data visualisations (including dashboards)

5. Concluding Remarks

The early history of data science holds important lessons; technology, mathematics and society are in a continuous state of rapid change. Students should be made aware that current knowledge will be superseded, and that there are social forces that can limit their creativity.

Technologies have created existential threats to humanity such as global warming and nuclear war. There is a pressing urgency to address such problems. Both statistics and data science have their roots in solving challenging problems, but have traditionally adopted somewhat different approaches. Statistics is characterised by sophisticated modelling using a small set of well-defined variables; data science is often a-theoretical. Data science adopts practices that should be applied across a wide range of disciplines, such as sharing data, code and workflows. Statistics is strong on discovery methods.

There is an urgent need to create an Epistemological Engine – a set of semi-automated tools to understand and support effective science. Statisticians and data scientists are the people best placed to create and maintain this Engine. We offer some ideas on the tool set that will comprise the EE, and some suggestions about the competences needed by future data wranglers.

And a final piece of advice for young minds: *make a wall poster of these words from Ada Augusta King, Countess of Lovelace…*

> "A new, a vast, and a powerful language is developed for the future of analysis… the theoretical and the practical in the mathematical world, are brought into more intimate and effective connexion with each other." (Lovelace, 1843, p3)

References

1. Babbage (1864). *Passages from the Life of a Philosopher.* (https://en.wikisource.org/wiki/Passages_from_the_Life_of_a_Philosopher/Chapter_VIII
2. Boole, G. (1854). *The Laws of Thought*. *An Investigation of The Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*, Originally published by Macmillan, London. Reprint by Dover, 1958. Cited at https://plato.stanford.edu/entries/boole/#LawsThou1854
3. Box, G., and Draper, N. (1987). *Empirical Model-building and Response Surfaces.* New York: Wiley. Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.*, **16**(3), 199–231.

4. Cobb, G. W. (2015). Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up. *The American Statistician, 69*(4), 266–282.
5. Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine* **2**(8): e124. doi:10.1371/journal.pmed.0020124
6. Lovelace, A. (1843). Notes on a translation of Sketch of the Analytical Engine invented by Charles Babbage by L.F. Menabrea (1842). https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by_Charles_Babbage,_Esq./Notes_by_the_Translator. Downloaded 5 April 2019.
7. Menabrea, L. (1842). On the mathematical principles of the Analytical Engine. *Bibliothèque Universelle de Génève*, No. 82. October 1842. Translated by A. Lovelace.https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by Charles_Babbage,_Esq. Downloaded 5 April 2019.
8. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 348(6251)*. DOI: 10.1126/science.aac4716.
9. Perez, C. (2019). *Invisible Women: exposing data bias in a world designed for men.* London:Penguin. Pullinger, J. (2013). Statistics making an impact. *J. R. Statistic. Soc. A, 176*(4), 819 – 836.
10. Ridgway, J. (1998). *The Modelling of Systems and Macro-Systemic Change - Lessons for Evaluation from Epidemiology and Ecology.* National Institute for Science Education Monograph 8. University of Wisconsin-Madison.
11. Retrieved from http://archive.wceruw.org/nise/Publications/Research_Monographs/Vol8.pdf. Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3). Retrieved from onlinelibrary.wiley.com/doi/10.1111/insr.12110/full.
12. Ridgway, J., Ridgway, R. and Nicholson, J. (2018) Data science for all: A stroll in the foothills. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward*. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute. http://icots.info/10/proceedings/pdfs/ICOTS10_3A1.pdf?1531364253
13. RSS (2019). Past Presidents https://www.rss.org.uk/RSS/About/About_the_RSS/President_and_Vice_Presidents/Past_presidents/ RSS/About_the_RSS/About_sub/President_and_vice_presidents/Past_presidents.aspx?hkey=eec573c_6-b10e-4027-8846-9802a07a1d28 Downloaded 8 April 2019.

Teaching data science - a user perspective or (preparation for a career as an official statistician)

Steve MacFeely[1,2]
[1]United Nations Conference on Trade and Development, Geneva, Switzerland
[2]Centre for Policy Studies, University College Cork, Ireland

Abstract
Within the broader panoply of statistics and data science, official statistics is a discipline of its own. While it requires the same core skills as most other statistical professions, it also requires other dimensions or elements unique to official statistics. Most formal statistical and data scientist training has to date, not addressed these other elements, and therefore have not equipped students very well for the complex and challenging world of official statistics. Introducing students to this fascinating world in university, would allow them to better understand how statistics can play an integral role in public policy and the important role they could play in shaping the way we think about our world. This paper discusses, from the perspective of an employer of data scientist and statistical graduates, some of the topics that universities might consider including in their curricula. Furthermore, the argument that a greater focus on nurturing competencies rather than on developing specific skills to prepare students for a rapidly changing world is explored.

Keywords
official statistics; statistical literacy; skills; competencies

1. Introduction

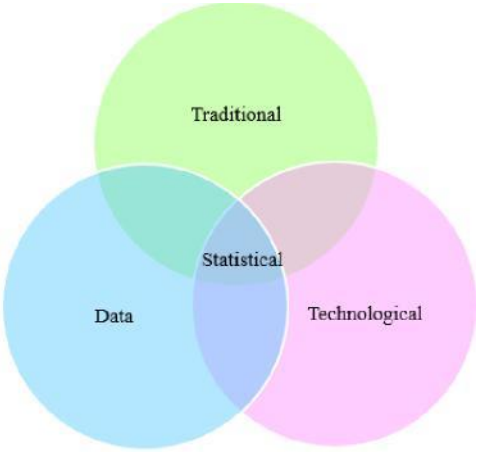
> Q: What's the difference between a data scientist and a statistician?
> A: 40K per annum

For the purposes of this paper, I am treating a data scientist and a statistician as being synonymous (I won't be addressing salary differentials). Perhaps if we dive deep enough there are some subtle differences, but for all practical purposes I believe they are essentially the same profession requiring the same skills and competencies. The real difference perhaps is that one – the data scientist has jumped on the 'data revolution' wave and has rebaptised him or herself for marketing reasons. For the remainder of the paper, I will refer to statisticians, but the reader can substitute data scientist if they prefer.

It might seem obvious, but a statistician hoping to work as an official statistician in a National Statistical Office (NSO) or National Statistical System (NSS) should be statistically literate or should at least be capable of becoming

statistically literate in a reasonably short space of time. Statistical literacy can mean different things to different people, so for the purposes of this paper, a very broad definition or concept of statistical literacy is used (see Figure 1). There are many definitions of statistical literacy [e.g. 1]. For the purposes of this paper, I explain statistical literacy as the nexus or intersection point between traditional literacy (which includes reading, writing and mathematics, but also a wider appreciation of history, geography, legal issues, ethics and culture); technological literacy (which includes knowledge of databases, software and increasingly, social media); and data literacy (which requires an understanding of primary and secondary sources, including big data and GIS, open data, and an understanding of the Generic Statistical Business Process Model). Statistical literacy straddles all of these, requiring elements from each of these dimensions, but in addition, requiring an appreciation of statistical techniques and modelling. Put together, a statistician should understand the broader context of what is being measured – including the history and the appropriateness of the measurement tool[1].

Figure 1. The Statistical Literacy Nexus



2. Required Skills

Thinking about NSOs and NSSs of the future, it is very hard to anticipate what specific skills will be required. That said, at this juncture, it is hard to imagine that the 'hard' skills expected of a statistician today not remaining relevant. Arguably, these core statistical skills will become more important, rather than less, as graduates become more accustomed to using blackbox software packages but perhaps less accustomed to thinking about the underlying concepts and methodologies. While hard skills will most likely

[1] This view of literacy overlaps considerably I think with the views expressed in the ProCivicStat (2018) - http://community.dur.ac.uk/procivic.stat/

remain an essential ingredient for any statistician in the future, perhaps the necessary skills mix will change. For the moment (and I think for the foreseeable future), three essential skills will be required: numerical skills; statistical skills; and increasingly, technological skills (see Figure 2). Irrespective of whether we are discussing a data scientist, an official statistician or a professional statistician in another field, the requirement for these basic skills is universal. Mathematical and numerical skills is I think self-explanatory, but crucially a statistician should be able to spot patterns, understand differences between stocks and flows and be comfortable reading and writing in scientific notation. Statistical skills means being able to work with real, often messy or incomplete data. Understanding bias; both the likely sources and what remedial actions can be taken. Statisticians should understand the subtle but important differences between accuracy and precision. They should also develop a good understanding of concepts like uncertainty and risk. A competent statistician should be able to select and use appropriate statistical techniques and models. Future technological skills are the area hardest to predict. Technology is changing rapidly, with consequences not only for the applications we will use, but also the types of data we may have access to. Here it is very hard for a university to prepare courses for the future and for statistical offices to say with any certainty what will be required. If current trends have anything useful to say, then it suggests a greater use to 'freeware' and combining packages. It also suggests a commitment to lifelong learning will be essential.

Statisticians must understand the underlying logic of theory, so that having acquired skills, they can apply them and put theory into practice in a variety of real-life situations (all invariably more complex and messy than the scenarios presented in text books). Other skills, perhaps neglected in the past, but now universally recognized as important, is the ability to communicate well and to present statistics in context. In a world awash with data and cluttered, incoherent babble, the ability to translate data into coherent statistics and understandable and digestible messages is absolutely essential. Data visualization is an important element subset of this skill, but perhaps one where too much emphasis is being placed at the moment. A statistician can only design an effective visualization if they are clear themselves what the key messages are.

Figure 2 – Skills and Competencies of a Data Scientist / Statistician

3. Required Competencies

Any discussion of future requirements should, I think, be broadened beyond skills to include competencies. Statisticians, like any other professional will need to continually update their skills over the lifetime of their career[2]. What is less likely to change over time are the basic characteristics or competencies necessary to be a good statistician. Specifically, I would argue a statistician must be creative, curious, critical, skeptical and resourceful (see Figure 2). These I think are self-explanatory and don't require any elaboration. Perhaps, less obvious, a statistician should also be aware of the cultural and civic or political environment in which they operate. It is very important that a statistician not only understands the context in which previous indicators and statistics were compiled, but also that they properly understand the environment in which they operate. For example, when contemplating the use of big data, MacFeely [1] notes that NSOs may be forced to confront issues before the law is clear or cultural norms have been established. Given the importance of public trust for an NSO it is essential that statisticians are sensitive to these issues and understand what is acceptable by the public they serve. The challenge for universities is how to nurture and develop competencies.

[2] Outside the scope of this paper, but perhaps there is a market for universities beyond preparation of early career statisticians to cater better for mid and senior career statisticians too?

4. What Universities could teach

A course that prepares statisticians for a career in official statistics would ideally cover 4 broad domains[3]: statistical systems; statistical techniques; policy issues; and for want of a better heading, the political economy of data. In the limited space available, a few elements of these domains are highlighted. 'Official statistical literacy', which comprises of this wider mix of issues would enrich the content of university courses, stimulate discussion and give statistics students a better flavor of what their future career might involve.

4.1 Statistical Systems: This would include an introduction to the official statistics eco-system comprising of national, regional and global statistical systems. Understanding the mandate and function of the UN Statistical Commission [4] including: how internationally recognized classifications, such as the ISIC[5], and framework methodologies, such as the SNA[6], are developed; and how other regional and national statistical systems, including international organisations coordinate their work. Statisticians should also be introduced to: the UN Fundamental Principles of Official Statistics [2], the Principles Governing International Statistical Activities [3] and any regionally or nationally appropriate adaptations; professional principles, such as, the ISI Declaration of Professional Ethics [4]; and internationally recognized statistical quality assurance frameworks and codes of practice, such as, the UN System statistical quality assurance framework [5], or the UN National QAF [6] and any regionally or nationally appropriate adaptations. An important element of any statistician's work (whether official or otherwise) will be dealing with legislation. Official statisticians will need to be conversant with the national statistical law (and those in Europe will also need to know their way around European statistical legislation). But increasingly, concerns regarding privacy and confidentiality[7] mean that all statisticians and data scientists must understand the legal environment in which they work, in particular, data protection and freedom of information legislation and how they interrelate. As we expect more and more secondary digital data to be generated, the legal issues of accessing and using these data are critical to ensuring a statistician does no harm[8].

4.2 Statistical (and technological) Techniques: Traditional topics like distribution theory, probability, sampling/weighting, multivariate analyses and

[3] Although the focus of this paper is preparation for a career as an official statistician, most of what is discussed in this section, would arguably be very valuable for any statistician.

[4] https://unstats.un.org/unsd/statcom/

[5] ISIC - International Standard Industrial Classification of All Economic Activities

[6] SNA - System of National Accounts

[7] UN Fundamental Principle of Official Statistics No. 6

[8] ISI Professional Value 1.2.

regression must be included, but so too should some basic, but often ignored, statistical techniques. For example, data cleaning (including treatment of outliers, imputation and interpolation) is an essential skill. Unfortunately, NSOs rarely get to work with the clean datasets favoured by university courses. Real life data are typically very messy. Another technique, too often ignored, is seasonal adjustment. Essential for analyzing and presenting sub-annual time series, yet frequently young career statisticians don't know how to seasonally adjust time series or how to test time series models to ensure they are appropriate. More could be done to examine real life issues, such as, how to seasonally adjust series in the aftermath of shocks, such as the 2008 financial crisis. Another area deserving of more attention is index theory and practice – the various index formula and where it is appropriate to use them. Fixed weight versus chain linking. These are essential for statisticians working in price, business and macroeconomic statistics. Increasingly NSOs and IOs are compiling leading, composite and sentiment (LCS)[9] indices - the merits and technical challenges of such indices could usefully be discussed. In a 'big data' world, with massive computing power facilitating the linking of records, safeguarding confidentiality and public key cryptography are becoming greater challenges. The importance of these subjects deserves a prominent place in any statistics curriculum. Students should also be introduced to programming logic, not just how to use the most fashionable software packages.

4.3 Policy Issues: Beyond technical statistical techniques and tools, universities could play an important role in highlighting some of policy issues that pose significant challenges for public policy (and by extension NSOs). For the purposes of this paper 4 issues are highlighted, but this is by no means exhaustive. The first is globalization. With the emergence of the internet, the fall of the Berlin wall and China joining the WTO, properly measuring the impacts of globalization is now a major issue for statisticians. Economic and social globalisation is impacting on employment, crisis contagion, trade policy, protectionism and migration [7]. It has also challenged the relevance of traditional trade, price and macro-economic statistics. The next issue is measuring wellbeing (aka progress).

Since the 1970's there have been several attempts to supplement or supplant GDP with other indicators of progress[10]. Since the 2008 financial

[9] Some examples: UNDP Human Development Index; OECD Better Life Index; WEF Global Competitiveness Index.

[10] The Measure of Economic Welfare (MEW); the Genuine Progress Index (GPI); the Human Development Index (HDI); and index of Gross National Happiness (GNH) to name a few.

crisis those efforts have intensified with a raft of new indices[11] attempting to measure and blend human, natural and production capital or the social, economic and environmental pillars of development. Linked, is the emergence of climate crisis and environmental degradation. For example, how can we put a value on the environment, on biodiversity, on ecosystems, or spirituality or aesthetic? Challenges regarding potential double counting, designing a single valuation methodology that works equally well across all domains, and problems with data availability all pose difficulties. For example, most economic valuations are based on marginal changes to ecosystems on the assumption that they are stable, but in reality, little is known about the stability of ecosystems and their response to change - a critical threshold could trigger structural changes, at which point the marginality assumption and the valuation may no longer hold. Another serious policy concern is 'end of work' or a 'jobless future' – the disruptive blend of technology, automation and task based labour that has been dubbed the 4th industrial revolution. The shift towards non-standard work patterns is blurring the distinction between formal and informal employment. The variety of new non-standard or on-demand employment contracts mixed with robotics and other technological advances are posing challenges for those trying to classify and measure this new gig economy.

4.4 Political Economy: Statisticians should also be aware of the wider context on how data are being used and the important debates underway in which data or statistics are in one way or another at the heart. At the core of official statistics is the ambition to provide impartial information[12], yet no statistic is truly impartial – every statistic is the product of many decisions and assumptions, conscious or subconscious. It is important that students understand this and that the choices they make have an impact. The selection of variables in a composite index; whether to weight or not; or the treatment of outliers all affect, not only the basic result but perhaps also the alignment to political or economic ideologies. This realization becomes especially important when thinking about 'evidence based decision making' versus 'evidence informed decision making' or 'Governance by numbers' [8]. This is a growing concern for many as algorithms are playing an increasingly greater role in our lives, from deciding whether we get a loan to whether we are short listed for an interview.

Many other interesting debates are underway, not least, the 'End of Theory' argument posited by Anderson [9] that causation no longer matters, and that with big data, only correlation matters[13]. Should 'open

[11] European Commission Beyond GDP; OECD Better Life Index; and UNEP Inclusive Wealth Index.

[12] UN Fundamental Principle of Official Statistics No. 1

[13] For the record, I think this argument is nonsense, but it is an interesting debate.

data' exclusively target public sector data, or should it also push for more openness with private sector data. This discussion is connected to the nascent but fascinating debate regarding who owns the data held by social media and search platforms that are essentially the product of our labour. Is 'Fake News' undermining the credibility of science and official statistics and how can NSOs tackle it without themselves becoming politicised? Has official statistics failed in its role – why for example, in the face of overwhelming statistical evidence is the climate crisis not being taken as seriously as it should? There are interesting ideas, such as, data infrastructure to be discussed where state information is organized into state registers supported by unique identifiers [10]. This is a very logical, effective and efficient approach but requires some sacrifice of privacy to the state. What are the cultural barriers to this? There are so many fascinating debates that students could and should be introduced to, so that they understand, the fascinating range of opportunities available to them. For example, there is much talk today of Data Revolution, but what does that mean? Has there really been a revolution or is the data deluge just another step in the evolution of data? Discuss...

5. Conclusion

I have argued that statisticians require a broad range of skills. One can anticipate technological skills will playing a greater role in the toolbox of a future statistician, but they will still be required to have all the core statistical skills. But in a rapidly changing environment, the requirement for particular technological skills may change or evolve. However, the core competencies will not. I have argued that a statistician, whether today or in fifty-year time, will need to be curious about what's going on. They must naturally critically analyse trends and be sceptical of results. The world is a complex, messy and resource constrained place, and so they must be creative and resourceful if they are to get the job done.

The role of universities should not only be to impart information and skills but ignite curiosity and stimulate discussion. Too often statistics is presented or taught as a black and white set of rules or solutions. But in reality, the application of statistics to the complex world we inhabit, is at best grey. Official statistics, in particular, has not been historically well covered in academia – perhaps as many academics themselves don't fully appreciate or understand the role. I have argued, that in addition to teaching the traditional elements of statistics, more practical, applied and contextual statistics could be taught.

The bigger challenge for universities is how to cultivate and nurture competencies. This I would suggest might be tackled by introducing students to what I have termed policy issues and the political economy of data. These issues are fascinating and could be used to encourage debate, stimulating

students to think about the complex role that data and statistics play in todays data driven world. The role of *Diploma in Official Statistics*[14] in Ireland or the *European Master in Official Statistics*[15] are interesting case studies as they were designed with input from official statisticians and take a broader perspective than traditional statistics courses. Nevertheless, more could be done to widen the scope of statistics courses, making them more interesting, linking them to the great debates of the day, and showcasing what a fascinating and rewarding career may be awaiting graduates.

References

1. Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, and Responsibilities. International Statistical Review, 70, 1, 1-25.
2. MacFeely, S. (2018). 'Big Data and Official Statistics' in Kruger, S. and Kruger, M. (Eds.) in *Big Data Governance and Perspectives in Knowledge Management*, IGI Global, Hershey, PA, pp.25 – 54.
3. United Nations (2014). Fundamental Principles of Official Statistics. Resolution 68/261 adopted by the General Assembly on January 29, 2014. A/RES/68/261.
4. Committee for the Coordination of Statistical Activities (2014). Principles Governing International Statistical Activities.
5. ISI (2010). Declaration of Professional Ethics.
6. Committee of the Chief Statisticians of the United Nations System (2018). United Nations Statistics Quality Assurance Framework.
7. UNSD (2012) Generic National Quality Assurance Framework (NQAF).
8. MacFeely, S. (2016). The Continuing Evolution of Official Statistics: Some Challenges and Opportunities, Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 789–810
9. Fukuda-Parr, S. and O'Neill, D. (2019). Knowledge and Politics in Setting and Measuring the SDGs: Introduction to Special Issue, *Global Policy*, Vol. 10 (1), pp.5 – 15.
10. Anderson (2008). The End of Theory: The data deluge makes the scientific method obsolete. *Wired* June 27, 2008.
11. MacFeely, S. and J. Dunne (2014). Joining up public service information: The rationale for a national data infrastructure. *Administration,* Vol.61, No.4, pp. 93–107.

[14] https://www.ipa.ie/public-management/professional-diploma-in-official-statistics-for-policy-evaluation.1763.html

[15] https://ec.europa.eu/eurostat/web/european-statistical-system/emos

Data Science programmes: Is there an ideal design?

Jeremiah D. Deng, Matthew Parry
University of Otago, New Zealand.

Abstract

"Data Science" has become a buzzword in recent years, and many universities have started offering undergraduate degrees in Data Science. Yet the shape of Data Science as an interdisciplinary field remains elusive for a clear definition, and the curriculum design varies from one programme to the other. In this paper we sample a few Data Science undergraduate programes offered in a number of institutes in the US, China, Australia, and New Zealand. The diversity of these programmes is revealed by using indices quantified on four dimensions: mathematical, statistical, computing, and programming. Furthermore, we use Machine Learning as the core subject within a Data Science programme in an effort to map out some key, prerequisite subjects as required for teaching mainstream algorithms effectively. We also argue that there are also other aspects of Data Science that can be easily neglected by most offerings, such as distributed database systems, and privacy preserving data mining, let alone domain-rooted subjects such as biomedicine and computational finance. With these considerations in mind, we call for a flexible curriculm design that incorporates a thin core and allows students to opt for endorsements in different specialties, e.g. statistics, information technology, and big data applications.

Keywords

Data science; curriculum design; diversity

1. Introduction

Data Science as an interdisciplinary subject has become an increasingly important area that attracts intensive efforts worldwide in teaching, research and development. As a relevant subject statistics as regained popularity and the number of undergraduate statistics degrees have trippled over the last decade largely due to the emergence of big data[1].

Having become a buzzword, "data science" still lacks a clear, widely adopted definition. The very nature of data science, its content, and perceptions regarding to its potentials and weakness etc., remain hot topics

[1] The Conversation.com, "Statistics and data science degrees: Overhyped or the real deal?", URL https://theconversation.com/statistics-and-data-science-degrees-overhyped-or-the-real-deal-102958, October 29, 2018. Retrieved April 30, 2019

debated in daily conversations, and in scientific discourses. Dhar (2013) pointed out that, as a multi-disciplinary subject, Data Science (DS) is distinct from statistics, referring to the growing needs of processing data that are increasingly heterogeneous and unstructured. There is an emerging consensus to see DS as an interdiscplinary field that incorporates mathematics/statistics, computer science and information science, and domain knowledge (WSU, 2016). Consequently, we consider it an interesting question to ask: is there an ideal design of DS undergraduate curricula?

2. Methodology

Profiling DS Programmes

In this paper, we will examine a few typical DS curricula as adopted by some institutions in the US, China, and New Zealand. The diversity of these DS programmes is demonstrated in the different types of insitutes (liberal-arts colleges, business schools, and universities) and different degress (BA, BSc, and BEng). To profile the curriculum designs from these programmes, we use a hybrid quantification approach to score their requirements on four dimensions, and we use visualization to indicate the differences: mathematics, statistics, programming, and computing. We concentrate on the prerequisite and core levels and leave the electives out to gain some understanding of the core structure of these programmes.

Using University of Columbia as an example for illustration, we take a look of its undergraduate DS programme design as shown in Table 1[2]:

Table 1. Programme structure of the DS major at Columbia

Prerequisites: 15 points • Calculus I – III • Linear Algebra (Math or Applied Math) • STAT 1201 (Calculus-Based Introduction to Statistics)
Core: 8 courses (STAT and COMS) STAT (12 points): 1) STAT 4203 (Probability Theory) 2) STAT 4204 (Statistical Inference) 3) STAT 4205 (Linear Regression Models) 4) STAT 4241 (Statistical Machine Learning) or COMS 4771 (Machine Learning) COM (12 points) 1) Introduction to Computer Science: COMS 1004, COMS 1005, ENGI 1006, or COMS 1007

[2] Columbia University, URL https://mice.cs.columbia.edu/c/d.php?d=245, August 8, 2018. Retrieved April 29, 2019

2) Data Structures: COMS 3134, COMS 3136, or COMS 3137 3) Discrete Math: COMS 3203 4) Analysis of Algorithms: CSOR
Electives: 5 Courses STAT: 2 from the following 1) STAT 3106 (Applied Data Mining) 2) STAT 4206 (Statistical Computing and Introduction to Data Science) 3) STAT 4243 (Applied Data Science) 4) STAT 4224 (Bayesian Statistics) 5) STAT 4242 (Advanced Machine Learning) COMS: 3 from the following 1) COMS 3261 (Computer Science Theory) 2) COMS 4111 (Introduction to Databases) 3) COMS 4130 (Principles and Practice of Parallel Programming) 4) COMS 4236 (Introduction to Computational Complexity) 5) COMS 4252 (Introduction to Computational Learning Theory) 6) Any COMS W47XX course except COMS 4771 (These are our AI/ML oriented courses.)

By matching and evaluating the course outlines we gathered from the Internet, we first score the subject intensities as represented in various DS progarmmes with a range from 1 (only rudimentary) to 3 (very strong presence and coverage). In Columbia's case, the scores are: Mathematics (3: calculus, algebra, discrete mathematics), Statistics (3: probability, inference, regression, machine learning), Programming (2: data structure, algorithms), Computing (1: introduction to CS). Clearly, the score on Computing could be improved if other CS subjects are better covered such as database and cloud computing. This shows that there might be some pedagogical or logistics factors that have contributed to this particular curriculum design.

The DS profiles each with four scores are ready to be contrasted.

The following DS profiles are included for comparison in this paper:

- Columbia University, BA
- Michigan University, BEng
- Rochester University, BA/BSc
- Kansas State University, BA/BSc
- Maryville Colege, BA with BusAdmin minor
- "ChineseU", a university in mainland China[3]
- Auckland University, BSc
- Canterbury University, BEng

[3] As the interim course information was obtained through private communications only, the university hence is made anonymous.

Secondly, we employ Kiviat diagrams to visually compare the DS profiles.

Mapping of Requisites

Besides measuring the overall DS programme structure, we will also look at the instrinsic requirements in terms of teaching DS, using the Machine Learning course as an example. Machine Learning is chosen because it is a subject that can be either taught by a Statistics or a Computer Science department, and is considered a core subject within a DS programme. By looking at some specific, core concepts and algorithms taught within the course, we plan to work out some desirable requisites that may be covered by some prerequisites or early core courses. This will show how a particular DS design may accommodate the needs for a successful delivery of Machine Learning.

3. DS Profiling: Findings

Using DS programme links listed at various online sources[4], we score a number of undergraduate DS programmes on four dimensions, using the afore-mentioned method. The total weight as the sum of the four scores, gives an indication about the overall technical intensity of these programmes. The results are shown in Table 2.

We can visualize these DS profiles using Kiviat diagrams (Kolence, 1973) as shown in Figure 1. It can be seen from Figure 1 that the DS profiles vary from each other in terms of sizes and orientations. Maryville and ChineseU both have envelops of larger sizes, while Auckland and Kansas State etc. give smaller envelops. Their focuses seem to differ: Maryville leans to statistics rather than computing, while Canterbury is on the contrary.

Therefore some interesting questions arise. Why is there some significant diversity as observed in these DS programmes? We do not have further information in this regard, but we suspect a number of factors may contribute to the diversity: the subjective understanding of what DS stands for, the availability and preference of the teaching staff with expertise, the industrial demands from multiple sectors, etc. Unlike traditional disciplines such as computer science and electronic engineering, there are no professional agencies such as IEEE and ACM yet to provide curriculum recommendations or accreditations. On the other hand, the multi-disciplinary nature of Data Science will be persistent, and institutions may see the diversity an opportunity as to promote the uniqueness of their own DS programmes as features desirable for student recruitment or appealing to particular niches of job markets. One may see it important to equip DS graduates with skills in developping new statistical algorithms for new data types; some may find it

[4] E.g., https://www.discoverdatascience.org/programs/bachelors-in-data-science/

demands may make DS curriculum design seem quite subjective and arbitrary.

crucial to train them on how to deal with large-scale problem solving utilizing distributed storage and cloud computing; and others may find both of them crucial, or irrelevant! The diversity in the understanding of DS and in industrial

Table 2. Profiling DS programmes' technical strength on four dimensions: Mathematics, Statistics, Programming, and Computing. The total weight is shown on the last column.

Institutions	Mathematics	Statistics	Programming	Computing	Total Weight
Columbia	3	3	2	1	9
Michigan	3	2	2	1	7
Rochester	3	2	2	0	7
Kansas	2	2	3	0	7
Maryville	3	3	3	1	10
ChineseU	3	2	3	3	11
Auckland	2	3	2	1	8
Canterbury	2	2	2	3	9

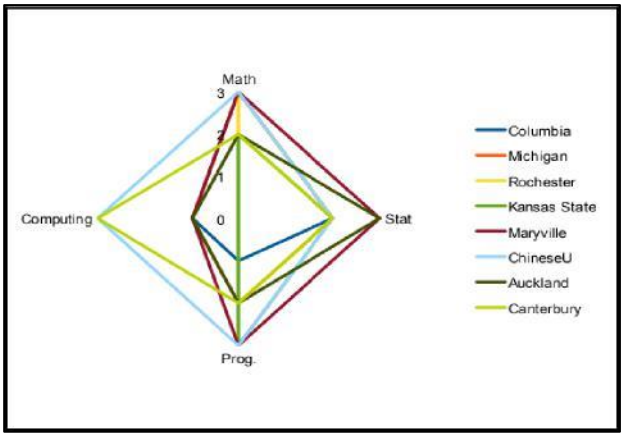


Figure 1. Kiviat diagrams of various DS programmes.

4. Mapping inside-out

To justify DS as an interdisciplinary but standalone discipline, there has to be some common understadning on the core requisites and essential components within a DS programme. In this section we will try to map out the

requisites on mathematics and statistics by looking at some DS course content at a higher level within the undergraduate programme

As explained in Section 3, Machine Learning is a core subject within a DS programme and serves a good example course for us to examine what particular prerequisites and core components are necessary for its successful delivery.

A Machine Learning course can be taught anywhere in the latter half of a DS programe. In the New Zealand or Australia context, it can be second-year or third-year "paper". Regardless, we look at what mathematical and statistical prerequisites may be necessary to enable effective teaching or learning of key Machine Learning concepts and algorithms.

Using the famous "top-10 data mining algorithms" (Wu et al., 2008) as a starting point, we examine the relevant preliminary subject studies as required by each of the algorithms. In 2006, ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners were asked to nominate key algorithms across all fields of data mining and machine learning, and the nominations were then voted by hundreds of the ICDM'06/SDM'06/KDD'06 Technical Programme Committee members, resulted in the top-10 algorithms listed in Table 3. Here for each algorithm, the relevant background mathematics or statistics knowledge as required is estimated, and matched to four generic courses with code and content outlined as follows:

- MATH100: Entry-level algebra and calculus
- MATH200: Linear algebra, discrete mathmatics, optimization
- STAT100: Basic statistics such as probabilities and tests
- STAT200: Statistical inference

The ticks in Table 3 are made in a rough estimation of a normal delivery of the algorithm. For instance, for the k-nearest neighbour (k-NN) algorithm its algorithmic operation is introduced and relvance to density estimation is hinted, but the connection to EM and Bayesian inference is not necessarily discussed (which then would require STAT200). Also, if a "200" option is ticked for an algorithm the "100" option will be omitted.

As seen from Table 3, it seems that the top-10 algoirthms can be delivered effectively without too much mathematical or statistical requisites.

On the other hand, DS is a discipline that undergoes rapid advances. To reflect the landscape of R&D a decade latter than the top-10, we may have a new list of key algorithms as given by Table 4. Clearly the algorithms have become more advanced, bearning complexities that require deeper mathematical or statistical understanding. Hence we come to the conclusion that both MATH200 and STAT200 are indispensable cores of a proper DS programme (that teaches Machine Learning effectively).

Arguably, we can adopt a similar approach to map out the core requisites of DS programmes by looking at other computing and statistics courses. This

will be helpful for us to derive a thin-core undergraduate DS curriculum, which may be flexibly augmented by electives in senior years, including some strongly domain-related courses such as bioinformatics and computational finance.

5. Discussion and Conclusion

In this paper, we have surveyed a number of representative data science undergraduate curricula and revealed some intersting diversity between these programmes. Perhaps the inherent interdisciplinary nature of data science itself justifies and demands its diversity, and hence it will be futile to keep a consistent curriculum design. Rather, institutes may find it more rewarding to install a thin-core but multi-facet programme that accommodates students' and employers' diverse interests.

This flexible design implies running classes with strong diversity and poses new challenges to mathematical teaching. The strong correlation between students' performances in mathematics and engineering subjects has been confirmed (Bishof et al., 2015). On the other hand, Ooi (2007) criticised the usual, administratvely efficient mode of mathemtics teaching delivered as separate subjects, resulting in low relevance perception among Engineering students. Are these findings relevant to DS as well? This is a topic to be investigated in our future work.

Table 3. Top-10 data mining algorithms and their corresponding prerequisites as required.

Algorithm	MATH100	MATH200	STAT100	STAT200
C4.5			x	
k-means	x			
SVM		x		
Apriori	x			
EM	x			x
PageRank		x		
AdaBoost	x		x	
kNN			x	
NB				x
CART			x	
# Ticks	**4**	**2**	**4**	**2**

Table 4. Requisite matching for newer algorithms

Algorithm	MATH100	MATH200	STAT100	STAT200
PCA/Kernel PCA		x	x	
Laplace eigenmap		x		
t-SNE		x		x
AdaBoost	x			x
MLP/CNN		x	x	
MCMC	x			x
Variational AutoEncoder		x		x
# Ticks	**2**	**5**	**2**	**4**

References

1. Dhar, V. (2013). "Data science and prediction". Communications of the ACM. 56 (12): 64–73. doi:10.1145/2500499.
2. Kolence, K. W. (1973). "The Software Empiricist". ACM SIGMETRICS Performance Evaluation Review. 2 (2). doi:10.1145/1113644.1113647.
3. WSU (2016). "Data Analytics", URL https://data-analytics.wsu.edu/197-2/, retrieved on April 30, 2019.
4. Bischof, G. (2015). Correlation between engineering students' performance in mathematics and academic success, 122nd ASEE Annual Conference & Exposition, paper 12476.
5. Ooi, A. (2007) An Analysis of the Teaching of Mathematics in Undergraduate Engineering Courses, Proceedings of the 2007 AaeE Conference, Melbourne. https://conference.eng.unimelb.edu.au/aaee2007/papers/paper-69.pdf

Sufficient conditions for càdlàg sample paths of stable superpositions of Ornstein–Uhlenbeck processes

Andreas Basse-O'Connor
Department of Mathematics, Aarhus University, Aarhus, Denmark

Abstract
In this paper we derive sufficient conditions for stable superpositions of Ornstein–Uhlenbeck processes to have cádlág sample paths with probability one.

Keywords
SupOU processes; cádlág; sample path properties; stable processes

1. Introduction

The aim of the present paper is to obtain regularity results for stable superpositions of Ornstein–Uhlenbeck (supOU) processes. These processes are, in general, non-Markovian and non-semi martingale and have a complicated dependence and path structure. The two main path regularities used in the context of stochastic processes are continuity and cádlág. The acronym cádlág comes from the French *à droite, limite à gauche*, which means right continuous with left limits, and in particular, continuity implies cádlág. Since stable supOU processes never have continuous sample paths (except in the Gaussian case), it is natural to focus on when they have cádlág sample paths, which is the content of Theorem 3.1.

In the following we will introduce supOU processes. A stable Ornstein–Uhlenbeck (OU) process $X = \{X(t): t \geq 0\}$ is the solution to the stochastic differential equation

$$X(t) = \int_0^t -\lambda X(s)\, ds + L(t), \qquad (1.1)$$

where $L = \{L(t) : t \geq 0\}$ is a stable L´evy process and $\lambda > 0$. An Ornstein–Uhlenbeck process is both Markovian, a semi martingale and has càdlàg sample paths. Furthermore, the stationary solution to (1.1), which always exists, can be represented on the form

$$X(t) = \int_{-\infty}^t e^{-\lambda(t-s)}\, dL_s, \qquad (1.2)$$

where L is extended to a two-sided L´evy process $\{L(t): t \in R\}$. Going beyond Ornstein–Uhlenbeck processes, convex combinations of independent Ornstein–Uhlenbeck processes (superpositions) are often used as models for

more complex phenomena, which are not captured by ordinarily Ornstein–Uhlenbeck processes. That is, we consider a process X given by

$$X(t) = \sum_{i=1}^{n} r_i X_i(t) = \sum_{i=1}^{n} r_i \int_{-\infty}^{t} e^{-\lambda_i (t-s)} \, dL_i(s) \qquad (1.3)$$

where $n \in \mathbb{N}$, X_i for $i = 1, \ldots, n$ are independent α-stable Ornstein–Uhlenbeck processes with parameters $\lambda_i > 0$, and $r_i \geq 0$ for $i = 1, \ldots, n$ are positive numbers satisfying $r_1 + \cdots + r_n = 1$. Generalizing this idea we arrive at the definition of superposition of Ornstein–Uhlenbeck processes (cf. Barndorff-Nielsen (2000)):

Definition 1.1. Let m denote a probability measure on $\mathbb{R}_+$ and $\alpha \in (0,2]$. Then X is called an α-stable superposition of Ornstein–Uhlenbeck (supOU) process if it has a representation of the form

$$X(t) = \int_{(-\infty,t] \times \mathbb{R}_+} e^{-\lambda(t-s)} \, \Lambda(ds, d\lambda), \qquad t \in \mathbb{R}, \qquad (1.4)$$

where Λ is a symmetric α-stable random measure on $\mathbb{R}^2$ with control measure $dsm(d\lambda)$.

We note that the supOU process X, given in (1.4), is well-defined if and only if $\int_0^1 \lambda^{-1} m(d\lambda) < \infty$, cf. Fasen and Klüppelberg (2007, page 343). The ordinary Ornstein– Uhlenbeck process (1.2) corresponds to $m = \delta_\lambda$, and the more general case (1.3) corresponds to $m = \sum_{i=1}^{n} r_i \delta_{\lambda i}$. Throughout the paper δ_x denotes the Dirac measure at $x \in \mathbb{R}$ given by $\delta_x(A) = \mathbf{1}_A(x)$, and $\mathbb{R}_+ := (0, \infty)$. An α-stable supOU process is a stationary α-stable process, however, it is neither Markovian nor semi martingale in general, opposite to ordinary OU processes. Moreover, supOU processes provide a flexible framework for modelling long-range dependence, let e.g. m be the $\Gamma(r,1)$-law where $r \in (0,1)$, cf. Barndorff-Nielsen (2000, Example 3.1), which is also opposite to ordinary OU processes. In the next section we will prove the main result of the paper (Theorem 3.1), which says that *if m has finite $r-th$ moment for some $r > 0$ and $\alpha \in (1,2)$, then the α-stable supOU process has càdlàg sample paths.*

2. Methodology

To show that an α-stable supOU process X, given in Definition 1.1, has càdlàg sample paths we decompose it as

$$X(t) = Y(t) + Z(t), \qquad (2.1)$$

where

$$Y(t) = \int_{(0,t] \times \mathbb{R}_+} e^{-\lambda(t-s)} \, \Lambda(ds, d\lambda), \qquad Z(t) = \int_{(-\infty,0] \times \mathbb{R}_+} e^{-\lambda(t-s)} \, \Lambda(ds, d\lambda). \qquad (2.2)$$

Our treatment of the two processes Y and Z in (2.1) requires different techniques. In particularly, under a moment condition on m, we will show, in

Lemma 2.1, that Y has cádlág sample paths, and in Lemma 2.2, we will show that Z has continuous samples paths. For the basic properties of stable random variables/processes/measures we refer to Samorodnitsky and Taqqu (1994). Throughout this section C will denote a finite constant which might vary from line to line.

Lemma 2.1. *Suppose that the measure m has finite $r-th$ moment for some $r > 0$. Then the process Y, given in* (2.2)*, has càdlàg sample paths almost surely.*

Proof. To ease the notation let $\mu(ds, d\lambda) = dsm(d\lambda)$ and $f_t(s,\lambda) = e^{\lambda(t-s)}\mathbf{1}_{(0,t]}(s)$. Moreover, fix $T > 0$. To prove the lemma, we will verify the existence of $p_1 > \alpha, p_2 > \alpha/2$, $\beta_1 > 1/2$ and $\beta_2 > 1$ such that

$$\int |f_{t_2} - f_{t_1}|^{p_1}\, d\mu \le C(t_2 - t_1)^{\beta_1} \tag{2.3}$$

$$\int |(f_{t_2} - f_t)(f_t - f_{t_1})|^{p_2}\, d\mu \le C(t_2 - t_1)^{\beta_2}, \tag{2.4}$$

hold for all $0 \le t_1 \le t \le t_2 \le T$, which due to Theorem 4.3 of Basse-O'Connor and Rosin´ski (2013) yields that Y has càdlàg sample paths with probability one. Let $\lambda > 0$ and $s \in (0,t]$. To verify (2.4) we use the mean-value theorem to obtain

$$\begin{aligned}
&|(f_{t_2}(s,\lambda) - f_t(s,\lambda))(f_t(s,\lambda) - f_{t_1}(s,\lambda))| \\
&\quad = |(e^{-\lambda(t_2-s)} - e^{-\lambda(t-s)})(e^{-\lambda(t-s)} - e^{-\lambda(t_1-s)})|\mathbf{1}_{(0,t_1]}(s) \\
&\qquad + |(e^{-\lambda(t_2-s)} - e^{-\lambda(t-s)})e^{-\lambda(t-s)}\mathbf{1}_{(t_1,t]}(s) \\
&\quad \le C\Big(\min\{|\lambda(t_2-t)|, e^{-\lambda(t-s)}\}\mathbf{1}_{(0,t_1]}(s) + \min\{|\lambda(t_2-t_1)|, 1\}\mathbf{1}_{(t_1,t]}(s)\Big) \\
&\quad = C\Big(g_1(s,\lambda) + g_2(s,\lambda)\Big),
\end{aligned} \tag{2.5}$$

Where

$$\begin{aligned}
g_1(s,\lambda) &= \min\{|\lambda(t_2-t)|, e^{-\lambda(t-s)}\}\mathbf{1}_{(0,t_1]}(s), \\
g_2(s,\lambda) &= \min\{|\lambda(t_2-t_1)|, 1\}\mathbf{1}_{(t_1,t]}(s).
\end{aligned} \tag{2.6}$$

First we will consider the g_1-term from (2.5). Since $0 \le t_1 \le t$ we have that

$$\int_0^{t_1} e^{-\lambda(t-s)p_2}\, ds = \frac{e^{-\lambda(t-t_1)p_2} - e^{-\lambda t p_2}}{\lambda p_2} \le C\lambda^{-1},$$

which implies that for all $\kappa > 0$,

$$\begin{aligned}
\int |g_1|^{p_2}\, d\mu &= \int_0^\infty \Big(\int_0^{t_1} \min\{|\lambda(t_2-t)|^{p_2}, e^{-\lambda(t-s)p_2}\}\, ds\Big)\, m(d\lambda) \\
&\le C\Big((t_2-t)^{p_2}\int_0^\kappa \lambda^{p_2}\, m(d\lambda) + \int_\kappa^\infty \lambda^{-1}\, m(d\lambda)\Big).
\end{aligned} \tag{2.7}$$

Since m has finite $r-th$ moment we have for all $p_2 \geq r$ that

$$\int_0^\kappa \lambda^{p_2}\, m(d\lambda) \leq \int_0^\kappa \lambda^{p_2}(\kappa\lambda^{-1})^{p_2-r}\, m(d\lambda) \leq \Big(\int_0^\infty \lambda^r\, m(d\lambda)\Big)\kappa^{p_2-r} \leq C\kappa^{p_2-r}, \quad (2.8)$$

and similarly,

$$\int_\kappa^\infty \lambda^{-1}\, m(d\lambda) \leq \int_\kappa^\infty \lambda^{-1}(\lambda\kappa^{-1})^{1+r}\, m(d\lambda) \leq \Big(\int_0^\infty \lambda^r\, m(d\lambda)\Big)\kappa^{-1-r} \leq C\kappa^{-1-r}. \quad (2.9)$$

First we choose a number q satisfying $(1+r)^{-1} < q < 1$, which is possible since $r > 0$. Next, we choose $p_2 > max\{\alpha/2, r\}$ such that $p_2(1-q) + qr > 1$, which is possible since $q < 1$. By setting $\kappa = (t_2 - t_1)^{-q}$ we obtain the following estimate from (2.7), (2.8) and (2.9)

$$\begin{aligned}\int |g_1|^{p_2}\, d\mu &\leq C\Big((t_2-t_1)^{p_2}\kappa^{p_2-r} + \kappa^{-1-r}\Big) \\ &\leq C\Big((t_2-t_1)^{p_2(1-q)+qr} + (t_2-t_1)^{q(1+r)}\Big) \leq C(t_2-t_1)^{1+\delta}, \end{aligned} \quad (2.10)$$

for some $\delta > 0$ only depending on r, p_2 and q.

For the g_2 –term from (2.5) we use the following estimate

$$\begin{aligned}\int |g_2|^{p_2}\, d\mu &\leq C(t_1-t)\int_0^\infty \min\{|\lambda(t_2-t_1)|^{p_2}, 1\}\, m(d\lambda) \\ &\leq C(t_2-t_1)^{1+r}\int_0^\infty \lambda^r\, m(d\lambda) \leq C(t_2-t_1)^{1+r} \end{aligned} \quad (2.11)$$

where the second inequality follows by $min\{|x|^{p_2}, 1\} \leq |x|^r$ for all $x \in \mathbb{R}$, which holds since $0 \leq r \leq p_2$. From (2.5), (2.10) and (2.11) we obtain (2.4).

In the following we will prove (2.3). By recalling the definition of the function g_1 in (2.6), in the case $t = t_1$, we have by the mean-value theorem

$$\begin{aligned}&|f_{t_2}(s,\lambda) - f_{t_1}(s,\lambda)| \\ &\quad = |e^{-\lambda(t_2-s)} - e^{-\lambda(t_1-s)}|\mathbf{1}_{(0,t_1]}(s) + e^{-\lambda(t_2-s)}\mathbf{1}_{(t_1,t_2]}(s) \\ &\quad \leq \min\{|\lambda(t_2-t_1)|, e^{-\lambda(t_1-s)}\}\mathbf{1}_{(0,t_1]}(s) + \mathbf{1}_{(t_1,t_2]}(s) \\ &\quad = g_1(s,\lambda) + \mathbf{1}_{(t_1,t_2]}(s). \end{aligned}$$

Hence, if we set $p_1 = p_2$ with p_2 chosen according to (2.10), we have according to (2.10) that

$$\begin{aligned}\int |f_{t_2} - f_{t_1}|^{p_1}\, d\mu &\leq C\Big(\int |g_1|^{p_1}\, d\mu + \int \mathbf{1}_{(t_1,t_2]}(s)\, \mu(ds, d\lambda)\Big) \\ &\leq C\Big((t_2-t_1)^{1+\delta} + (t_2-t_1)\Big) \leq C(t_2-t_1), \end{aligned}$$

which completes the proof of (2.3), and hence the lemma.

Lemma 2.2. *Let Z be given by* (2.2) *and assume that $\alpha \in (1,2)$ and that the measure m has finite $r-th$ moment for some $r > 0$. Then Z has continuous sample paths with probability one.*

Proof. For each $\alpha-$stable random variable U let $\|U\|_\alpha$ denote the scale parameter of U. By the isometric property of stable integrals, cf. Proposition 3.4.1 of Samorodnitsky and Taqqu (1994), we have that

$$\|Z(t)-Z(u)\|_\alpha^\alpha = \int_0^\infty \left(\int_{-\infty}^0 \left| e^{-\lambda(t-s)} - e^{-\lambda(u-s)} \right|^\alpha ds \right) m(d\lambda). \tag{2.12}$$

Choose $q \in [0,1]$ such that $1/\alpha < q \leq (r+1)/\alpha$, which is possible since $r > 0$ and $\alpha > 1$.
By the inequality

$$|e^{-\lambda(t-s)} - e^{-\lambda(u-s)}| \leq C e^{\lambda s} \min\{|\lambda(t-u)|, 1\},$$

and (2.12) we deduce that

$$\begin{aligned}
\|Z(t)-Z(u)\|_\alpha^\alpha &\leq C \int_0^\infty \left(\int_{-\infty}^0 e^{\alpha\lambda s}\, ds \right) \min\{|\lambda(t-u)|^\alpha, 1\}\, m(d\lambda) \\
&\leq C \int_0^\infty \lambda^{-1} \min\{|\lambda(t-u)|^\alpha, 1\}\, m(d\lambda) \\
&\leq C \int_0^\infty \lambda^{-1} |\lambda(t-u)|^{q\alpha}\, m(d\lambda) = C|t-u|^{q\alpha} \int_0^\infty \lambda^{q\alpha-1}\, m(d\lambda) \\
&\leq C|t-u|^{q\alpha}
\end{aligned}$$

where we have used that $min\{|x|,1\} \leq |x|^q$ for all $x \in \mathbb{R}$, in the third inequality (recall $q \in [0,1]$), and that $0 < q\alpha - 1 \leq r$ in the last inequality together with the fact that m is a finite measure. Hence,

$$\|Z(t) - Z(u)\|_\alpha \leq C|t-u|^q. \tag{2.13}$$

Fix $T > 0$ and consider the metric $d(t,u) = \|Z(t) - Z(u)\|_\alpha$ on $[0,T]$ induced by the $\alpha-$stable process Z. For each > 0 we let $N(\epsilon)$ denote the smallest number of open $d-$balls of radius needed to cover $[0,T]$. From the estimate (2.13) we deduce that

$$N(\epsilon) \leq C\epsilon^{-1/q},$$

and hence

$$\int_0^T N(\epsilon)^{1/\alpha}\, d\epsilon \leq C \int_0^T \epsilon^{-1/(\alpha q)}\, d\epsilon < \infty \tag{2.14}$$

where the last inequality follows by the fact that $q > 1/\alpha$. By (2.14) and Theorem 12.2.1 of Samorodnitsky and Taqqu (1994) we deduce that Z has continuous sample paths with probability one, which concludes the proof. □

3. Result

In this section we state the main result of the paper, which gives sufficient conditions for stable supOU processes to have càdlàg sample paths.

Theorem 3.1. *Let X be an $\alpha-$stable supOU process given in Definition 1.1 with $\alpha \in (1,2)$ and such that the measure m has finite $r-$th moment for some $r > 0$. Then X has c`adl`ag sample paths with probability one.*

Proof. The theorem follows by the decomposition (2.1) and Lemmas 2.1 and 2.2. □

4. Discussion and Conclusion

In the following we fix $\alpha \in (1,2)$ and let X be an α −stable supOU process given in Definition 1.1. We recall that due to the discontinuity of the integrand $t \mapsto e^{-\lambda(t-s)}\mathbf{1}_{(-\infty,t]}(s)$ in (1.4) at $t = s$, the process X will never have continuous sample paths, cf. Rosiński (1989). However, the above Theorem 3.1 shows that if m has finite $r-th$ moment for some $r > 0$ then X has càdlàg sample paths with probability one. To compare this condition, with the literature we note that Example 4.1 in Basse-O'Connor and Rosiński (2016), shows that X is a semi martingale if and only if m has finite $(\alpha - 1) - th$ moment, and if X is a semi martingale then it has càdlàg sample paths. Since $\alpha - 1 > 0$ in our setting, the conditions in Theorem 3.1 are weaker than the ones we can derive from Basse-O'Connor and Rosiński (2016).

References

1. Barndorff-Nielsen, O. E. (2000). Superposition of Ornstein-Uhlenbeck type processes. *Teor. Veroyatnost. i Primenen. 45*(2), 289–311.
2. Basse-O'Connor, A. and J. Rosiński (2013). On the uniform convergence of random series in Skorohod space and representations of c`adl`ag infinitely divisible processes. *Ann. Probab. 41*(6), 4317–4341.
3. Basse-O'Connor, A. and J. Rosiński (2016). On infinitely divisible semimartingales. *Probab. Theory Related Fields 164*(1-2), 133–163.
4. Fasen, V. and C. Klüppelberg (2007). Extremes of supOU processes. In *Stochastic analysis and applications*, Volume 2 of *Abel Symp.*, pp. 339–359. Springer, Berlin.
5. Rosiński, J. (1989). On path properties of certain infinitely divisible processes. *Stochastic Process. Appl. 33*(1), 73–87.
6. Samorodnitsky, G. and M. S. Taqqu (1994). *Stable Non-Gaussian Random Processes*. Stochastic Modeling. New York: Chapman & Hall. Stochastic models with infinite variance.

Extreme value theory for long range dependent stable random fields

Zaoli Chen, Gennady Samorodnitsky
Cornell University, USA

Abstract
We study the extremes for a class of a symmetric stable random fields with long range dependence.

Keywords
Random field; extremal limit theorem; random sup measure; random closed set; long range dependence; stable law; heavy tails.

1. Introduction

Extreme value theorems describe the limiting behaviour of the largest values in increasingly large collections of random variables. The classical extremal theorems, beginning with Fisher and Tippett (1928) and Gnedenko (1943), deal with the extremes of i.i.d. (independent and identically distributed) random variables. The modern extreme value theory techniques allow us to study the extremes of dependent sequences; see Leadbetter et al. (1983) and the expositions in Coles (2001) and de Haan and Ferreira (2006). The effect of dependence on extreme values can be restricted to a loss in the effective sample size, through the extremal index of the sequence. When the dependence is sufficiently long, more significant changes in extreme value may occur; see e.g. Samorodnitsky (2004), Owada and Samorodnitsky (2015b). The present paper aims to contribute to our understanding of the effect of memory on extremes when the time is of dimension larger than 1, i.e. for random fields.

We consider a discrete time stationary random field $X = \left(X_t, \boldsymbol{t} \in \mathbb{Z}^d\right)$. For $\boldsymbol{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$ we would like to study the extremes of the random field over growing hypercubes of the type

$$[\mathbf{0}, \mathbf{n}] = \left\{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}\right\}, \ \mathbf{n} \to \boldsymbol{\infty},$$

1991 *Mathematics Subject Classification.* Primary 60G60, 60G70, 60G52.

Key words and phrases. Random field, extremal limit theorem, random sup measure, random closed set, long range dependence, stable law, heavy tails.

This research was partially supported by the NSF grant DMS-1506783 and the ARO grant W911NF-18 - 10318 at Cornell University.

where $\mathbf{0}$ is the vector with zero coordinates, the notation $\boldsymbol{s} \leq \boldsymbol{t}$ for vectors $\boldsymbol{s} = (s_1, \ldots, s_d)$ and $\boldsymbol{t} = (t_1, \ldots, t_d)$ means that $s_i \leq t_i$ for all $i = 1, \ldots, d$, and the notation $\boldsymbol{n} \to \infty$ means that all d components of the vector $\mathbf{n}$ tend to infinity. Denote

$$M_{\mathbf{n}} = \max_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}} X_{\mathbf{t}} .$$

What limit theorems does the array $(M_{\boldsymbol{n}})$ satisfy? It was shown by Leadbetter and Rootzén (1998) that under appropriate strong mixing conditions, only the classical three types of limiting distributions (Gumbel, Fréchet and Weibull) may appear (even when forcing $\boldsymbol{n} \to \infty$ along a monotone curve). In the case when the marginal distributions of the field $\boldsymbol{X}$ have regularly varying tails, this allows only the Fréchet distribution as a limit.

In this paper we will discuss only random fields with regularly varying tails, in which case the experience from the classical extreme value theory tells us to look for limit theorems for the type

$$\frac{1}{b_{\mathbf{n}}} M_{\mathbf{n}} \Rightarrow Y \quad \text{as } \mathbf{n} \to \boldsymbol{\infty} \tag{1.1}$$

for some nondegenerate random variable Y. The regular variation of the marginal distributions means that

$$P(X(\mathbf{0}) > x) = x^{-\alpha} L(x),\ \alpha > 0,\ L \text{ slowly varying}, \tag{1.2}$$

see e.g. Resnick (1987). Notice that the assumption is only on the right tail of the distribution since, in most cases, one does not expect a limit theorem for the partial maxima as in (1.1) to be affected by the left tail of $X(\mathbf{0})$.

If the random field $\boldsymbol{X}$ consists of i.i.d. random variables satisfying the regular variation condition (1.2), then the classical extreme value theory tells us that the convergence in (1.1) holds if we choose

$$b_n = \inf\{x > 0 : P(X(\mathbf{0}) > x) \leq (n_1 \ldots n_d)^{-1}\}, \tag{1.3}$$

in which case the limiting random variable Y has the standard Fréchet distribution. We are interested in understanding how the spatial dependence in the random field $\boldsymbol{X}$ affects the scaling in and the distribution of the limit not only in (1.1), but in its functional versions, which can be stated in different spaces, for example in the space $D(\mathbb{R}_+^d)$ of right continuous, with limits along monotone paths, functions (see Straf (1972)), or in the space of random sup measures $\mathcal{M}(\mathbb{R}_+^d)$; see O'Brien et al. (1990). We will describe the relevant spaces below.

If the time is one-dimensional, and the memory in the stationary process is short, then the standard normalization (1.3) is still the appropriate one, and the limits both in (1.1) and its functional versions change only through a change in a multiplicative constant; see Samorodnitsky (2016) and references therein. However, when the memory becomes sufficiently long, both the order

of magnitude of the normalization in the limit theorems changes, and the nature of the limit changes as well; see Samorodnitsky (2004) and Owada and Samorodnitsky (2015b). Furthermore, the limit may even stop having the Fréchet distribution (or Fréchet marginal distributions, in the functions limit theorems); see Samorodnitsky and Wang (2017). It is reasonable to expect that similar phenomena happen for random fields, but because it is harder to quantify how long the memory is when the time is not one-dimensional, less is known in this case.

In this paper we will concentrate on the case where the random field $\boldsymbol{X}$ is a symmetric α-stable ($S\alpha S$) random field, $0 < \alpha < 2$. Recall that this means that every finite linear combination of the values of the values of the random field has a one-dimensional SαS distribution, i.e. has a characteristic function of the form $\exp\{-\sigma^{\alpha}|\theta|^{\alpha}\}, \theta \in \mathbb{R}$, where $\sigma \in [0, \infty)$ is a scale parameter that depends on the linear combination; see Samorodnitsky and Taqqu (1994). The marginal distributions of $S\alpha S$ random fields satisfy the regular variation assumption (1.2) with $0 < \alpha < 2$ that coincides with the index of stability. In this case a series of results on the relation between the sizes of the extremes of stationary SαS random fiels and certain ergodic-theoretical properties of the Lévy measures of these fields is due to Parthanil Roy and his coworkers; see Roy and Samorodnitsky (2008), Chakrabarty and Roy (2013), Sarkar and Roy (2016). These results are made possible because of the connection between the structure of the $S\alpha S$ random fields and ergodic theory established by Rosiński (2000).

This paper contributes to understanding the extremal limit theorems for $S\alpha S$ random fields and their connection to the dynamics of the Lévy measures. In this sense our paper is related to the ideas of Rosiński (2000). However, we will restrict ourselves to certain Markov flows. This will allow us to avoid, to a large extent, the language of ergodic theory, and state everything in purely probabilistic terms. There is not doubt, however, that our results could be extended to more general dynamical systems acting on the Lévy measures of $S\alpha S$ random fields. The generality in which work is sufficient to demonstrate the new phenomena that may arise in extremal limit theorems for random fields with long range dependence. We will exhibit new types of limits, some of which will have non-Fréchet distributions, both in the space of random sup measures and in the space $D(\mathbb{R}_+^d)$.

2. A $S\alpha S$ random field with long range dependence

We start with a construction of a family of stationary SαS random fields, $0 < \alpha < 2$, whose memory has a natural finite-dimensional parameterization. It is an extension to random fields of models considered before in the case of one-dimensional time; see e.g. Resnick et al. (2000),

Samorodnitsky (2004), Owada and Samorodnitsky (2015a,b), Owada (2016) and Lacaux and Samorodnitsky (2016).

We start with d σ-finite, infinite measures on $(\mathbb{Z}^{\mathbb{N}_0}, \mathcal{B}(\mathbb{Z}^{\mathbb{N}_0}))$ defined by

$$\mu_i := \sum_{k\in\mathbb{Z}} \pi_k^{(i)} P_k^{(i)}, \tag{2.1}$$

where for $i = 1, \dots, d, P_k^{(i)}$ is the law of an irreducible aperiodic null-recurrent Markov chain $(Y_n^{(i)})_{n\geq 0}$ on $\mathbb{Z}$ starting at $Y_0^{(i)} = k \in \mathbb{Z}$. Further, $(\pi_k^{(i)})k \in \mathbb{z}$ is its unique (infinite) invariant measure satisfying $\pi_0^{(i)} = 1$. Given this invariant measure, we can extend the probability measures $P_k^{(i)}$ from measures on $\mathbb{Z}^{\mathbb{N}_0}$ to measures on $\mathbb{Z}^{\mathbb{Z}}$ which, in turn, allows us to extend the measure μ_i in (2.1) to $\mathbb{Z}^{\mathbb{Z}}$ as well. We will keep using the same notation as in (2.1).

We will work with the product space

$$(E, \mathcal{E}) = \left(\mathbb{Z}^{\mathbb{Z}} \times \cdot \times \mathbb{Z}^{\mathbb{Z}},\ \mathcal{B}(\mathbb{Z}^{\mathbb{Z}}) \times \cdot \times \mathcal{B}(\mathbb{Z}^{\mathbb{Z}})\right)$$

of d copies of $\left(\mathbb{Z}^{\mathbb{Z}}, \mathcal{B}(\mathbb{Z}^{\mathbb{Z}})\right)$, on which we put the product, σ-finite, infinite, measure

$$\mu = \mu_1 \times \cdot \times \mu_d .$$

The key assumption is a regular variation assumption on the return times of the Markov chains $(Y_n^{(i)})_{n\geq 0}, i = 1, \dots, d$. For $\boldsymbol{x} = (\dots, x_{-1}, x_0, x_1, x_2, \dots) \in \mathbb{Z}^z$ we define the first return time to the origin by $\varphi(x) = \inf\{n \geq 1 : x_n = o\}$. We assume that for $i = 1, \dots, d$ we have

$$P_0^{(i)}(\varphi > n) \in RV_{-\beta_i} \tag{2.2}$$

for some $0 < \beta i_i < 1$. This implies that

$$\mu(\{x : x_k = 0 \text{ for some } k = 0, 1, \dots, n\}) \tag{2.3}$$

$$\sim \sum_{k=1}^{n} P_0^{(i)}(\varphi > k) \sim (1 - \beta_i)^{-1} n P_0^{(i)}(\varphi > n) \in RV_{1-\beta_i} .$$

See Resnick et al. (2000).

On $\mathbb{Z}^{\mathbb{Z}}$ there is a natural left shift operator

$$T\left((\dots, x_{-1}, x_0, x_1, x_2 \dots)\right) = (\dots, x_0, x_1, x_2, x_3 \dots).$$

It is naturally extended to a group action of $\mathbb{Z}^d$ on E as follows. Writing an element $x \in E\ as\ x = \left(x^{(1)}, \dots, x^d\right)$ with $x^i = (\dots, x_{-1}^{(i)}, x_1^{(i)}, x_2^{(i)}, \dots \in \mathbb{Z}^{\mathbb{Z}})$ for $i = 1, \dots, d$, we set for $\boldsymbol{n} = (n_1, \dots, n_d,) \in \mathbb{Z}^d$,

$$T^n \mathrm{x} = \left(T^{n1} x^{(1)}, \dots . T^{nd} x^{(d)}\right) E . \tag{2.4}$$

Even though we are using the same notation T for operators acting on different spaces, the meaning will always be clear from the context. Note that each individual left shift T on $(\mathbb{Z}^{\mathbb{N}_o}, \mathcal{B}(\mathbb{Z}^{\mathbb{N}_o}), \mu_j)$ is measure preserving (because each $(\pi_i^{(j)})_{i\in\mathbb{Z}}$ is an invariant measure.) It is also conservative and ergodic by Theorem 4.5.3 in Aaronson (1997). Therefore, the group action $\mathcal{T} = \{T^n : n \in \mathbb{Z}^d\}$ is conservative, ergodic and measure preserving on $(E, \mathcal{E}, \mu)$.

Equipped with a measure preserving group action on the space $(E, \mathcal{E})$ we can now define a stationary symmetric α-stable random field by

(2.5) $$X_n = \int_E f \circ T^n(x) M(dx), n \in \mathbb{Z}^d,$$

where M is a $S\alpha S$ random measure on$(E, \mathcal{E})$ with control measure μ, and

(2.6) $$f(x) = 1\big(x^{(i)} \in A, i = 1, \dots, d\big), x = (x^{(1)}, \dots, x^{(d)}).$$

where $A = \{x \in \mathbb{Z}^{\mathbb{Z}} : x_0 = 0\}$. Clearly, $f \in L^{\propto}(\mu)$, which guarantees that the integral in (2.5) is well defined. We refer the reader to Samorodnitsky and Taqqu (1994) for general information on stable processes and integrals with respect to stable measures, and to Rosiński (2000) on more details on stationary stable random fields and their representations.

The random field model defined by (2.5) is attractive because the key parameters involve in its definition have a clear intuitive meaning: the index of stability $0 < \alpha < 2$ is responsible for the heaviness of the tails, while $0 < \beta_i < 1, i = 1, \dots, d$ (defined in (2.2)) are responsible for the "length of the memory". The latter claim is not immediately obvious, but its (informal) validity will become clearer in the sequel.

The following array of positive numbers will play the crucial role in the extremal limit theorems in this paper. Denote for $n = 1, 2, \dots$ and $i = 1, \dots, d$

$$b_n^{(i)} = (\mu_i(\{\mathrm{x}: x_k = 0 \text{ for some } k = 0, 1, \dots, \mathrm{n}\}))^{1/\alpha},$$

and let

(2.7) $$b_{\mathbf{n}} = \prod_{i=1}^{d} b_{n_i}^{(i)}, \ \mathbf{n} = (n_1, \dots, n_d) \in \mathbb{N}_0^d.$$

Then $b_n^{\alpha} = \mu(B_n)$, where

$B_n = \{\mathrm{x} = (\mathrm{x}^{(1)}, \dots, \mathrm{x}^{(\mathrm{d})}) \in E : x_{k_i}^{(i)} = 0$ for some $0 \le k_i \le n_i$, each $i = 1, \dots, d\}$.

Therefore, we can define, for each $\mathbf{n} \in \mathbb{N}_o^d$, a probability measures η_{n} on $(E, \mathcal{E})$ by

(2.8) $$\eta_{\mathrm{n}}(\cdot) = b_{\mathrm{n}}^{-\alpha} \mu(\cdot \cap B\_n).$$

This probability measure allows us to represent the restriction of the stationary $S\alpha S$ random field $\mathbf{X}$ in (2.5) to the hypercube $[\mathbf{0}, \mathbf{n}] = \{\mathbf{0} \le \mathbf{k} \le \mathbf{n}\}$ as a series, described below, and that we will find useful in the sequel. It is useful to note also that the measure $\eta_{\mathbf{n}}$ is the product measure of d probability measures

on$\left(\mathbb{Z}^{\mathbb{Z}}, \mathcal{B}\left(\mathbb{Z}^{\mathbb{Z}}\right)\right)$: $\eta_{\boldsymbol{n}} = \eta_{n1}^{(1)} \times . \times \eta_{n_d}^{(d)}$ for $\mathbf{n} = (n_1, \ldots, n_n) \in \mathbb{N}_o^d$,for where for $i = 1, \ldots, d$ and $n \geq 0$,

$$\eta_n^{(i)}(\cdot) = \left(b_n^{(i)}\right)^{-\alpha} \mu_i(\cdot \cap \{\mathrm{x} \in \mathbb{Z}^{\mathbb{Z}} : x_k = 0 \text{ for some } 0 \leq k \leq n\}). \tag{2.9}$$

The restriction of the stationary *SαS* random field **X** in (2.5) to the hypercube [**0**,**n**] admits, in law, the series representation

$$X_{\mathbf{k}} = b_{\mathbf{n}} C_\alpha^{1/\alpha} \sum_{j=1}^{\infty} \epsilon_j \Gamma_j^{-1/\alpha} \mathbf{1}_{A^d} \circ T^{\mathbf{k}}(U_{j,\mathbf{n}}), \ \mathbf{0} \leq \mathbf{k} \leq \mathbf{n}, \tag{2.10}$$

with $A^d = A \times \cdots \times A$ the direct product of d copies of A and A is in (2.6), where the constant C_α is the tail constant of the α-stable random variable:

$$C_\alpha = \left(\int_0^\infty x^{-\alpha} \sin x dx \right)^{-1} = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} & \alpha \neq 1 \\ 2/\pi & \alpha = 1 \end{cases}$$

Furthermore,$\{\epsilon_j\}$ is a iid sequence of Rademacher random variables, $\{\Gamma_j\}$ is the sequence of the arrival times of a unit rate Poisson process on $(0, \infty)$, and $\{U_{j,n}\}$ are iid E-valued random elements with common law η_n. The sequence $\{\epsilon_j\}$,$\{\Gamma_j\}$ and $\{U_{j,n}\}$ are independent. See Samorodnitsky and Taqqu (1994) for details.

References

1. J. Aaronson (1997): *An Introduction to Infinite Ergodic Theory*, volume 50 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence.
2. Chakrabarty and P. Roy (2013): Group theoretic dimension of stationary symmetric α-stable random fields. *Journal of Theoretical Probability* 26:240–258.
3. S. Coles (2001): *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York.
4. L. de Haan and A. Ferreira (2006): *Extreme Value Theory: An Introduction*. Springer, New York.
5. R. A. Fisher and L. Tippett (1928): Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings of Cambridge Philisophical Society* 24:180–190.
6. Gnedenko (1943): Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics* 44:423–453.
7. Lacaux and G. Samorodnitsky (2016): Time-changed extremal process as a random sup measure. *Bernoulli* 22:1979–2000.

8. M. Leadbetter, G. Lindgren and H. Rootzén (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer Verlag, New York.
9. M. Leadbetter and H. Rootzén (1998): On extremes values in stationary random fields. In *Stochastic Processes and related Topics*. Birkhäuser, Boston, pp. 275–285.
10. G. O'Brien, P. Torfs and W. Vervaat (1990): Stationary self-similar extremal processes. *Probability Theory and Related Fields* 87:97–119.
11. T. Owada (2016): Limit theory for the sample autocovariance for heavy tailed stationary infinitely divisible processes generated by conservative flows. *Journal of Theoretical Probability* 29:63–95.
12. T. Owada and G. Samorodnitsky (2015a): Functional Central Limit Theorem for heavy tailed stationary infinitely divisible processes generated by conservative flows. *Annals of Probability* 43:240–285.
13. T. Owada and G. Samorodnitsky (2015b): Maxima of long memory stationary symmetric α-stable processes, and self-similar processes with stationary max-increments. *Bernnoulli* 21:1575–1599.
14. S. Resnick (1987): *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.
15. S. Resnick, G. Samorodnitsky and F. Xue (2000): Growth rates of sample covariances of stationary symmetric α-stable processes associated with null recurrent Markov chains. *Stochastic Processes and Their Applications* 85:321–339.
16. J. Rosiński (2000): Decomposition of stationary α–stable random fields. *Annals of Probability* 28:1797–1813.
17. P. Roy and G. Samorodnitsky (2008): Stationary symmetric α-stable discrete parameter random fields. *Journal of Theoretical Probability* 21:212–233.
18. G. Samorodnitsky (2004): Extreme value theory, ergodic theory, and the boundary between short memory and long memory for stationary stable processes. *The Annals of Probability* 32:1438–1468.
19. G. Samorodnitsky (2016): *Stochastic Processes and Long Range Dependence*. Springer, Cham, Switzerland.
20. G. Samorodnitsky and M. Taqqu (1994): *Stable Non-Gaussian Random Processes*. Chapman and Hall, New York.
21. G. Samorodnitsky and Y. Wang (2017): Extremal theory for long range dependent infinitely divisible processes. Technical report.
22. S. Sarkar and P. Roy (2016): Stable random fields indexed by finitely generated free groups. *Annals of Probability*, to appear.
23. M. Straf (1972): Weak convergence of stochastic processes with several parameters. In *Proceedings of the Sixth Berkeley Symposium on*

Mathematical Statistics and Probability, volume 2. University of California Press, Berkeley, CA, pp. 187–221

Infinitely divisible processes and their random translations

Jan Rosinski´
University Of Tennessee, Usa

Abstract

The celebrated Cameron-Martin formula relates a Gaussian process to its translation by a deterministic function and shows an isomorphism between moments of nonlinear functionals of the original process and the translated one. We show that a similar phenomenon occurs for more general infinitely divisible processes when we allow random translations. Precise understanding of Lévy measures of processes on path spaces is the key to produce rich classes of admissible random translations. We illustrate this approach on examples of squared Bessel processes, Feller diffusions, permanental processes, as well as Lévy processes.

Keyword

path Lévy measure; admissible shifts; permanental processes; Lévy processes.

1. Introduction

Let $G = (G_t)_{tET}$ be a centered Gaussian process over a set T. The Cameron-Martin Formula says that for every random variable ξ in the L^2-closure of the subspace spanned by G and for any measurable functional $F: \mathbb{R}^T \mapsto \mathbb{R}$

$$\mathbb{E}[F((G_t + \phi(t))_{t\in T})] = \mathbb{E}\left[F((G_t)_{t\in T})\, e^{\xi - \frac{1}{2}\mathbb{E}\xi^2}\right] \tag{1}$$

Where $\emptyset(t) = \mathbb{E}(\xi G_t)$. This formula has many applications, including stochastic differential equations and stochastic partial driven by Gaussian random fields.

It is well-known that (1) does not extend to the Poissonian case. Indeed, it is easy to show that if $Y = (Y_t)_{t\in[0,1]}$ is a Poisson process, then there is no function $\psi: [0,1] \to \mathbb{R}, \psi \not\equiv 0$, such that

$$\mathbb{E}\left[F\left((Y_t + \psi(t))_{t\in[0,1]}\right)\right] = \mathbb{E}\left[F\left((Y_t)_{t\in[0,1]}\right)\eta\right]$$

for all measurable functionals $F: \mathbb{R}^{[0,1]} \to \mathbb{R}$ and some random variable $\eta \geq 0$ with $\mathbb{E}_\eta = 1$.

There is however a rich class of random translations associated with each infinitely divisible process. By definition, a process $X = (X_t)_{t\in T}$ over a general set T is said to be infinitely divisible if for every $n \geq 2$ there exist i.i.d. processes $(Y_t^{n,i})_{t\in T}, i = 1, \dots, n$ such that

$$(X_t)_{t\in T} \stackrel{d}{=} \left(Y_t^{n,1} + \cdots + Y_t^{n,n}\right)_{t\in T}$$

where $\stackrel{d}{=}$ denotes the equality in finite dimensional distributions. Every infinitely divisible process $X = (X_t)_{t\in T}$ has a Lévy measure ν defined on the path space $\mathbb{R}^T$, which characterizes the nonGaussian part of X. Suppose that a stochastic process $Z = (Z_t)_{t\in T}$ is independent of X and the distribution $\mathcal{L}(Z)$ of Z is absolutely continuous with respect to ν. Then there exists a measurable function $g : \mathbb{R}^T \mapsto \mathbb{R}_+$ such that for any measurable functional $F: \mathbb{R}^T \mapsto \mathbb{R}_+$

$$\mathbb{E}\left[F\left((X_t + Z_t)_{t\in T}\right)\right] = \mathbb{E}\left[F\left((X_t)_{t\in T}\right) g(X)\right]. \tag{2}$$

There are two basic directions of applying identity (2). The first one is to start with a process $Z = (Z_t)_{t\in T}$ of interest, associate with it (possibly easier to handle) infinitely divisible process $X = (X_t)_{t\in T}$ whose Lévy measure dominates the law of Z, and transfer path properties of X to Z via identity (2). Using Dynkin's Isomorphism Theorem, Marcus, M.B., &Rosen, J. (1992), &(2006) derived many results for local times of Markov processes, including Lévy processes. Another direction of applications of (2) is much harder, to derive information about X by utilizing Z. One way to approach it is to consider the "converse" version of (2) which expresses X as the process $X + Z$ with changed measure.

Recall that a stochastic process $X = (X_t)_{t\in T}$ is said to be Poissonian infinitely divisible if its all finite dimensional marginal distributions are infinitely divisible without Gaussian part.

Throughout this short paper, an identity as (2) reads: if one side exists then the other does and they are equal.

2. Methodology

Successful implementation of identities like (2) requires precise understanding of Lévy measures of processes, which are defined on path spaces with the usual cylindrical σ-algebras (as opposed to σ-rings in Lee, P.M. (1967) and Maruyama, G. (1970)). We view Lévy measures as "laws of processes" defined on possibly infinite measure spaces and call such "processes" representations of Lévy measures. Properties of Lévy measures are defined by properties of their representations. Transfer of regularity property puts the Lévy measure on the same Borel function space where paths of the corresponding infinitely divisible processes belong. This allows to relate path properties of processes and representations of their Lévy measures.

Let $\mathbb{R}^T$ be the space of all functions $x : T \mapsto \mathbb{R}$, and let $\mathcal{B}^T$ denote its cylindrical (product) σ −algebra. Let 0_T denotes the origin of $\mathbb{R}^T$, considerd as a point or the one-point set, depending on the context. We give the following definition of path Lévy measure.

Definition 1. A measure ν on $(\mathbb{R}^T, \mathcal{B}^T)$ is said to be a Lévy measure if the following two conditions hold

(L1) for every $t \in T$ $\int_{\mathbb{R}^T} |x(t)|^2 \wedge 1 \nu(dx) < \infty$,
(L2) for every $A \in \mathfrak{B}^T \nu(A) = \nu_*(A \backslash O_T)$, where ν_* is the inner measure.
If only condition (L1) is assumed, then ν is called a pre Lévy measure.

Condition (L1) is a technical one, needed for the integral in the Lévy-Khintchine formula (3) to be well-defined. Condition (L2) gives a rigorous meaning to "ν has no mass at the origin". Indeed, if T is countable, then $O_T \in \mathfrak{B}^T$ and (L2) is equivalent to $\nu(0_T) = 0$, which is the usual condition for Lévy measures. If T is uncountable, then $O_T \notin \mathfrak{B}^T$, so that $\nu(0_T)$ is undefined. However, (L2) still makes sense and it ensures uniqueness of ν. Indeed, we will show that every infinitely divisible process has a unique Lévy measure satisfying the above definition.

We will now give the Lévy-Khintchine representation for an arbitrary infinitely divisible process obtained in Rosinski, J. (2018). The key to this representation is the Definition 1,´ which encompassed any infinitely divisible process. Special cases of the representation, under additional assumptions on the underlying process, were obtained in Barndorff-Nielsen et al. (2015) and Kabluchko, Z., &Stoev, S. (2016). Below and in what follows, $\hat{T}$ will denote the family of all finite nonempty subsets of the index set T,

$$\hat{T} = \{I \subset T : 0 < \mathrm{Card}(I) < \infty\},$$

so that for any $I \in \hat{T}$, $\mathbb{R}^I$ can be identified with the Euclidean space $\mathbb{R}^{Card(I)}$. $[\![.]\!]$ will denote a fixed truncation functions, see Rosinski, J. (2018) for details.´

Theorem 2. Let $X = (X_t)_{t\in T}$ be an infinitely divisible process. Then there exist a unique triplet $(\sum, \nu, b)$ consisting of a non-negative definite function $\sum$ on $T \times T$, a Lévy measure ν on $(\mathbb{R}^T, \mathfrak{B}^T)$ and a function $b \in \mathbb{R}^T$ such that for every $I \in \hat{T}$ and $a \in \mathbb{R}^I$

$$\mathbb{E}\exp i\sum_{t\in I} a_t X_t = \exp\left\{-\frac{1}{2}\langle a, \Sigma_I a\rangle + \int_{\mathbb{R}^T} (e^{i\langle a, x_I\rangle} - 1 - i\langle a, [\![x_I]\!]\rangle)\, \nu(dx) + i\langle a, b_I\rangle\right\}, \quad (3)$$

where $\sum_I$ is the restriction of $\sum$ to $I \times I$. $(\sum, \nu, b)$ is called the generating triplet of X. Conversely, given a generating triplet $(\sum, \nu, b)$ as above, there exists an infinitely divisible process $X = (X_t)_{t\in T}$ satisfying (3).

By (3) one can decompose $X \overset{d}{=} G + Y$, where $G = (G_t)_{t\in T}$ is a symmetric Gaussian process with covariance function $\Sigma : T \times T \mapsto \mathbb{R}$ and $Y = (Y_t)_{t\in T}$ is an independent of G Poissonian process. By their independence, in many situations, we can consider both processes separately. The Poissonian process Y admits the following canonical spectral representation.

Proposition 3. Let $Y = (Y_t)_{t\in T}$ be a Poissonian infinitely divisible process with Lévy measure ν and a shift function b. Let N be a Poisson random measure on $(\mathbb{R}^T, \mathfrak{B}^T)$ whose intensity measure equals to the Lévy measure ν of Y. Let χ be a (cutoff) function on $\mathbb{R}$ satisfying $[\![u]\!] = u\chi(u), u \in \mathbb{R}$. Then the process $\tilde{Y} = (\tilde{Y}_t)_{t\in T}$ given by

$$\tilde{Y}_t = \int_{\mathbb{R}^T} y(t)[N(dy) - \chi(y(t))\nu(dy)] + b(t), \quad t \in T$$

has the same distribution as Y. $\tilde{Y}$ will be called a canonical spectral representation of Y.

Notice that this is a non-linear stochastic integral but is well defined due to condition (L1) of Definition 1 (see Rosinski, J. (2018) for details.).

Finally, we should mention transfer of regularity for Lévy measures. In short, this property says that path regularities of infinitely divisible processes are inherited by supports of their path Lévy measures. A precise statement follows (cf. Rosinski, J. (2018)).

Theorem 4 (Transfer of regularity). Let $Y = (Y_t)_{t\in T}$ be an infinitely divisible process with a σ-finite Lévy measure ν. Assume that paths of Y lie in a set U that is a standard Borel space for the σ-algebra $\mathcal{U} = \mathcal{B}^T \cap U$ and also that U is an algebraic subgroup of $\mathbb{R}^T$ under addition. Then ν is concentrated on U in the sense that $\nu * (\mathbb{R}^T \setminus U) = 0$. Therefore, both $\mathcal{L}(X)$ and its Lévy measure ν are carried by U.

This implies, for instance, that the Lévy measure of an infinitely divisible process on $T = [0,1]$ having twice continuously differentiable sample paths must be concentrated on $C^2[0,1]$.

3. Result

Now we are ready state the results.

Theorem 5. Let $X = (X_t)_{t\in T}$ be a Poissonian infinitely divisible process with a σ-finite Lévy measure ν and given by its canonical spectral representation

$$X_t = \int_{\mathbb{R}^T} x(t)[N(dx) - \chi(x(t))\nu(dx)] + b(t), \quad t \in T,$$

where N is a Poisson random measure with intensity ν. Let $Z = (Z_t)_{t\in T}$ be an arbitrary process independent of N such that $\mathcal{L}(Z) \ll \nu$. Put $q := \frac{d\mathcal{L}(Z)}{d\nu}$.

Then for any measurable functional $F: \mathbb{R}^T \mapsto \mathbb{R}$

$$\mathbb{E}F(\{X_t + Z_t : t \in T\}) = \mathbb{E}[F(\{X_t : t \in T\})N(q)]$$

where

$$N(q) = \int_{\mathbb{R}^T} q(x)N(dx).$$

Theorem 6. Let $X = (X_t)_{t\in T}$ be a Poissonian infinitely divisible process with a σ-finite Lévy measure ν and given by its canonical spectral representation

$$X_t = \int_{\mathbb{R}^T} x(t)[N(dx) - \chi(x(t))\nu(dx)] + b(t), \quad t \in T,$$

where N is a Poisson random measure with intensity ν. Let $Z^{(j)} = (Z_t^{(j)})_{t\in T}$ be independent processes that are also independent of N such that $\mathcal{L}(Z^{(j)}) \ll \nu$. Then for any measurable functional $F: \mathbb{R}^T \mapsto \mathbb{R}$

$$\mathbb{E}F\left(\{X_t + \sum_{j=1}^{m} Z_t^{(j)} : t \in T\}\right) = \mathbb{E}[F(\{X_t : t \in T\})H_m(N)] \tag{4}$$

where

$$H_m(N) = N(H_{m-1}(N-\delta_\cdot)q_m(\cdot)), \quad H_0 = 1. \tag{5}$$

We have

$$H_m(N) = \sum_{P\in\mathcal{P}_m} c_P \prod_{j=1}^{|P|} N(q^{P_j}) \tag{6}$$

where $\mathcal{P}_m$ is the family of all partitions $P = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, m\}$, and $|P|$ is the number of sets in partition P; $q^{P_j} = \prod_{i\in P_j} q_i$.

In particular, if the supports of q_i's are pairwise disjoint modulo ν, then

$$H_m(N) = \prod_{i=1}^{m} N(q_i).$$

If the supports of q_i's are triple-wise disjoint modulo ν, then

$$H_m(N) = \sum_{P\in\mathcal{P}_{m,2}} (-1)^{\kappa(P)} \prod_{j=1}^{|P|} N(q^{P_j}).$$

where $\mathcal{P}_{m,2}$ is the family of all partitions $P = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, m\}$, such that $|P_j| = 1$ or 2, $|P|$ is the number of sets in P, and $q^{P_j} = \prod_{i\in P_j} q_i$. $\kappa(P) = \mathrm{Card}\{j : |P_j| = 2\}$.

4. Discussion and Conclusion

We illustrate this Theorem 5 on a familiar case of Lévy processes.

Example 7. Let $X = (X_t)_t \geq 0$ be a Lévy process such that $\mathbb{E}e^{iuX_t} = e^{tK(u)}$, where K is a cumulant function given by

$$K(u) = \int_{\mathbb{R}} (e^{iux} - 1 - iux\chi(x))\,\rho(dx) + icu.$$

Let $q : \mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}_+$ be a measurable function such that $\int_{\mathbb{R}_+\times\mathbb{R}} q(r,v)dr\rho(dv) = 1$.Then for any measurable functional $F : \mathbb{R}^{(0,\infty)} \mapsto \mathbb{R}$

$$\mathbb{E}\int_{\mathbb{R}_+\times\mathbb{R}} F\left(\{X_t + \mathbf{1}_{\{r\leq t\}}v : t \geq 0\}\right) q(r,v)\,dr\rho(dv)$$
$$= \mathbb{E}[F\left(\{X_t : t \geq 0\}\right) g(X)],$$
$$\text{where } g(X) = \textstyle\sum_{\{r>0:\Delta X_r\neq 0\}} q(r,\Delta X_r) \text{ and } \Delta X_r = X_r - X_{r-}.$$

In this example $Z_t((r,v)) = \mathbf{1}_{\{r\leq t\}}v$, where $(r,v) \in \widetilde{\Omega} = \mathbb{R}_+ \times \mathbb{R}$ and $\widetilde{\Omega}$ is equipped with probability measure $\widetilde{\mathbb{P}}(dr,dv) = q(r,v)dr\rho(dv)$. The Lévy measure of X is the "distribution" of the process Z on $\mathbb{R}^{\mathbb{R}+}$ under possibly infinite infinite measure $dr\rho(dv)$ on $\widetilde{\Omega}$. Therefore, one has as many random translations of a Lévy process as the number of choices of a nonnegative function q satisfying $\int_{\mathbb{R}_+\times\mathbb{R}} q(r,v)dr\rho(dv) = 1$. This is already a rich class.

The possibility of selecting of many independent translations increases their applicability but that also increases the complexity of $H_m(N)$ in the change of

measure formula (4), see (5) and (6). Still, relatively easy to handle is case $m = 2$, where

$$H_2(N) = N(q_1)N(q_2) - N(q_1 q_2).$$

The infinite divisibility of squared Bessel processes was not known until Shiga, T., & Watanabe, S. (1973) paper. It still took several years until the Lévy measures of such processes were found by Pitman, J.W., & Yor, M. (1982). The description of such measures is in terms of Itô measure of the Brownian positive excursions and the total accumulated local time of such excursions. Therefore, it seems that Lévy measures here are relatively more complicated than the infinitely divisible processes. This is completely opposite to the case of Lévy processes, where Lévy measures are simple but processes can get involved.

Squared Bessel processes and Feller processes belong to the class of permanental processes. Permanental distributions were first considered in statistics as an extension of gamma distributions to the multivariate case. Vere-Jones, D. (1967) established the infinite divisibility of bivariate permanental distributions and found their Lévy measures. Significant progress to characterize the infinite divisibility of multivariate permanental distributions was made by Bapat, R.B. (1989). For a complete discussion of infinite divisibility of permanental distributions see Eisenbaum, N. and Kaspi, H. (2009) and references therein. The celebrated Dynkin Isomorphism Theorem (Dynkin, E.B. (1984)) can now be viewed in the framework of admissible translations (2), see Rosinski, J. (2018).

References

1. Bapat, R.B. (1989). Infinite divisibility of multivariate gamma distribution and Mmatrices, Sankhya 51 73-78.
2. Barndorff-Nielsen, O.E., Sauri, O., and Szozda, B. (2015) Selfdecomposable Fields. *J. Theor. Probab.*, Online First, Springer.
3. Dynkin, E.B. (1984). Gaussian and non-Gaussian random fields associated with Markov processes, *J. Funct. Anal.* 55 344–376.
4. Eisenbaum, N. (2003). On the infinite divisibility of squared Gaussian processes, *Probab. Theory Related Fields* 125 381–392.
5. Eisenbaum, N. (2008). A Cox Process Involved in the Bose–Einstein Condensation, *Ann. Henri Poincaré* 9 1123–1140.
6. Eisenbaum, N. and Kaspi, H. (2006). A characterization of the infinitely divisible squared Gaussian processes, *Ann. Probab.* 34(2) 728–742.
7. Eisenbaum, N. and Kaspi, H. (2009). On permanental processes, *Stochastic Process. Appl.* 119(5) 1401–1764.
8. Kabluchko, Z. and Stoev, S. (2016). Stochastic integral representations and classification of sum– and max-infinitely divisible processes, *Bernoulli* 22(1), 107–142.

9. Lee, P.M. (1967). Infinitely divisible stochastic processes. *Z. Wahrsch. verv. Geb.*, 7 147– 160.
10. Marcus, M.B. and Rosen, J. (1992). Sample path properties of the local times of strongly symmetric Markov processes via Gaussian processes, *Ann. Probab.* 720 1603–1684 (special invited paper).
11. Marcus, M.B. and Rosen, J. *Markov Processes, Gaussian Processes, and Local Times*. Cambridge Studies in Advanced Mathematics 100. Cambridge University Press, 2006.
12. Maruyama, G. (1970). Infinitely divisible processes. *Theory Probab. Appl.* 15, (1970), 1–22.
13. Pitman, J.W. and Yor, M. (1982). A decomposition of Bessel bridges, *Z. Wahrsch. verv. Geb.*, 59 425–457.
14. Rajput, B.S. and Rosinski, J. (1989). Spectral representations of infinitely divisible pro-´ cesses, *Probab. Theory Related Fields* 82(3) 451–487.
15. Rosinski, J. (1990). On series representations of infinitely divisible random vectors, *Ann. Probab.* 18 405–430.
16. Rosinski, J. (2000). Decomposition of stationary´ α-stable random fields. *Ann. Probab.* 28 1797–1813.
17. Rosinski, J. (2001). Series representations of Lévy processes from the perspective of point´ processes. In *Lévy processes*, pp. 401–415. Boston, MA: Birkhäuser Boston.
18. Rosinski, J. (2007). Lévy and Related Jump-type Infinitely Divisible Processes. Lecture´ Notes, Cornell University.
19. Rosinski, J. (2018). Representations and isomorphism identities for infinitely divisible´ processes, Ann. Probab. 46 (2018), no. 6, 3229-3274.
20. Sato, K. *Lévy Processes and Infinitely Divisible Distributions*, vol. 68 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1999.
21. Shiga, T. and Watanabe, S. (1973). Bessel diffusions as a one-parameter family of diffusion processes. *Z. Wahrsch. verv. Geb.* 27 37–46.
22. Vere-Jones, D. (1967). The infinite divisibility of a bivariate gamma distributions. *Sankhya˜* Ser. A 29 421–422.

A survey of recent progress in free infinite divisibility

Steen Thorbjørnsen
Depart. of Math., University of Aarhus, Ny Munkegade 118, 8000 Aarhus C, Denmark

Abstract

The aim of this paper is to provide a brief introduction to the theory of the additive convolution associated to the notion of free independence and the derived concept of free infinite divisibility. We shall emphasize many parallels and some differences to the classical theory of infinitely divisible probability measures. We aim further to provide an overview of some of the recent developments in this field with an ample amount of references.

1. Introduction

With the work of Speicher [19], Ben Ghorbal and Schu¨rmann [5] and Muraki [16] it was clarified that there are only five notions of "probabilistic independence" which satisfy some naturally required conditions (associativity, universality, extension and normalization). These five notions of independence are: Classical (or tensor) independence, Free independence, Boolean independence, Monotone independence and Anti-monotone independence. Two classical random variables X and Y cannot satisfy any of the four last notions of independence, unless either X or Y is a constant. In fact, disregarding constant random variables, the four last notions of independence will all entail that the product XY is distinct from YX. Thus these last four notions of independence are (in essence) only realizable in the framework of non-commutative (or quantum) probability, where the "random variables" x and y are realized as Hermitian (possibly unbounded) operators on an infinite dimensional Hilbert space H, and the expectation functional can be realized as a vector state corresponding to a (fixed) unit vector ξ on $\mathcal{H}$:

$$\mathbb{E}_\xi[x] = \langle x\xi, \xi\rangle,$$

where $\langle .,. \rangle$ denotes the inner product on $\mathcal{H}$. In case x is a continuous (i.e. bounded) Hermitian operator on $\mathcal{H}$, there exists a unique compactly supported probability measure μ_x on $\mathbb{R}$, satisfying that

$$\mathbb{E}_\xi[p(x)] = \int_{\mathbb{R}} p(t)\, \mu_x(dt)$$

for any polynomial $p\colon \mathbb{R} \to \mathbb{R}$ and with $p(x)$ defined in the obvious way. This measure μ_x is referred to as the (spectral) distribution of x (with respect to the

chosen vector state). If x is unbounded (i.e. non-continuous) the spectral distribution μ_x is similarly defined by the equation

$$\mathbb{E}_\xi[f(x)] = \int_{\mathbb{R}} p(t)\,\mu_x(dt)$$

which is then required to hold for any bounded Borel function $f : \mathbb{R} \to \mathbb{R}$, and where $f(x)$ is defined in terms of spectral calculus (see e.g. [17]). In this case μ_x generally has unbounded support.

Each of the five notions of independence mentioned above gives rise to a corresponding notion of (additive) convolution of two probability measures μ and ν on $\mathbb{R}$, roughly defined as the (spectral) distribution of $x + y$, where x and y are two Hermitian operators such that $\mu_x = \mu, \mu_y = \nu$, and x and y are independent in the considered sense (see Theorem 2.1 below).

Among the four non-classical notions of independence, free independence, introduced by Voiculescu in the 1980's, is by far the most developed and well-studied. This is mainly due to its applications in the theory of operator algebras (which was Voiculescu's original motivation for introducing the concept; see [22]) and its striking connection to the theory of random matrices, which we shall indicate below. In the Hilbert space framework outlined above, two Hermitian bounded operators x and y are *freely independent*, if, for any sequence $p_1, p_2, p_3, \ldots$ of polynomials in one variable and for any n in N, it holds that

$$\mathbb{E}_\xi\left[\left(p_1(x) - \mathbb{E}_\xi[p_1(x)]\right) \cdot \left(p_2(y) - \mathbb{E}_\xi[p_2(y)]\right) \cdots \left(p_{2n}(y) - \mathbb{E}_\xi[p_{2n}(y)]\right)\right] = 0,$$

and that

$$\mathbb{E}_\xi\left[\left(p_1(x) - \mathbb{E}_\xi[p_1(x)]\right) \cdot \left(p_2(y) - \mathbb{E}_\xi[p_2(y)]\right) \cdots \left(p_{2n+1}(x) - \mathbb{E}_\xi[p_{2n+1}(x)]\right)\right] = 0,$$

and similar conditions for products starting with polynomial expressions in *y*. In words the requirement is that any product of centered polynomial expressions, alternating in *x* and *y*, must have expectation equal to 0.

2. Free additive convolution

2.1 Theorem & Definition ([21],[7]). Let μ and ν be (Borel-) probability measures on $\mathbb{R}$. Then there exists a Hilbert space $\mathcal{H}$, a unit vector ξ in $\mathcal{H}$ and (possibly unbounded) Hermitian operators x and y acting on $\mathcal{H}$, such that the following conditions hold (with the notation introduced above)

(i) $\mu_x = \mu, \mu_y = \nu$.

(ii) x and y are freely independent with respect to E_ξ.

The conditions (i) and (ii) uniquely determine (in particular) the spectral distribution of the Hermitian operator $x + y$, which may thus be denoted by $\mu \boxplus \nu$ and referred to as the free (additive) convolution of μ and ν.

2.2 Connection to random matrices. As mentioned above one of the most important aspects of free independence is its significance for large random matrices. To briefly indicate this, recall that for a Hermitian $n \times n$ −matrix $\mathcal{H}$, the empirical spectral distribution μ_H is the probability measure on $\mathbb{R}$ given by: $\mu_H = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda k}$, where $\lambda_1 \leq \cdots \leq \lambda_n$ are the n real eigenvalues of $\mathcal{H}$ (counted with multiplicity), and δ_c denotes the Dirac measure at a real number c. Now, let μ and ν be probability measures on $\mathbb{R}$, and for each n in $\mathbb{N}$ assume that A_n and B_n are two Hermitian random matrices such that the entries of A_n are (jointly) independent of those of B_n, and such that $\mu_{An(\omega)} \longrightarrow \mu$ and $\mu_{Bn(\omega)} \longrightarrow \nu$ weakly as $n \to \infty$ for almost all ω. Then under various additional (but rather general) conditions on the distribution of the entries of A_n and B_n it holds that $\mu_{A_n}(\omega) + B_n(\omega) \longrightarrow \mu \boxplus \nu$ weakly as $n \to \infty$ for almost all ω (see e.g. [22] or [1]). This (meta-) result illustrates the phenomenon that as dimension increases the assumed (classical) independence between the random matrices A_n and B_n is transformed into free independence. Thus free probability provides a concrete model for the asymptotics of large (classically independent) random matrices, and the analytic function tools of free probability (described in the following) can be used to determine these asymptotics; a fact that has been exploited recently e.g. in the theory of wireless communication.

As we shall demonstrate, the theory of free additive convolution is in many respects completely parallel to the classical theory. For example we have the following analog of the classical CLT:

2.3 Free Central Limit Theorem ([20], [8], [23], [24]). Let μ be a probability measure on R with zero mean and finite variance σ^2. Then

$$D_{1/\sqrt{n\sigma^2}}(\mu^{\boxplus n}) \longrightarrow \frac{1}{2\pi}\sqrt{4-t^2}\,1_{[-2,2]}(t)\,\mathrm{d}t \qquad \text{weakly as } n \to \infty.$$

In fact, $D_{1\sqrt{n\sigma^2}}(\mu^{\boxplus n})$ is Lebesgue absolutely continuous for large n, and the densities converge uniformly to that of the semi-circle distribution.

In Theorem 2.3 we use the notation $D_c\mu$ for the scaling (or dilation) of a measure μ by the constant c. Theorem 2.3 illustrates the phenomenon that the role of the Gaussian distribution in classical probability is played by the semicircle distribution in free probability. In a similar fashion the role of the Poisson distribution in classical probability theory is in many respects played by the Marchenko-Pastur distribution in free probability.

2.4 Free Poisson Limit Theorem ([20]). For any positive number λ, it holds that

$$\left((1-\tfrac{\lambda}{n})\delta_0 + \tfrac{\lambda}{n}\delta_1\right)^{\boxplus n} \longrightarrow \nu \qquad \text{weakly as } n \to \infty,$$

where ν is the Marchenko-Pastur Law:

$$\nu(\mathrm{d}x) = \begin{cases} (1-\lambda)\delta_0 + \frac{1}{2\pi x}\sqrt{(x-a)(b-x)}\cdot 1_{[a,b]}(x)\ \mathrm{d}x, & \text{if } 0 \le \lambda \le 1, \\ \frac{1}{2\pi x}\sqrt{(x-a)(b-x)}\cdot 1_{[a,b]}(x)\,\mathrm{d}x, & \text{if } \lambda > 1, \end{cases}$$

with $a = (1-\sqrt{\lambda})^2$ and $b = (1+\sqrt{\lambda})^2$.

By a standard computation one may verify that the square of a semicircular distributed random variable has a Marchenko-Pastur distribution. This curious fact provides a link between the free analogs of the Gaussian and Poisson distributions which is not paralleled in the classical theory (but in accordance with random matrix theory). This is also the case for the following theorem due to Bercovici and Voiculescu, which illustrates some of the regularizing properties of free convolution.

2.5 Theorem ([9]). Let μ and ν be (Borel-) probability measures on R. Then for any atom γ for $\mu \boxplus \nu$, there exist atoms α and β for μ and ν, respectively, such that

(a) $\gamma = \alpha + \beta$,

(b) $\mu(\{\alpha\}) + \nu(\{\beta\}) > 1$,

(c) $\mu \boxplus \nu(\{\gamma\}) = \mu(\{\alpha\}) + \nu(\{\beta\}) - 1$.

As corollaries of this result it follows that a free convolution $\mu\nu$ can have at most finitely many atoms and that a free convolution square $\mu \boxplus \mu$ can have at most one atom!

2.6 Examples. (1) Denote by C the (standard) Cauchy distribution, i.e. the probability measure on $\mathbb{R}$ with Lebesgue density $t \mapsto \frac{1}{\pi}\frac{1}{1+t^2}$. Then for any (Borel-) probability measure μ on $\mathbb{R}$, it holds that $\mu \boxplus C = \mu * C$, where the "$*$" on the right hand side denotes classical convolution. A proof of this intriguing "folklore result" can be found e.g. in [13].
(2) In [8] Bercovici and Voiculescu established that there exist non-semicircular probability measures μ and ν, such that $\mu \boxplus \nu = \frac{1}{2\pi}\sqrt{4-t^2}1_{[-2,2]}(t)dt$. In classical probability Cramér's Theorem asserts that if the convolution of two probability measures μ and ν is a Gaussian distribution, then both μ and ν have to be Gaussian distributions themselves. Thus Cramér's Theorem fails in free probability.

3. Free Infinite divisibility

The classes of infinitely divisible, stable and self-decomposable probability laws in free probability are obtained by replacing classical convolution by free convolution in the definitions of the corresponding classical classes. By P(R) we denote in the following the class of all Borel probability measures on R.

3.1 Definition. (a) A measure μ from P(R) is $\boxplus$-infinitely divisible if the following condition is satisfied:

$$\forall_n \in \mathbb{N} \exists_{\mu_n} \in \mathcal{P}(\mathbb{R}): \mu = \mu_n \boxplus \mu_n \boxplus \ldots \boxplus \mu_n \quad (n \text{ terms}).$$

The class of $\boxplus$-infinitely divisible probability measures is denoted by $\mathcal{ID}(\boxplus)$.

(b) A measure μ from $\mathcal{P}(\mathbb{R})$ is -stable if the class of (increasing) affine transformations

$$\mathcal{A}(\mu) = \{D_c\mu \boxplus \delta_b \mid c > 0, b \in \mathbb{R}\}$$

is stable under $\boxplus$. The class of $\boxplus$-stable probability measures is denoted by $\mathcal{S}(\boxplus)$.

(c) A measure μ from P(R) is $\boxplus$-self-decomposable if the following condition is satisfied:

$$\forall c \in (0,1)\ \exists \mu_c \in \mathcal{P}(\mathbb{R}): \mu = D_c\mu \boxplus \mu_c.$$

The class of all $\boxplus$-self-decomposable probability measures is denoted by $\mathcal{L}(\boxplus)$.

In the following we denote the classical counterparts of $\mathcal{ID}(\boxplus)$, $\mathcal{L}(\boxplus)$ and $\mathcal{S}(\boxplus)$ by, respectively, $\mathcal{ID}(*)$, $\mathcal{L}(*)$ and $\mathcal{S}(*)$. As in the classical case, we have the following hierarchy of classes of probability measures (see [7] and [2]):

$$\mathcal{P}(\mathbb{R}) \supseteq \mathcal{ID}(\boxplus) \supseteq \mathcal{L}(\boxplus) \supseteq \mathcal{S}(\boxplus) \supseteq \mathcal{G}(\boxplus),$$

where $\mathcal{G}(\boxplus)$ denotes the class of semi-circular distributions.

As in classical probability the main tool for studying the class $\mathcal{ID}(\boxplus)$ is a Lévy -Khintchine type representation. In order to describe this in detail we first have to introduce the free analog of the logarithm of the Fourier transform of a probability measure. By $\mathbb{C}^+$ (resp. $\mathbb{C}^-$) we denote the set of complex numbers with strictly positive (resp. negative) imaginary part.

3.2 Theorem & Definition ([7]). Let μ be a probability measure on $\mathbb{R}$, and consider its Cauchy (or Stieltjes) transform:

$$G_\mu(z) = \int_{\mathbb{R}} \frac{1}{z-t}\,\mu(\mathrm{d}t), \qquad (z \in \mathbb{C}^+).$$

Then the range of G_μ contains an open region D_μ in the form:

$$\mathcal{D}_\mu = \left\{z \in \mathbb{C}^- \;\middle|\; |z| < \delta \text{ and } \mathrm{Arg}(z) \in (-\tfrac{\pi}{2} - \epsilon, -\tfrac{\pi}{2} + \epsilon)\right\}$$

for suitable ϵ, δ in $(0, \infty)$. On this region the (right) inverse G_μ^{-1} is well-defined, and the free cumulant transform $\mathcal{C}_\mu$ may subsequently be defined as

$$\mathcal{C}_\mu(z) = zG_\mu^{-1}(z) - 1, \qquad (z \in \mathcal{D}_\mu).$$

The key property of the free cumulant transform is that it linearizes free additive convolution:

3.3 Theorem ([21],[15],[7]). For any (Borel-) probability measures μ_1, μ_2 on $\mathbb{R}$ we have that

$$\mathcal{C}_{\mu_1 \boxplus \mu_2}(z) = \mathcal{C}_{\mu_1}(z) + \mathcal{C}_{\mu_2}(z),$$

for all z in a region of $\mathbb{C}^-$ where all three free cumulant transforms are defined.

We can now state a Lévy-Khintchine type representation for the free cumulant transform of a free infinitely divisible probability law:

3.4 Theorem ([7]). For a measure μ in P(R) the following conditions are equivalent:

(i) $\mu \in \mathcal{ID}(\boxplus)$.

(ii) $\mathcal{C}_\mu$ may be extended to an analytic function $\mathcal{C}_\mu : \mathbb{C}^- \to \mathbb{C}$.

(iii) There exist unique a in $[0, \infty)$, η in $\mathbb{R}$ and a Lévy measure ρ on $\mathbb{R}$, such that

$$\mathcal{C}_\mu(z) = \eta z + a z^2 + \int_{\mathbb{R}} \left(\frac{1}{1 - tz} - 1 - tz 1_{[-1,1]}(t) \right) \rho(\mathrm{d}t), \quad (z \in \mathbb{C}^-).$$

The triplet (a, ρ, η) appearing in (iii) of Theorem 3.4 is called the free characteristic triplet for μ. In terms of the characteristic triplet, we have the following characterizations of the classes of stable and self-decomposable probability laws in free probability:

3.5 Theorem ([7], [2]). For μ in $\mathcal{ID}(\boxplus)$ with free characteristic triplet (a, ρ, η) it holds that:

(i) $\mu \in \mathcal{S}(\boxplus)\Gamma(\boxplus)$, if and only if a = 0, and ρ has the form:

$$\rho(\mathrm{d}t) = \left(c_- |t|^{-1-\alpha} 1_{(-\infty,0)}(t) + c_+ t^{-1-\alpha} 1_{(0,\infty)}(t) \right) \mathrm{d}t$$

for suitable constants $c^+ c^-$ in $[0,\infty)$ and α in (0,2).

(ii) $\mu \in \mathcal{L}(\boxplus)$, if and only if ρ has the form: $\rho(\mathrm{d}t) = \frac{k(t)}{|t|}\mathrm{d}t$, where k is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$.

3.6 The Bercovici-Pata bijection. The theorem above demonstrates a complete analogy to the characterizations of the classical stable and selfdecomposable distributions in terms of the classical Lévy-Khintchine representation. This is no coincidence. In fact the mapping Λ that maps a measure in $\mathcal{ID}(*)$ with characteristic triplet (a, ρ, η) onto the measure in $\mathcal{ID}(\boxplus)$ with free characteristic triplet (a, ρ, η) is (obviously) a bijection, but it also preserves scaling of measures and satisfies that $\Lambda(\delta_c) = \delta_c$ (for all c in $\mathbb{R}$) and that $\Lambda(\mu_1 * \mu_2) = \Lambda(\mu_1) \boxplus \Lambda(\mu_2)$ for all μ_1, μ_2 in$\mathcal{ID}(*)$). These properties immediately imply that $\Lambda(\mathcal{S}(*)) = \mathcal{S}(\boxplus)$ and $\Lambda(\mathcal{L}(*)) = \mathcal{L}(\boxplus)$, and hence the characterizations in Theorem 3.5 follow readily from the corresponding classical results (see e.g. [18]). The mapping Λ was introduced in [6] and is

widely referred to as the Bercovici-Pata bijection. It was shown in [2] that Λ is also a homeomorphism with respect to weak convergence. This mapping further supports the roles of the semi-circular distribution and the Marchenko-Pastur distribution as the free analogs of the Gaussian and Poisson distributions, respectively. Indeed, Λ maps the Gaussian distributions onto the semi-circular ones and the Poisson distributions onto the Marchenko-Pastur distributions. It is also noteworthy that the Cauchy distribution (see Example 2.6) is a fixed point with respect to Λ.

The Lebesgue decomposition of measures in $\mathcal{ID}(\boxplus)$

From the perspective of the Lebesgue-decomposition there are fundamental differences between the classes $\mathcal{ID}(*)$ and $\mathcal{ID}(\boxplus)$. In [3] it was proved that if $\nu \in \mathcal{ID}(\boxplus)$, then ν has no continuous singular part. Furthermore, Theorem 2.5 implies that ν has at most one atom, so that $\mathcal{ID}(\boxplus)$ contains no nondegenerate discrete probability laws. We mention also that it was proved in [7] that for all sufficiently large n, the convolution power $\nu^{\boxplus n}$ of a (non-degenerate) measure ν from $\mathcal{ID}(\boxplus)$ has no atoms. This is in contrast to the fact that for a measure μ in $\mathcal{ID}(*)$ the convolution power μ^{*n} either has an atom for all n or is atom-less for all n (see e.g. [18]).

The complete picture of the Lebesgue-decomposition of a measure μ from $\mathcal{ID}(\boxplus)$ is given in the following theorem which was proved only recently by Hasebe and Sakuma (based in part on [14]).

3.7 Theorem ([11]). For a measure ν in $\mathcal{ID}(\boxplus)$ with free characteristic triplet (a,ρ,η) it holds that:

(i) If a > 0 or ρ(ℝ) ∈ (1,∞], then ν is absolutely continuous (with respect to Lebesgue measure) and has a continuous density.

(ii) If a = 0 and ρ(ℝ) = 1, then ν is absolutely continuous.

(iii) If a = 0 and ρ(ℝ) ∈ [0,1), then $c := \log_{\epsilon \downarrow 0} F_\mu^{-1}(i\epsilon)$ exists in ℝ, and ν({c}) = 1−ρ(ℝ). Here the function F_μ is the reciprocal Cauchy transform: $F_\mu = 1/G_\mu$.

Prominent probability laws in $\mathcal{ID}(\boxplus)$

In this final subsection, we list a number a prominent probability laws (from classical probability theory), which in recent years have been shown to belong to $\mathcal{ID}(\boxplus)$.

(a) In [4] it was proved that the classical Gaussian distribution belongs to $\mathcal{ID}(\boxplus)$. This (at the time rather surprising) result was obtained by a deep complex analysis argument establishing that the free cumulant transform of $N(0,1)$ can be extended analytically to all of $\mathbb{C}^-$ (cf. Theorem 3.4). In [12] it was subsequently established (based on [4]) that in fact $N(0,1) \in \mathcal{L}(\boxplus)$.

(b) For a positive number p we let γ_p denote the Gamma distribution given by

$$\gamma_p(\mathrm{d}x) = \tfrac{1}{\Gamma(p)} x^{p-1}\,\mathrm{e}^{-x}\,1_{[0,\infty)}(x)\,\mathrm{d}x.$$

It was proved in [10] that$\gamma_p \in \mathcal{ID}(\boxplus)$,if $p \in \left(0,\frac{1}{2}\right] \cup \left[\frac{3}{2},\infty\right)$,whereas $\gamma_p \notin \mathcal{ID}(\boxplus)$, if p belongs to the set

$$\mathcal{I} := \bigcup_{n\in\mathbb{N}} \left(\tfrac{2n-1}{2n}, \tfrac{2n}{2n+1}\right) \cup \left(\tfrac{2n+2}{2n+1}, \tfrac{2n+1}{2n}\right) \subseteq \left(\tfrac{1}{2}, \tfrac{3}{2}\right). \qquad (3.1)$$

(c) For any positive number p we denote by γ_p^{-1} the inverse Gamma-distribution given by

$$\gamma_p^{-1}(\mathrm{d}x) = \tfrac{1}{\Gamma(p)} x^{-p-1}\,\mathrm{e}^{-1/x}\,1_{(0,\infty)}(x)\,\mathrm{d}x.$$

It was proved in [10] that $\gamma_p^{-1} \in \mathcal{ID}(\boxplus)$ for all p in $(0,\infty)$.

(d) For positive numbers p, q we let $\beta_{p,q}$ denote the Beta-distribution given by

$$\beta_{p,q}(\mathrm{d}x) = \tfrac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} 1_{[0,1]}(x)\,\mathrm{d}x.$$

In [10] it was proved that, $\beta_{p,q} \in \mathcal{ID}(\boxplus)$, if $p, q \geq \frac{3}{2}$, or if $p + q \geq 2$ and either p or q belongs to $\left(0,\frac{1}{2}\right]$.

If $p, q \in (0,1]$, or if either p or q belongs to the set $\mathcal{I}$ given in (3.1), then $\beta_{p,q} \notin \mathcal{ID}(\boxplus)$.

(e) For p, q in $(0,\infty)$ we let $\beta'_{p,q}$ denote the Beta-distribution of second kind given by

$$\beta'_{p,q}(\mathrm{d}x) = \tfrac{1}{B(p,q)} x^{p-1}(1+x)^{-p-q} 1_{[0,\infty)}(x)\,\mathrm{d}x.$$

It was proved in [10] that$\beta'_{p,q} \in \mathcal{ID}(\boxplus)$, if $p \in \left(0,\frac{1}{2}\right] \cup \left[\frac{3}{2},\infty\right)$, whereas $\beta'_{p,q} \notin \mathcal{ID}(\boxplus)$, if $p \in \mathcal{I}$.

(f) For any number q in $(\frac{1}{2},\infty)$ we denote by $\boldsymbol{t}_q$ Student's $\boldsymbol{t}$-distribution given by

$$\mathbf{t}_q(\mathrm{d}x) = \tfrac{1}{B(\frac{1}{2}, q-\frac{1}{2})} (1+x^2)^{-q}\,\mathrm{d}x.$$

It was proved in [10] that $\boldsymbol{t}_q \in \mathcal{ID}\,(\boxplus)$, if $q \in (\frac{1}{2}, 2]$ or if $q \in \bigcup_{n\in\mathbb{N}}[2n+\frac{1}{4}, 2n+2]$.

Some of the results from [10] listed above were obtained previously in special cases; we refer to [10] for a full bibliographical account on such partial results.

References

1. G. Andersen, A. Guionnet and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge studies in advanced Math. **118**, Cambridge University Press (2010).
2. O.E. Barndorff-Nielsen and S. Thorbjørnsen, *Selfdecomposability and Lévy processes in free probability*, Bernoulli **8** (2002), 323-366.
3. S.T. Belinschi and H. Bercovici, *Atoms and regularity for measures in a partially defined free convolution semigroup*, Math. Z. **248** (2004), 665-674.
4. S.T. Belinschi, M. *Bożejko*, F. Lehner and R. Speicher, *The normal distribution is $\boxplus$-infinitely divisible*, *Adv. Math.* **226** (2011), 3677–3698.
5. A. Ben Ghorbal and M. Schürman, *Non-commutative notions of stochastic independence*, Math. Proc. Cambridge. Philo. Soc. **133** (2002), 531-561.
6. H. Bercovici and V. Pata, *Stable Laws and Domains of Attraction in Free Probability Theory*, Ann. Math. **149** (1999), 1023-1060.
7. H. Bercovici and D.V. Voiculescu, *Free Convolution of Measures with Unbounded Support*, Indiana Univ. Math. J. **42** (1993), 733-773.
8. H. Bercovici and D.V. Voiculescu, *Super convergence to the central limit and failure of the Cramér's theorem for free random variables*, Probab. Theory Related Fields **103** (1995), 215-222.
9. H. Bercovici and D.V. Voiculescu, *Regularity questions for free convolution*, in "Nonself adjoint operator algebras, operator theory, and related topics" (H. Bercovici and C. Foias editors), Oper. Theory Adv. Appl. **104**, Birkhäuser (1998), 37-47.
10. T. Hasebe, *Free infinite divisibility for beta distributions and related ones*, Electr. J. Prob. **19** (2014), 1-33.
11. T. Hasebe and N. Sakuma, *Unimodality for free Lévy processes*, Ann. l'Institut Henri Poincaré **53** (2017), 916-936.
12. T. Hasebe, N. Sakuma and S. Thorbjørnsen, *The normal distribution is freely self decomposable*, Intern. Math. Research Notices, Volume 2019, Issue 6 (2019), 1758-1787.
13. T. Hasebe and Y. Ueda, *Large time unimodality for classical and free Brownian motions with initial distributions*, ALEA Lat. Am. J. Probab. Math. Stat. **15** (2018), 353-374.
14. H.-W. Huang, *Supports, regularity and -infinite divisibility for measures of the form* $\left(\mu^{\boxplus p}\right)^{\uplus q}$, arXiv:1209.5787v1.
15. H. Maassen, *Addition of freely independent random variables*, J. Funct. Anal. **106** (1992), 409-438.
16. N. Muraki, *The five independences as natural products*, Inf. Dim. Anal. Quant. Prob. **6** (2003), 337-371.
17. W. Rudin, *Functional Analysis (second edition)*, McGraw-Hill Inc. (1991).

18. K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge studies in advanced math. **68** (1999).
19. R. Speicher, *On universal products*, Fields Institute Communications **12**, D. Voiculescu (ed), 257-266 (1997).
20. D.V. Voiculescu, *Symmetries of some reduced free product C*-algebras*, in "Operator algebras and their Connections with Topology and Ergodic Theory", Springer Lecture Notes in Mathematics **1132** (1985), 556-588.
21. D.V. Voiculescu, *Addition of certain non-commuting random variables*, J. Funct. Anal. **66** (1986), 323-346.
22. D.V. Voiculescu, K.J. Dykema and A. Nica, *Free Random Variables*, CRM Monograph Series **1**, AMS (1992).
23. J. Wang, *Local limit theorems in free probability theory*, Ann. Probab. **38** (2010), 1492-1506.
24. J. Williams, *Uniform Convergence and the Free Central Limit Theorem*, Complex Anal. Oper. Theory. **6** (2012), 23-31.

Infinitely divisible probability measures under generalized convolutions

B.H. Jasiulis - Gołdyn[1]; M. Arendarczyk[1]; M. Borowiecka-Olszewska[2]; J.K. Misiewicz[3]; E. Omey[4]; J. Rosiński[5]

[1] Institute of Mathematics, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland

[2] Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra, ul. Prof. Z. Szafrana 4A, 65-516 Zielona Góra, Poland

[3] Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warszawa, Poland

[4] KU Leuven, Warmoesberg 26, 1000 Brussels, Belgium

[5] Department of Mathematics, 227 Ayres Hall, University of Tennessee, Knoxville TN 37996, USA

Abstract

Kingman, in his seminal work [13], introduced a new type of convolution of distributions that is naturally related to spherically symmetric random walks. Motivated by this paper, Urbanik in a series of papers [17] established a theory of generalized convolutions $\diamond$ as certain binary commutative and associative operations that include classical and Kingman's convolutions as a special case. This theory was further developed by Bingham ([2, 3]) in the context of regularly varying functions. There is a rich class of examples of generalized convolutions that are motivated by problems in applications of probability theory. For instance, the distribution of the maximum of two independent random variables is a generalized convolution fundamentally associated with the extreme value theory, and extensively applied to model events that rarely occur, but the appearance of which causes large losses. Similarly, to the classical theory, we define infinite divisibility with respect to generalized convolution $\diamond$ and establish Lévy-Khintchine representation [11]. Lévy and additive stochastic processes under generalized convolutions are constructed as the Markov processes in ([5]). In this paper we survey examples of generalized convolutions and related Lévy-Khintchine representation. Results on Kendall convolution and extreme Markov chains driven by the Kendall convolution ([1, 5, 10]) using Williamson transform ([18]) are also presented.

Keywords

infinitely divisible probability measure; generalized convolution; Kendall random walk; LévyKchintchine representation; regular variation

1. Introduction

Notation:

Throughout this paper, the family of all probability measures on the Borel subsets of $\mathbf{R}_+$ is denoted by $\mathbf{P}_+$. For a probability measure $\lambda \in \mathbf{P}_+$ and a $\in R_+$ the rescaling operator is given by $T_a = \mathcal{L}(aX)$ if $\lambda = \mathcal{L}(X)$ denotes the distribution of the random element X.

Finally a measurable function f(·) is regularly varying at infinity with index β (notation $f \in RV_\beta$) if, for all $x > 0$, it satisfies $lim_{t \to \infty} f(tx)/f(t) = x^\beta$ (see, e.g., [4]).

2. Methodology

The main unconventional tool used here is generalized convolution ([17]), which is a generalization of the classical convolution corresponding to the sum of independent random elements. Generalized convolutions were explored with the use of regular variation ([2,3]) and were applied to construct Lévy processes and stochastic integrals ([5]). Their origin can be found in delphic semigroups ([12]). The development of generalized convolutions was motivated by spherically symmetric random walks (see [13]). Hence generalized convolutions are closely related to multidimensional distributions.

Definition 1.

A **generalized convolution** is a binary, symmetric, associative and commutative operation on $\diamond$ $\mathcal{P}_+$ having the following properties:

(i) $\lambda \diamond \delta_0 = \lambda \; for \; all \; \lambda \in \mathcal{P}_+$;

(ii) $(p\lambda_1 + (1-p)\lambda_2 \diamond \lambda = p(\lambda_1 \diamond \lambda) + (1-p)(\lambda_2 \diamond \lambda)$ for each $p \epsilon [0,1]$ and $\lambda, \lambda_1, \lambda_2 \epsilon \mathcal{P}_+$;

(iii) $T_a(\lambda_1 \diamond \lambda_2) = (T_a\lambda_1) \diamond (T_a\lambda_2)$ for all $a \geq 0$ and $\lambda_1, \lambda_2 \epsilon \mathcal{P}_+$;

(iv) if $\lambda_n \longrightarrow \lambda$ and $v_n \longrightarrow v$, then $(\lambda_n \diamond v_n) \longrightarrow (\lambda \diamond v)$, where $\longrightarrow$ denotes weak convergence;

(v) there exists a sequence of positive numbers c_n such that $T_{cn}\, \delta_1^{\diamond n}$ converges weakly to a measure $v \neq \delta_0$ (here $\lambda^{\diamond n} = \lambda \diamond \lambda \diamond \ldots \diamond \lambda$ denotes the generalized convolution of n identical measures λ).

The pair $(\mathcal{P}_+, \diamond)$ is called a generalized convolution algebra. We define a continuous mapping h: $\mathcal{P}_+ \rightarrow \mathbf{R}_+$, called the homomorphism of the algebra $(\mathcal{P}_+, \diamond)$, such that for $\lambda, v \in \mathcal{P}_+$, $p \in [0,1]$, we have:

- $h(p\lambda + (1-p)v) = p\,h(\lambda) + (1-p)\,h(v)$,
- $h(\lambda \diamond v) = h(\lambda)\,h(v)$.

The homomorphism in $(\mathcal{P}_+, \diamond)$ plays an important role in the theory of generalised convolutions and if it is not trivial, then it defines, for any measure

$\lambda \in \mathcal{P}_+$ a counterpart of a classical characteristic function called generalized characteristic function

$$\Phi_\lambda(t) = h(T_t\,\lambda) = \int_{[0,\infty)} h(\delta_{tx})\,\lambda(dx).$$

Each generalized convolution is uniquely determined by the probability kernel $\delta_x \diamond \delta_y$, i.e.

$$\lambda_1 \diamond \lambda_2(A) = \int_{[0,\infty)} \int_{[0,\infty)} \delta_x \diamond \delta_y(A)\,\lambda_1(dx)\,\lambda_2(dy)$$

for every $\lambda_1, \lambda_2 \in \mathcal{P}_+$.

Example 1. The α-convolution, $\alpha > 0$, is defined, for $a, b, c \geq 0$, by $\delta_a \diamond \delta_b = \delta_c$, where $c^\alpha = a^\alpha + b^\alpha$ and with homomorphism $h(\delta_x) = exp\{-x^\alpha\}$.

Example 2. The Kendall convolution Δ_α is defined in the following way:

$$\delta_x \Delta_\alpha \delta_1 = x^\alpha\,\pi_{2\alpha} + (1 - x^\alpha)\,\delta_1$$

for $0 \leq x \leq 1$ and $\alpha > 0$, where $\pi_{2\alpha}$ denotes a Pareto distribution measure with the density $\pi_{2\alpha}(dx) = 2\alpha\, x^{-(2\alpha+1)}\,\mathbf{1}_{(1,\infty)}(x)dx$. In this case we have

$$h(\delta_t) = (1 - t^\alpha)_+,$$

where $a_+ = a$ if $a > 0$ and $a+ = 0$ if $a \leq 0$. The corresponding generalized characteristic function is the Williamson transform (for more details on the transform see, e.g., [14, 15, 16, 18])

$$\Phi_\lambda(t) = h(T_t\,\lambda) = \int_{[0,\infty)} (1 - x^\alpha\,t^\alpha)_+\,\lambda(dx).$$

Example 3. For every $p \geq 2$ and properly chosen $c > 0$ the function

$$h(\delta_t) = (1 - (c+1)t + ct^p)\mathbf{1}_{[0,1]}(t)$$

is the kernel of a Kendall type (see [14]) generalized convolution $\diamond$ defined for $x \in [0,1]$ by the formula:

$$\delta_x \diamond \delta_1 = h(\delta_x)\,\delta_1 + x^p\,\lambda_1 + (c+1)(x - x^p)\,\lambda_2,$$

where λ_1, λ_2 are probability measures absolutely continuous with respect to the Lebesgue measure and that does not depend on x. For example if $c = (p-1)^{-1}$ then

$$\lambda_1(du) = 2cu^{-3}\left[(c+1)(p+1)u^{1-p} + (c+1)(p-2) + cp(2p-1)u^{-2p-2}\right]\mathbf{1}_{[1,\infty)}(u)\,du,$$

and

$$\lambda_2(du) = c[2(p-2) + (p+1)u^{-p+1}]u^{-3}\,\mathbf{1}_{[1,\infty)}(u)\,du.$$

It is natural to consider infinitely divisible measures with respect to generalized convolutions.

Definition 2.
A measure $\lambda \in \mathcal{P}_+$ is said to be **infinitely divisible** with respect to the generalized convolution $\diamond$ (-$\diamond$ infinitely decomposable) in the algebra $(\mathcal{P}_+, \diamond)$ if for every $n \diamond \in \boldsymbol{N}$ there exists a probability measure $\lambda_n \in \mathcal{P}_+$ such that

$$\lambda = \lambda_n \diamond n.$$

One of the most important examples of $\diamond$-infinitely divisible distribution is $\diamond$-compound Poisson measure $Exp \diamond (a\lambda)$ defined in [5, 9, 10, 11]. In next example the Poisson probability measure in the Kendall generalized convolution algebra is presented.

Example 4.
$$\mathrm{Exp}_{\Delta\alpha}(a\delta_1)(du) = e^{-a}\,\delta_0(du) + ae^{-a}\,\delta_1(du) + \alpha\, a^2\, u^{-2\alpha-1} \exp\{-au^{-\alpha}\}\, u^{-2\alpha-1}\mathbf{1}_{[0,\infty)}(u)\, du.$$

It is worth to notice that Poisson measures with respect to generalized convolutions are not strictly discrete and have usually a continuous part. In [1] we proved that the last measure at the above convex linear combination is stable in the Kendall convolution algebra in the sense of Definition 3.

Definition 3.
Let $\lambda \in \mathcal{P}_+$. We say that λ is **stable** in the generalized convolution algebra $(\mathcal{P}_+, \diamond)$, if for all a, b ≥ 0 there exists c ≥ 0 such that

$$T_a\,\lambda \diamond T_b\,\lambda = T_c\,\lambda.$$

Similarly to the classical theory stable distributions in the generalized convolutions sense are $\diamond$ - infinitely divisible. The generalized characteristic function of $\diamond$ - infinitely divisible distribution is exponent of t^p for some $p > 0$. For every generalized convolution $\diamond$ on $\mathcal{P}_+$ there exists a constant $\kappa(\diamond)$, called a **characteristic exponent**, such that for every $p \in (0, \kappa(\diamond)]$ there exists a measure $\sigma_p \in \mathcal{P}_+$ with the $\diamond$-generalized characteristic function $\Phi\sigma_p(t) = \exp\{-t^p\}$ if $p < \infty$ and $\Phi\sigma_p(t) = \mathbf{1}_{[0,1]}(t)$ otherwise. For example $\kappa(\diamond) = 2$ for classical convolution and $\kappa(\Delta_\alpha) = \alpha$ for the case of Kendall convolution. Moreover, the set of all $\diamond$ - stable measures coincides with the set $\{T_a\,\sigma_p\colon a > 0, p \in (0, \kappa(\diamond)]\}$.

3. Results

Let λ be an infinitely divisible measure with respect to generalized convolution ◊. In [16] Urbanik found an analogue to the Lévy-Khintchine formula for the generalized characteristic function

$$\Phi_\lambda(t) = \exp\{-At^{\kappa(\diamond)} + \int_{[0,\infty)} (h(\delta_{tx}) - 1)/\omega(x)\, m(dx)\},$$

where m is a finite Borel measure on $[0,\infty)$, $\omega(x) = 1 - h(\delta_{min\{x,y\}})$ and $y > 0$ is such that $h(\delta_x) < 1$ whenever $0 < x \leq y$.

The extension of this result for the case of generalized convolutions on R connected with weakly stable measures (see, e.g., [11]) one can found in [7]. In [8] some connections with non-commutative probability theory are studied. Moreover some examples of measure m being an analog of the Lévy measures one can found in [5,11].

Using the Kołmogorov theorem we prove the existence of Lévy processes with respect to generalized convolution (see [5]) and show that they are Markov processes with the transition probabilities given by distributions that are infinitely divisible with respect to generalized convolution.

Theorem 1. Let $0 < s < t < u, x > 0$. There exists a Markov process $\{X_t : t > 0\}$ with $\mathcal{L}(X_1) = \lambda \in \mathcal{P}_+$ and transition probability:

$$P_{s,t}(x, A) := (\delta_x \diamond \lambda^{\diamond(t-s)})(A), \quad A \in Bor(R_+).$$

Proof. We show that the probability kernels $P_{s,t}(x, A)$ satisfy the Chapman-Kołmogorov equations, i.e.

$$P_{s,u}(x, A) = \int_{[0,\infty)} P_{s,t}(x, dy)\, P_{t,u}(y, A).$$

Indeed, we have:

$$\begin{aligned} P_{s,u}(x, A) &= (\delta_x \diamond \lambda^{\diamond(u-s)})(A) = (\delta_x \diamond \lambda^{\diamond(t-s)} \diamond \lambda^{\diamond(u-t)})(A) \\ &= \int_{[0,\infty)}\int_{[0,\infty)} (\delta_y \diamond \delta_z)(A)\ (\delta_x \diamond \lambda^{\diamond(t-s)})(dy)\ \lambda^{\diamond(u-t)}(dz) \\ &= \int_{[0,\infty)}\int_{[0,\infty)}\int_{[0,\infty)} (\delta_x \diamond \lambda^{\diamond(t-s)})(dy)\ (\delta_w \diamond \delta_z)(A)\ \delta_y(dw)\ \lambda^{\diamond(u-t)}(dz) \\ &= \int_{[0,\infty)} (\delta_x \diamond \lambda^{\diamond(t-s)})(dy)\ (\delta_y \diamond \lambda^{\diamond(u-t)})(A) = \int_{[0,\infty)} P_{s,t}(x, dy)\, P_{t,u}(y, A), \end{aligned}$$

which ends the proof.□

All these results are applied to the Kendall convolution case. We consider infinitely divisible distributions with respect to the Kendall convolution since except the classical and stable case this seems to be most applicable for modeling real processes.

In particular in [1] and [10] we prove a limit theorem for Markov chains $\{X_n : n \in N\}$ driven by the Kendall convolution (called also Kendall random walks

introduced in [6]) with $\mathcal{L}(X_1) = \lambda \in \mathcal{P}_+$ assuming that $E\,X_1^\alpha = \mathrm{m}_\alpha$ is finite or the truncated α – moment $\mathrm{H(t)} := \int_{[0,t)} x^\alpha\ \lambda\,(dx)$ is regularly varying.

Theorem 2. Let $\{X_n : n \in N\}$ be a Kendall random walk with parameter $\alpha > 0$, unit step distribution $\mathcal{L}(X_1) = \lambda \in \mathcal{P}_+$ and $\mathrm{G(x)} := \int_{[0,\infty)}(1 - x^{-\alpha}\,t^\alpha) + \lambda(dt)$

(i) If $E[X_1^\alpha] = m_\alpha < \infty$, then as $n \to \infty$,

$$\mathrm{n}^{-1/\alpha}\,\mathrm{X_n} \to \mathrm{X},$$

where the cdf of random variable X is given by

$$\mathrm{F_0(x) = (1 + m_\alpha\, x^{-\alpha})\, exp\{-m_\alpha\, x^{-\alpha}\}\ \ for\ \ x > 0,\ \ and\ \ F_0(0)=0.}$$

(ii) Suppose that $\mathrm{H} \in \mathrm{RV_\theta}$ where $0 \le \theta < \alpha$. Then there exists an increasing function U(x) such that $\mathrm{U(1/(1-G(x)) \sim x}$ and

$$\mathrm{X_n/U(n)\ \to Z^{-1/(\alpha-\theta)},\ n \to \infty,}$$

where Z has distribution, which is a convex linear combination of an exponential and a gamma distribution

$$\mathcal{L}(\mathrm{Z}) = \alpha^{-1}\,\theta\,\Gamma(1,1) + (1-\alpha^{-1}\,\theta)\,\Gamma(2,1)\,,$$

where $\Gamma(a,b)$ denotes the measure with the density $\mathrm{b^a}/\ \Gamma(\mathrm{a})\ x^{\,a-1}\mathrm{exp\{-bx\}1_{[0,\infty)}(u)}$.

Proof.

(i) Let F denotes the cdf of the unit step X_1. First notice that $\mathrm{H}(n^{\,1/\alpha}/\mathrm{x}) \to \mathrm{m}_\alpha$ and $\mathrm{F}(n^{\,1/\alpha}/\mathrm{x}) \to 1$, as $n \to \infty$. Since the Williamson transform for $n^{\,-1/\alpha}X_n$ is given by the following formula (see [1]):

$$\mathrm{G}(\mathrm{n}^{1/\alpha}/\mathrm{x})^{\mathrm{n}} = (\ \mathrm{F}(\mathrm{n}^{1/\alpha}/\mathrm{x}) - \mathrm{x}^\alpha\ \mathrm{H}(\mathrm{n}^{1/\alpha}/\mathrm{x})\ /\ \mathrm{n}\ )^{\mathrm{n}},$$

then we obtain $\mathrm{G}(n^{\,1/\alpha}/\mathrm{x})^{\mathrm{n}} \to \mathrm{exp\{-m_\alpha\ x^\alpha\}}$ as $n \to \infty$. To complete the proof it suffices to check that the limiting measure has exactly the Williamson transform $\mathrm{exp\{-}m_\alpha\ \mathrm{x}^\alpha\}$.

(ii) The second part of theorem can be proved using limit theorem for renewal process $\mathrm{N(t)} := \inf\{\,\mathrm{n}: X_{\,n+1} > \mathrm{t}\}$ constructed by the Kendall convolution (see Theorem 6 in [10]). We use the result that

$$\mathrm{(1\text{-}G(t))N(t) \to Z,\ t \to \infty.}$$

Since $\alpha - \theta > 0, 1/(1-G(x))$ is asymptotically equal to a strictly increasing function $\mathrm{V(x)} \in \mathrm{RV}_{\alpha-\theta}$ (see [4], Section 1.5.2, Theorem 1.5.4, p.23) and

$N(t)/V(t) \to Z$ in the sense of distribution. We denote the inverse of $V(x)$ by $U(x)$ and then $U \in RV_{1/(\alpha-\theta)}$. Clearly $N(t)/V(t) \to Z$ implies that $N(U(t))/t \to Z$.

Now we have

$$P(X_n \leq U(n)t) = P(N(U(n)t) > n) = P(N(U(n)t)/V(U(n)t) > n/V(U(n)t)).$$

Since $V(U(n)t)/n \sim V(U(n))t^{\alpha-\theta}/n \to t^{\alpha-\theta}$, we have $P(X_n \leq U(n)t) \to P(Z \geq t^{\alpha-\theta}) = P(Z^{-1/(\alpha-\theta)} \leq t)$.

Finally we conclude that $X_n/U(n) \to Z^{-1/(\alpha-\theta)}$ in the sense of distribution. The construction shows that $U(Q(x)) \sim U(V(x)) \sim x$, which ends the proof. □

It is worth to notice that exponential transform of random variable X fit to maximal daily concentration of nitrogen dioxide for data from USA and Poland in some cases. Above stable distributions are infinitely divisible in the Kendall generalized convolution algebra.

Since random walks with respect to the generalized convolutions form a class of extremal Markov chains (see [1, 5, 10]), studying them in the appropriate algebras will be a meaningful contribution to extreme value theory.

More about regular variation context for extremal Markov chains driven by the Kendall convolution one can find in [1, 10].

4. Discussion and Conclusion:

Even though the family of generalized convolutions is pretty rich by now we are still interested in constructing new examples and finding new methods of constructing them on the base of these which we already know.

The next open problem lie on proposing a statistical methods to recognize which stochastic processes are the Lévy processes with respect to some generalized convolution. Could we recognize this generalized convolution on the base of some empirical data?

Acknowledgements. This paper is a part of project "First order Kendall maximal autoregressive processes and their applications", which is carried out within the POWROTY/REINTEGRATION programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.

References

1. Arendarczyk M., Jasiulis-Gołdyn B.H. & Omey E.A.M. (2019). Asymptotic properties of Kendall random walks, submitted, arXiv: https://arxiv.org/pdf/1901.05698.pdf

2. Bingham N. H. (1971). Factorization theory and domains of attraction for generalized convolution algebra. Proc. London Math. Sci., Infinite Dimensional Analysis. *Quantum Probability and Related Topics* 23(4), 16-30.
3. Bingham N. H. (1984). On a theorem of Kłosowska about generalized convolutions. *Coll. Math.* 48(1), 117-125.
4. Bingham N. H., Goldie C. M. & Teugels J. L. (1987). *Regular variation.* Cambridge University Press, Cambridge.
5. Borowiecka-Olszewska M., Jasiulis-Gołdyn B.H., Misiewicz J.K. & Rosiński J. (2015). Lévy processes and stochastic integral in the sense of generalized convolution. *Bernoulli* 21(4), 2513-2551.
6. Jasiulis-Gołdyn B.H. (2016) Kendall random walks. *Probab. Math. Stat.* 36(1), 165-185.
7. Jasiulis B.H. (2010). Limit property for regular and weak generalized convolution. *J. Theoret. Probab*. 23(1), 315-327,
8. Jasiulis-Gołdyn B.H., Kula A. (2012). The Urbanik generalized convolutions in the noncommutative probability and a forgotten method of constructing generalized convolution. *Proc. Math. Sci.* 122(3), 437-458.
9. Jasiulis-Gołdyn B.H., Misiewicz J.K. (2015). Classical definitions of the Poisson process do not coincide in the case of weak generalized convolution. *Lith. Math. J.* 55(4), 518-542.
10. Jasiulis-Gołdyn B.H., Misiewicz J.K., Naskręt K. & Omey E.A.M. (2018). Renewal theory for extremal Markov sequences of the Kendall type, submitted, arXiv:https://arxiv.org/pdf/1803.11090.pdf
11. Jasiulis-Gołdyn B.H., Misiewicz J.K. (2015). Weak Lévy-Khintchine representation for weak infinite divisibility. *Theor. Probab. Appl.* 60(1), 45-61.
12. Kendall D. G. (1968). Delphic semi-groups, infinitely divisible regenerative phenomena, and the arithmetic of p-functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 9(3), 163-195.
13. Kingman J. G. C. (1963). Random Walks with Spherical Symmetry. *Acta Math.* 109(1), 11-53.
14. Misiewicz J. K. (2018). Generalized convolutions and Levi-Civita functional equation. *Aequationes Math.* 92(5), 911-933.
15. A.J. McNeil, J. Nešlehová. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *The Annals of Statistics* 37 (5B), 3059-3097.
16. McNeil A.J., Nešlehová J. (2010). From Archimedean to Liouville Copulas. *J. Multivariate Analysis* 101(8), 1771-1790.

17. Urbanik K., Generalized convolutions I-V, *Studia Math.*, 23(1964), 217-245, 45(1973), 57-70, 80(1984), 167-189, 83(1986), 57-95, 91(1988), 153-178.
18. Williamson R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Math. J.*, 23, 189-207.

Research on method of population prediction by big data from mobile phones

Xin Zheng[1], Qing Shen[1], Mingcui Du[*1], Guangzhi Zhang[2], Changcheng Kan[3]
[1]Beijing Municipal Bureau of Statistics, Beijing, China
[2]RongxinZhilian Network Technology Co., Ltd., Beijing, China
[3]Baidu Times Network Technology Co., Ltd. Beijing, China

Abstract
The signaling and APP data from mobile phones, as the most representative type of data among the spatial big data, enable the population analysis based on the personal behavior of mobile phone. By combining the traditional statistical method with the AI deep learning technology, this paper uses the time series correlation, spatial sequence correlation and deep residual model to predict the population size, and good results have been achieved. As the change in population size is periodic in time, the short-term prediction effect is good, but the effect brought by periodicity should be eliminated for the long-term prediction.

Keywords
Time series correlation; Spatial sequence correlation; Deep residual model

1. Introduction

In recent years, with the rapid development of mobile internet technology, mobile phones have become indispensable items in people's daily life. The signalling data and APP usage records from mobile phones, as the most representative type of data among the spatial big data, record the massive and diversified crowd time and space location information at short intervals, enable the analysis of urban spatial characteristics based on the personal behaviour of mobile phone and are of great significance for urban planning, transportation, public resource allocation, and business information mining, etc. By combining the traditional statistical method with the AI deep learning technology, this paper uses the time series correlation, spatial sequence correlation and deep residual model to predict the population size.

2. Methodology

(1) Using the time series correlation

If the time during a day is divided into several time periods $t_1, t_2, \cdots, t_T$ by 10min, 30min or 60min, the time series correlation model can be used to predict the population size in this area. If it is assumed that the population size at the next time period is related to that at the previous p time periods,

the ARIMA model is used to predict the population size at the next time period with the historical population size at p time periods.

(2) Using the spatial sequence correlation

Generally, the spatial correlation is given in space according to the distance. If the Euclidean distance is used as the measure of distance, the distance between any two places can be defined as follows:

$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

By taking $1 / d(x_i - x_j)$ as a weight, the population size of the central grid can be estimated by the weighted mean value of population size of nearby grids.

(3) Using the AI deep learning method

The deep residual ResNet method is a classical AI deep learning model. Its basic principle is that the increase in the number of layers will improve the learning effect of network in the neural network model, but at the same time this may result in the increase of errors. In the deep residual ResNet method, the residuals are learned instead of the mapping relation, so this problem is well improved.

3. Result

(1) By taking Tiantongyuan area in Beijing City as an example, the population stock recorded at 14:00 on the first Monday of each month is used as the sample data, and then the related structure in time is verified by time series. First, the data series will be subject to the white noise test. If the p value is smaller than 0.05, it will be judged as a non-white noise series. Second, the stationarity test is carried out to observe the changes in autocorrelation coefficient and partial autocorrelation coefficient. The autocorrelation coefficient does not attenuate rapidly, so we have reasons to believe that this time series is a non-stationary series. To further verify our speculation, the unit root test is introduced to judge the stationarity of the series. The p value of 0.9585 is larger than the significance level of 0.05, so this series is judged as a non-stationary series. Through logarithmic transformation and first-order difference transformation, the new series will be subject to the unit root test. The result shows that the p value is smaller than 0.05, so it is judged that the series is a stationary series after processing. The model is automatically determined by means of AIC and BIC statistics, and the ARIMA (0, 1, 1) model is used for the new series. The first-order MA coefficient is significantly lower than 0.05. The sample data is fit, and the result is shown in Figure 3. The blue color represents the original value, and the red color represents the predicted value. The autocorrelation coefficient and white noise test are adopted for residuals, and the p value of autocorrelation test over residuals is larger than

0.05, indicating that the residuals have no autocorrelation and are a white noise series. To sum up, the population size is predicted by using the correlation of population stock in time, and the outcomes of model fitting are good.

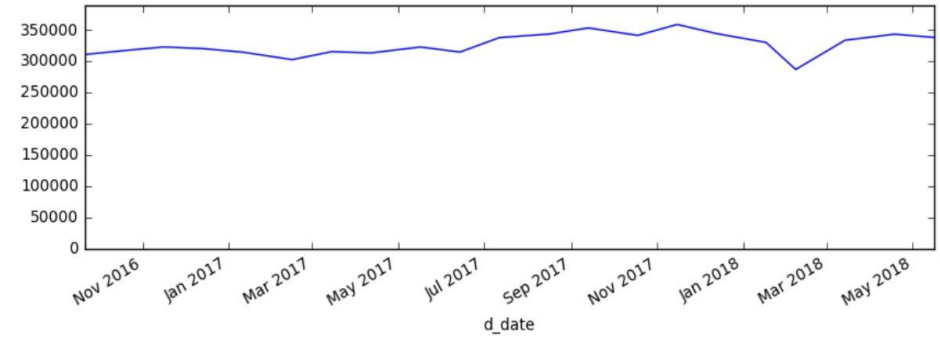


Figure 1. Historical Data about the Number of Mobile Phone Users in Tiantongyuan Area.

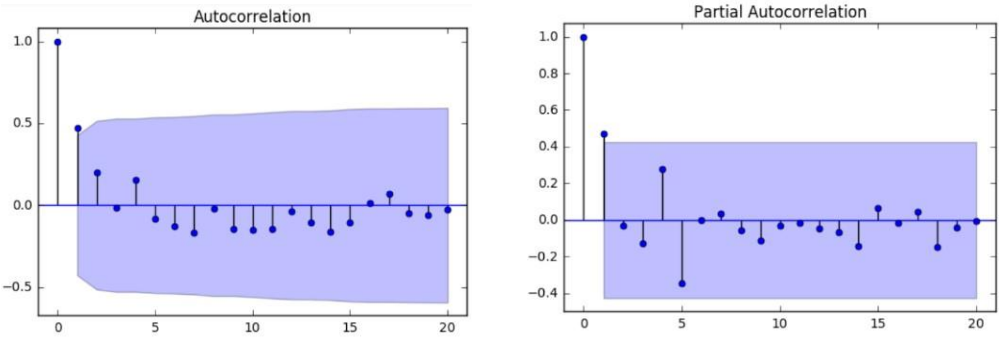


Figure 2. Autocorrelation Coefficient and Partial Autocorrelation Coefficient.

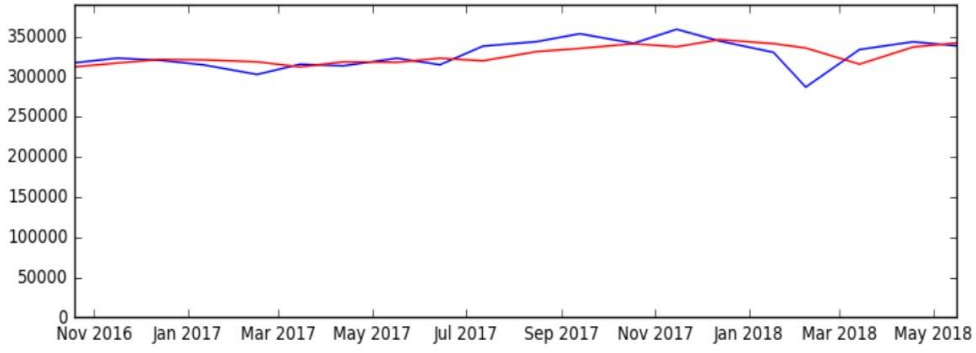


Figure 3. Fitting Result Based on the Time Series Correlation

(2) Similarly by taking Tiantongyuan area as an example, the geographic area is divided into the grids according to the area size of 100m*100m. The Euclidean distance between the target grid and the surrounding grids is calculated, and then the reciprocal of this distance is taken as a weight. The daily population size around the target grid from March 1 to March 21, 2019

is taken. The data from March 1 to March 11 is used as the training data size, and the data from March 12 to March 21 is used as a testing data set. In the training data set, the weighted mean value of population of the surrounding grids is calculated on a daily basis to get the daily non-standardized predicted value of population of the target grid (Table 1). Dividing the daily actual value of population of the target grid by the nonstandardized predicted value of population to get the standardized coefficient series (Table 2), and then the mean value of this standardized coefficient series is used as the standardized coefficient (here it is 0.032708827). So far, the training of model is finished.
In the testing data set, the population values of the grids around the target grid are taken. After the weighted mean value is calculated, it is multiplied by the standardized coefficient to get the predicted value of population of the target grid. The prediction results are shown in Figure 1. The daily prediction error is between -11.4 and 9.8, and the average error is 0.2 person. It can be seen that the prediction effect is good.

Date	Actual value of population	Predicted value of population (for training)
3-1	156	4767.418261
3-2	159	4859.243348
3-3	155	4633.040112
3-4	155	4981.36014
3-5	155	4487.921028
3-6	161	5009.078722
3-7	168	4883.601868
3-8	158	5012.276925
3-9	163	5091.594486
3-10	166	5114.32029

Table 1. Predicted Value of Population for Training.

Date	Standardized coefficient
3-1	0.032722113
3-2	0.032721144
3-3	0.033455355
3-4	0.031116
3-5	0.03453715
3-6	0.032141639
3-7	0.034400839
3-8	0.0315226
3-9	0.032013547
3-10	0.032457881

Table 2. Standardized Coefficient Series.

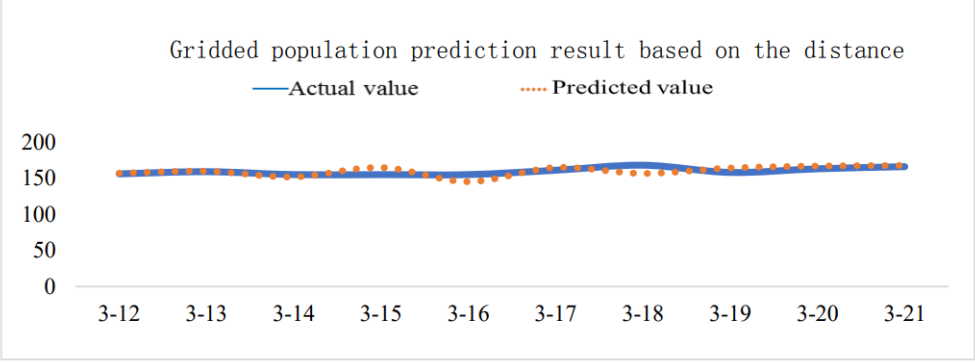


Figure 4. Prediction Based on Distance Correlation.

(3) The deep residual model is used to reckon and predict the population within the sixth ring road of Beijing. This area is divided into the grids according to the area size of 1KM*1KM, 72*72 grids in total. Observed from the actual people flow and prediction result during the morning peak from 07:00 to 09:00 on one working day in 2018, most of the grid losses are less than 10%, and more than 80% of the grid losses are less than 20%. According to the distribution situation of errors, the prediction effect of our model is good.

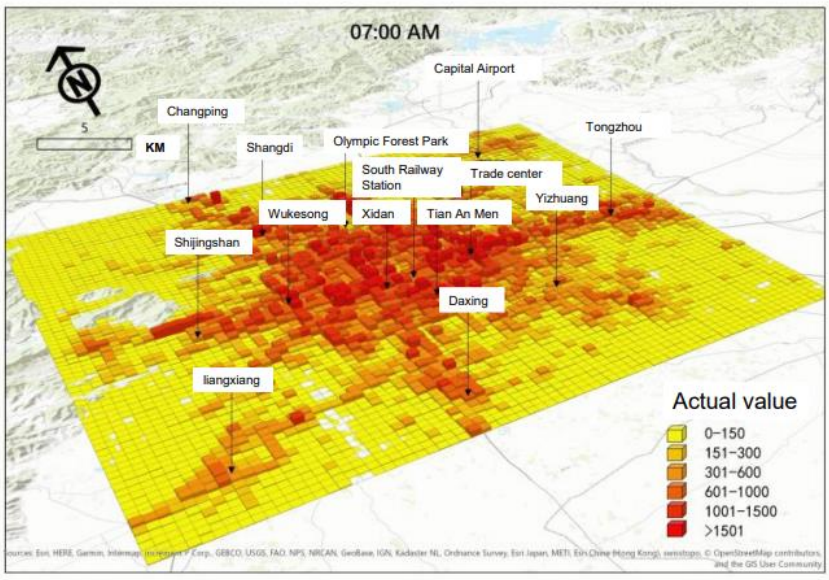


Figure 5. Distribution of Population within the Sixth Ring Road of Beijing.

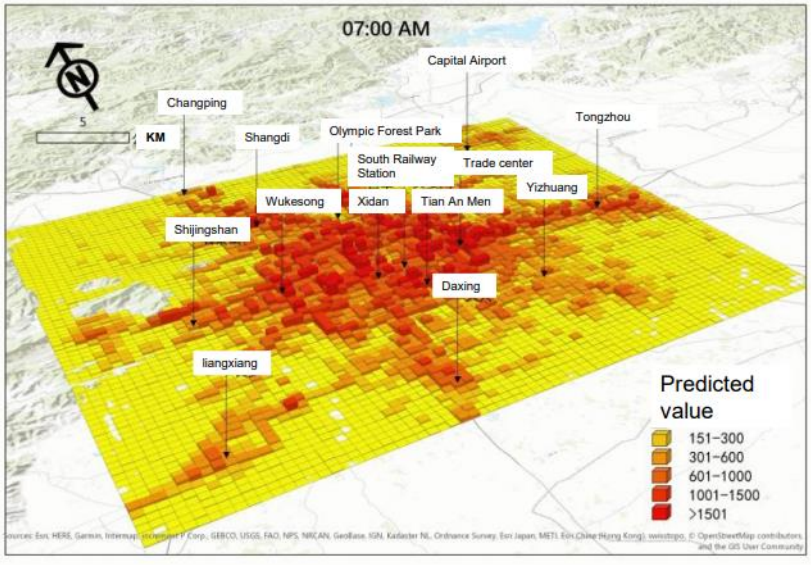


Figure6. Prediction Based on the Deep Residual Model.

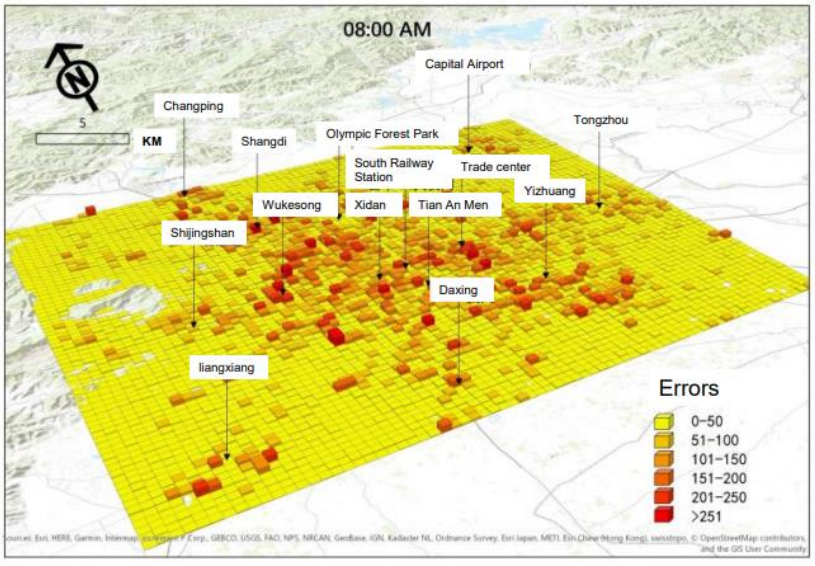


Figure 7. Errors of Prediction Based on the Deep Residual Model.

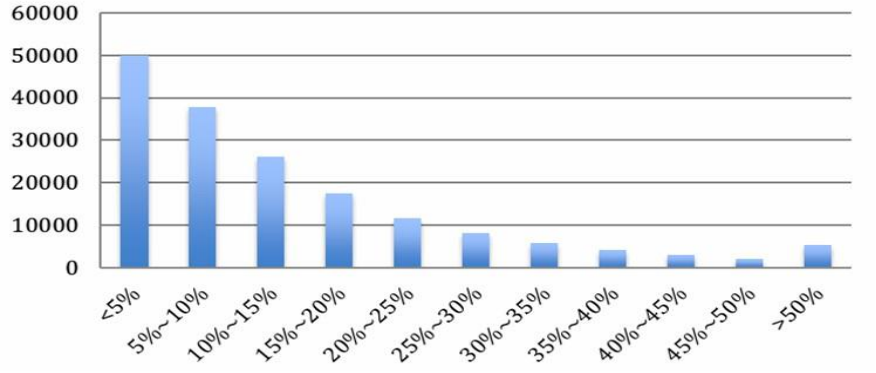


Figure 8. Distribution of Errors.

4. Discussion and Conclusion

As the change of population size is periodic in time, if it is observed on a daily basis, the population fluctuates from morning to night every day; if it is observed on a weekly basis, there are also significant differences on working days and at weekends; if it is observed on a yearly basis, the climate of four seasons and the holidays exert an obvious impact on people flow. Therefore, when the time correlation is used to predict the population size, to acquire a better prediction effect, either short-term prediction or time series with fixed intervals should be used to eliminate the impact brought by periodic factors.

If the deep residual model is used, part of the areas that show relatively large errors are the areas where the people flow is large, and the other part of the areas are the areas where the people flow is small, especially the urban border areas. It can be seen that this model still needs to be improved for the learning training and prediction of extreme value.

References

1. Zhang J, Zheng Y, Qi D. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction [C] //AAAI. 2017: 1655 - 1661.
2. Calabrese F, Lorenzo G D, Liu L, et al. Estimating Origin-Destination Flows Using Mobile Phone Location Data[J]. IEEE Pervasive Computing, 2011, 10(4):36-44.
3. Blondel VD, Decuyper A, Krings G. A survey of results on mobile phone datasets analysis [J]. Epj Data Science, 2015, 4(1):1-55.
4. 4.Xin Zheng, Qing Shen, Mingcui Du, Guojuan Li. Dynamic Monitoring of Population in Beijing City under the Big Data from Mobile Phones. Moving towards a High-quality Chinese Economy, 449- 467, 2018, Beijing: China Statistics Press.
5. Danhui Yi. Time Series Analysis: Methods and Applications. 2018, Beijing: Renmin University Press.

Harnessing the value of administrative data: A Central Bank's experience

Shazura Zainol Abidin, Basyirah Mohd Khairi, Chen Tze Ling
Bank Negara Malaysia, Kuala Lumpur

Abstract

Over the past few decades, the use of administrative data has witnessed a growing demand and interest globally, due to the large volume of data available and the low-cost nature of the collection process. In Malaysia, the establishment of institutional arrangements between Bank Negara Malaysia (the Bank) and several government agencies, mainly National Registry Department (JPN), National Property Information Centre (NAPIC), etc. has enabled the Bank to access a larger, population-based administrative records collected by these official compilers. The enhanced and enlarged scope of information resulting from integration of these administrative data with periodical reports and published information, is indeed a value add to support policy formulation, analysis and surveillance by the Bank. This paper aims to provide an overview of the Bank's experience in leveraging on administrative records as an alternative and supplementary source of data to further enhance and strengthen official findings, analyses and interpretations as well as to facilitate the Bank to assume the role as a central bank. It also aims to highlight the usage of administrative data by the Bank based on the existing arrangements as well as upcoming initiatives. The challenges including legal, policy, confidentiality, and technical issues will be discussed to address limitations, improve efficiency and strengthen institutional engagements in maximising the use of administrative data for the benefit of the Bank and the nation as a whole.

Keywords

Supplementary; institutional arrangement; entity; property; individual

1. Introduction

Many countries worldwide are increasingly using administrative data to improve statistical coverage in analysis and statistical publications. Administrative data refers to data collected by government bodies or other organisations in their day-to-day operations, where the data collection is not primarily meant for statistical purposes, which can be leveraged to enhance the quality of national statistics significantly, thus contributing to better policymaking and economic performance. The benefits of using administrative data as an alternative or complementary source includes reduction of cost,

maximised coverage and usage, increased timeliness and data frequency as well as lessening reporting burden in data compilation.

There are many successful analyses by other countries in using administrative data as sources for statistics on socioeconomic, healthcare, education, business and many other areas. For example, Singapore estimates its labour statistics based on contributor's records in the Central Provident Fund (CPF), while several countries such as Canada, Chile, and United Kingdom have been using their income tax records in compiling their monthly indices of economic activity, quarterly national accounts, annual GDP and business registers.

In view of its advantages, the Bank has been actively pursuing to acquire administrative data from various government agencies/institutions as an alternative source to complement the structured data being compiled by the Bank. To date, the Bank has obtained individual profiles from the National Registry Department (JPN) and granular data on property from the National Property Information Centre (NAPIC). All these micro-level, large-volume and high-frequency data, together with other sources of data, are value add to support policy formulation, analysis and surveillance by the Bank.

2. Use of Administrative Data at the Bank:

2.1 Individual Profiles from JPN

2.1.1 Data Checking Arrangement via Takaful and Insurance Deceased System (TIDeS)

New and still in progress, the data checking arrangement with JPN is one of the projects handled by the Bank which involves integration between information submitted by life insurers and family takaful operators (REs) on 12-digit National Registration Identity Card (NRIC) number of policy/certificate holders and individual profiles with deceased status from JPN's Registration of Death. The objective of this initiative is to improve efficiency of the claim process for death-related benefits and to strengthen pro-activeness of REs to reach out to the rightful beneficiaries for claim payment.

According to Section 130 of the Financial Services Act (FSA) read together with Schedule 10 of the FSA, an RE shall immediately notify the beneficiary of his entitlement to claim the policy moneys or takaful benefits where such beneficiary fails to claim the moneys/benefits within sixty days of the RE becoming aware of the death of the policy/certificate holder. However, in reality, the moneys/benefits may remain unclaimed in the combined situation of the beneficiary being unaware of the moneys/benefits and the RE being unaware of the death of the policy/certificate holder. Based on the important needs to address this, REs through the industry association i.e. Malaysia Takaful Association (MTA) and Life Insurance Association Malaysia (LIAM) have

engaged with JPN to explore the possibilities of establishing strategic collaboration to obtain the deceased status of policy/certificate holders from JPN's Registration of Death to facilitate death claims.

However, the information from JPN's Registration of Death can only be shared with government agencies, government-linked institutions and statutory bodies, such as the Bank. In line with one of the Bank's primary function to promote sound, progressive and inclusive financial system, the Bank has agreed to establish collaboration with JPN through the establishment of the "Integration and Data Sharing Specification" between both parties, where the initiative is deemed as an ancillary service offered by the Bank to the industry. To support this initiative, the Bank has internally developed Takaful and Insurance Deceased System (TIDeS), a secured platform to facilitate efficient submission and dissemination of 12-digit NRIC number between REs and JPN via the Bank.

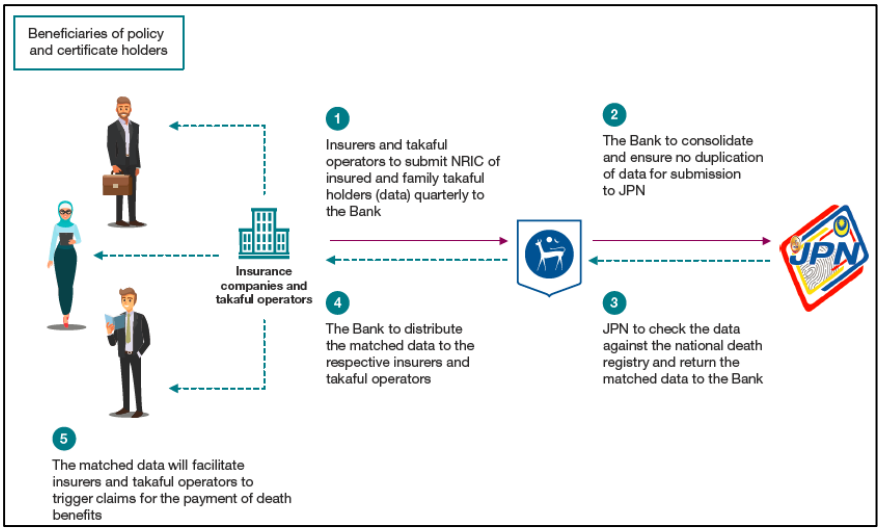


Diagram 2.1 Overall Process of TIDeS

Through the implementation of TIDeS (refer Diagram 2.1), REs shall submit the 12-digit NRIC number of life insurance policy and family takaful certificate holders to the Bank. Upon receiving data submission from REs, TIDeS will perform data validation, consolidate the NRIC numbers and ensure no duplication of NRIC numbers across the REs for submission to JPN. JPN will subsequently perform the matching of NRIC numbers against individual profiles with deceased status from JPN's Registration of Death and provide the list of matched NRIC numbers of deceased individuals to the Bank. TIDeS would rematch the list of deceased individuals by JPN against NRIC numbers submitted by REs earlier and disseminate the matched NRIC numbers to respective REs.

This initiative will benefit both the REs and policy/certificate holders where processing of claims for death related benefits could be triggered and expedited, thus, enabling REs to pro-actively reach out to the beneficiaries,

who are unaware of the insurance policy/takaful certificate of deceased holders. This improves the confidence of policy/certificate holders on the trustworthiness of insurance and takaful industry in Malaysia. Hence, policy/certificate holders can be assured that their inheritance are in good hands and will be distributed to the rightful beneficiaries upon death.

Nevertheless, there are risks associated with the implementation of TIDeS, due to the possibility of occurrences of incorrect death status in JPN's Registration of Death, arising from the declaration of death by court for a missing person and the registration of death for a deceased accident victim using stolen NRIC. The Bank may wrongly disseminate information to the REs, given that a person who is alive is reported as deceased by JPN. This is in addition to reputational risks involving the credibility and professionalism of the Bank. To mitigate the risks associated to this, the Bank through the policy issuance on Data Checking Arrangement with JPN – Guidance on Data Submission dated 12 April 2019, requires REs to establish a framework providing for an appropriate communication strategy in approaching the beneficiaries of a deceased policy/certificate holder given the sensitivity of the information relating to the death status and also the possibility of any inaccurate death status of such policy/certificate holder. REs also must not rely solely on the death status obtained from the Bank and should conduct their own due diligence to confirm the accuracy of the death status of a policy/certificate holder after obtaining such status under this arrangement and before making any communication with the beneficiaries of such policy/certificate holder.

2.1.1 Updating the 12-digit NRIC Number of Customers of REs

The benefits of leveraging on the alternative data sources have proven to be a value-add to increase the data quality at the Bank. In the Bank's attempt to further enhance data quality of statistical reporting by the REs in parallel with the request by the REs for the Bank's assistance via Financial Institutions Steering Committee (FISC[1]) to obtain the 12-digit NRIC numbers of their borrowers and accountholders from the Bank or directly from JPN, the Bank in early 2019 has embarked on a one-off exercise in updating the 12-digit NRIC number of customers of REs, who initially registered their identification (ID) number using the non-12-digit NRIC number while performing financial transactions with the REs.

This one-off exercise requires the REs to submit information on the names and old Individual NRIC/Army ID/Police ID/birth certificate numbers of their customers to the Bank, where the Bank will provide to JPN the customer's ID

[1] FISC is a high-level oversight committee and communication platform between the industry and the Bank.

to obtain the 12-digit NRIC number, if the individual profiles do not exist in the Bank's database. Upon receiving matched information on the 12-digit NRIC of the customer from JPN, the Bank will disseminate the information to the respective REs.

The purpose of this exercise is to facilitate accurate and consistent reporting of ID numbers of the customers in REs' statistical reporting submitted to the Bank, where it aims to establish a unique identification of individuals across the REs and at the Bank's database. This integrated information between REs, JPN and the Bank will be used to perform analysis and surveillance as well as to formulate prudent credit policies by the Bank to assist the REs in making informed and responsible lending decisions in a timely manner. Furthermore, this information can facilitate the assessment on the level of financial inclusion in Malaysia and support policy formulation to ensure all economic sectors and segments of the society have access to financial services.

As this exercise will establish a unique identification of individuals, it has benefitted Centralised Credit Reference Information System (CCRIS), a system developed by the Bank which collects credit information on borrowers from participating REs and supplies the information back to the REs, where the information will be used to assess customers' creditworthiness from the credit histories of potential/current borrowers. The unique identification provides accurate reporting of credit exposure of individuals, hence, enables participating REs in making faster and informed lending decisions. This facilitates to further improve the quality of REs credit assessment and risk modelling. In the national context, evidence-based lending decisions will reduce instances of loan defaults and household indebtedness, both of which are currently notable issues in the Malaysian economy.

Besides, the need for unique identifiers is reflected through the Financial Inclusion Survey (FIS), a statistical reporting by REs to the Bank, established in line with the Bank's mandate in promoting a sound, progressive and inclusive financial sector. The data collected via FIS, integrated with unique identifier from this exercise, enables the Bank to accurately obtain the actual number of Malaysian individuals who have access to various types of essential financial services, namely deposit accounts, financing and insurance, and eliminate the risks of double counting of individuals with multiple financial accounts/products within and across the REs. Hence, this exercise enhances the REs reporting to enable accurate assessment on the level of financial inclusion in Malaysia and drive financial inclusion policies in targeting financially-excluded segments of the population. This improves the overall well-being of communities on the aspects of convenient accessibility, high take-up, responsible usage and high satisfaction of financial services, which

contributes to a balanced as well as sustainable economic growth and development.

Notwithstanding the various benefits of leveraging on JPN administrative records to obtain unique identifier to enhance CCRIS and FIS reporting, there are also risks involved in the compilation and usage of these records. As these data are highly confidential and comprise personalised data, it is of utmost importance to manage and safeguard the confidentiality and integrity of the data being compiled and disseminated by providing adequate control, access and protection. This is to prevent misuse of information for fraudulent activities, phishing, identity theft as well as for marketing purposes.

2.1 Property Data

2.2.1 The Use of National Property Information Centre (NAPIC) Data

The Bank has been aggressively pursuing to source administrative data from various agencies/institutions, including property data from NAPIC. NAPIC is a property centre to monitor the growth of the property market in the country, where it was established at the insistence of the National Economic Action Committee (MTEN). NAPIC is responsible to collect property demand and supply data from various parties, develop and maintain a national property stock warehouse, provide accurate, comprehensive and timely information to government agencies and all other parties involved in the property industry as well as advice the government on property development in the country. The data compiled by NAPIC enables consumers, developers and regulators to make better analysis for informed decisions in property-related transaction, business strategy and policy formulation.

In view of its advantages, the Bank has established data sharing arrangement with NAPIC since 2014, which comprises aggregated data on property market, such as Malaysia House Price Index, average house prices, property transactions by state, price range and sub-sector, newly launched residential, overhang property by type and state, existing supply of hotel rooms by star rating and number of hotel by number of rooms. In early 2017, the Bank has further reviewed the existing data shared by NAPIC and requested to obtain additional data due to the increased demand from the Bank to analyse more detailed information related to the property transaction, to enable better usage, analysis, surveillance and research on housing. Upon further engagement and deliberation, NAPIC has agreed to share with the Bank granular data on housing, mainly data on property transactions, inventory, market status and new launches, where the first batch of additional data was received in December 2017.

The data sourced from NAPIC has facilitated the production of report for internal analysis and publication in the Bank's website. Besides, NAPIC data enables analyses from various perspectives of the property market to gain

better insight and understanding on the behaviour of the property market as well as to gauge where the demand lies, particularly by the property location and type for greater clarity. This enables policy formulation for a balanced property supply and demand as well as promoting affordable housing. Some of these analyses were published in the Bank's Quarterly Bulletin, which includes the volume and value of residential property transaction as well as the vacancy and rental rates of office space and shopping complexes. The Bank has also provided analysis on the affordability of property in the Financial Stability and Payment Systems Report published annually.

The Bank envisages for all the micro-level and large-volume administrative data, including NAPIC data and together with other sources of Big Data, to be stored and integrated in the Big Data Analytics Platform (BDAP), to facilitate data science and analytics initiatives in the Bank. This is to enable the larger dataset to build better, more accurate models and analysis for the Bank to be a leading evidence-based decision making institution with advanced data capabilities across all functions of the Bank.

Despite the value-add of this arrangement which facilitates comprehensive and timely property-related publications and analysis by the Bank, NAPIC requires three months after the end of the reporting quarter for data quality process to be conducted prior to data sharing, in line with NAPIC's mission to always provide comprehensive, quality and up-to-date property data for its stakeholders. However, there is no major implication since the Bank is still able to meet the property-related data needs.

3. Discussion and Conclusion

In a nutshell, administrative data sourced from government bodies and other institutions has proven valuable in enriching the Bank's operations through insightful implementation. Administrative data from JPN enables the insurance/takaful companies to pro-actively reach out to the beneficiaries of deceased policy/certificate holders, which increases the level of trust and confidence in the safety and soundness of the insurance and takaful industry, thus contributing towards a progressive financial sector. Besides, the use of administrative data from JPN improves the quality of industry credit assessment and risk modelling through timely and informed lending decisions, which lead to sustainable economy aiming towards declined loan defaults contributed by household indebtedness. JPN records also facilitate accurate assessment on financial inclusion level in Malaysia, which enables implementation of focused strategies in promoting financial inclusion as mandated to the Bank. Furthermore, property-related administrative data sourced from NAPIC allows policy formulation for a balanced property supply and demand as well as promoting affordable housing.

The usage of large-volume, micro-level and granular administrative data together with the periodical data compilation provides more statistical power to official statistics, thus bringing surveillance, analysis and regulatory obligations to a whole new level. Furthermore, by using advanced technologies, administrative data with other sources of Big Data could be stored and integrated in BDAP, to facilitate data science and analytics initiatives in the Bank and to enable better analysis and more accurate models in supporting the Bank's surveillance. The provision of administrative records as supplementary source of data has fostered cost-effectiveness since these data are already collected routinely for administrative purposes, hence, it reduces the reporting burden of the industry.

However, in the course of leveraging on administrative data, the Bank faces challenges and limitations arising from various aspects, namely legal, data confidentiality and access control, security as well as technical issues. In the Bank's attempt to obtain administrative data from multiple sources, organisational challenges become a major issue since different institutions are governed by different respective legal and institutional frameworks. The data sharing arrangement entails authority approval and should be permissible by the respective law. It also requires adequate time and effort for involved organisations to conduct assessment and agree with the terms, as well as to comply with the governance and procedure, which includes obtaining approval from relevant authorities prior to institutional agreement. The institution's vision and perception in supporting the idea of data sharing and transparency among the organisations have significant influence over the decision on the agreement for collaboration.

Since majority of the administrative data involves micro-level and personalised data, it is important to establish a set of policies, procedures and processes, and standards to ensure data are formally managed and to instil discipline in managing the whole lifecycle of these data. This includes managing and safeguarding the confidentiality and integrity of the data being compiled and disseminated by providing adequate control, access and protection to prevent misuse of information for fraudulent activities, phishing, identity theft as well as for marketing purposes. Given that the technology sector is an inescapable exception in advancing the use of administrative data by the Bank, information technology (IT) investment strategies are essential to manage the data through the development of a secured data submission, processing and dissemination platform to stay vigilant of technology frauds. As such, cost is one of the challenges in obtaining allocation for IT budgets, which is required to facilitate an organisation to drive the needs for the development of a safe and secured system, concurrently maintaining operational efficiencies.

While there are challenges and risks associated with sourcing and establishing institutional sharing of administrative data, the Bank is positive to pursue these initiatives to lavish in the benefits of the low-cost, micro-level administrative data in supporting policy formulation, analysis and surveillance. Moving forward, the Bank envisages broader needs and demand for the usage of administrative data with greater collaboration among the organisations, including Inland Revenue Board (LHDN), SSM etc., to leverage on the huge potential of administrative data. However, rigorous effort to pre-emptively address issues and challenges arising from the data sharing arrangement needs to be established and agreed collectively among the organisations in Malaysia towards improving the quality of official statistics for better collaboration, transparency, accessibility and overall society well-being.

References

1. Administrative Data Liaison Service -Economic and Social Research Council (2019). Retrieved from https://esrc.ukri.org/research/our-research/administrative-data-liaison-service/
2. Asian Development Bank (2010). Administrative Data Sources for Compiling Millennium Development Goals and Related Indicators.
3. Bank Negara Malaysia (n.d.). Retrieved from http://www.bnm.gov.my/
4. Borhan Nordin, S. H., Lim, S. L. & Abd Aziz, M. K. M. (2018). Indebted to Debt: An Assessment of Debt Levels and Financial Buffers of Households. Retrieved from http://www.bnm.gov.my/index.php?ch=en_publication&pg=en_work_papers&ac=65&bb=file
5. Credit Bureau (2014). Retrieved from http://creditbureau.bnm.gov.my/
6. Management of Customer Information and Permitted Disclosures. (2017). Retrieved from http://www.bnm.gov.my/index.php?ch=57&pg=144&ac=632&bb=file
7. National Property Information Centre (2019). Retrieved from http://napic.jpph.gov.my/portal
8. Overview of Financial Inclusion in Malaysia (2019). Retrieved from http://www.bnm.gov.my/index.php?ch=fi&pg=fi_ovr&ac=471&lang=en
9. Rivas, L. & Crowley, J. (2018). Using Administrative Data to Enhance Policymaking in Developing Countries: Tax Data and the National Accounts.
10. United Nations (2011). Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.
11. Yusoff, I. (2008). Managing Issues Addressing the Challenges of Using Administrative Data for Statistical Purposes. United Nations Statistical Institute: Seventh Management Seminar for the Heads of National

Statistical Officers in Asia and the Pacific, Shanghai. Retrieved from https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2437

Sourcing of administrative record for official compilers

Wan Zarazillah, Farid Ahmad
Companies Commission of Malaysia

Abstract

Recent advances in technology have resulted in more agencies linking up with Suruhanjaya Syarikat Malaysia (SSM) and other government bodies who have been compiling detailed official records at the point of registration or when updated by individuals or companies. Although the information is not primarily collected for statistical purposes, the details and on-going data collection on their clients can be a valuable source of information for statistical reports. SSM has been continuously providing its administrative records when requested to complement statistics collected by the official compilers, such as Department of Statistics Malaysia (DOSM) and Bank Negara Malaysia (BNM). While some requests may require resources to meet specific requirements by the compilers, the exercise is a value-add for both parties towards an in-depth understanding of data compilation and better data quality.

Keywords

SSM, MBRS, Core System, Data Warehouse, Data Manipulation.

1. Introduction

The Companies Commission of Malaysia or SSM is a result of the merger between the Registrar of Companies (ROC) and the Registrar of Business (ROB) in Malaysia which regulates companies and businesses. SSM came into operation on 16 April 2002.

The main function of SSM is to serve as an agency to incorporate companies and register businesses and limited liability partnerships (LLPs) as well as to provide information on companies, LLPs and businesses to the public. As the leading authority for corporate governance, SSM fulfills its function to ensure compliance with business and corporate legislation through comprehensive enforcement and monitoring activities to sustain positive developments in the corporate and business sectors of the Nation.

As the regulator of the registration of business entities in Malaysia, the Companies Commission of Malaysia or SSM is constantly improving its services and products in order to accommodate the needs from the business community and create a conducive business environment.

SSM receives submission of documents (statutory documents) in a paper-based format via its counters or any channel of submission (i.e. online or kiosk) throughout Malaysia on a daily basis. Further, SSM faces an influx of documents twice a year i.e. during the peak period when there is a high volume of annual returns and financial statements. Due to this fact, SSM is in the midst of developing a new system to enhance its service delivery under the SSM Transformation Program (SSMTP) (which is a project under the SSM Second and Third Direction Plans) which will see the introduction of a new core system replacing the current legacy system.

With this new core system, expected to be fully implemented in the fourth quarter of this year, it will allow SSM to better utilize the data stored within the business registry. Furthermore, under the SSMTP, one (1) of the highly anticipated projects is the introduction of the eXtensible Business Reporting Language (XBRL) or as it is referred to in Malaysia as the Malaysian Business Reporting System (MBRS) which is an open international standard for digital business reporting. Technological advances in recent years have given a new digital platform for financial information and XBRL has been identified as an enabling technology that allows paperless financial reporting.

MBRS in Malaysia will be primarily used for the filing of financial statements. It is strongly believed that MBRS is the way forward to ensure that accessibility of information is made easier for investors, regulators and the public at large is improved and enhanced. The key advantage of this new "language" is that an identifying tag applies for each individual item of data as opposed to storing block text of more than one million companies nationwide.

2. Methodology

With the development of the SSM core system, this will improve the sharing of information and statistical data with other agencies such as Department of Statistics Malaysia (DOSM), Central Bank of Malaysia (BNM), and Inland Revenue Board of Malaysia (LHDNM). As the custodian of corporate data, the data on business, company, and limited liability partnerships (LLPs) reside in SSM's Data Warehouse and this includes historical data.

The data warehouse is a system where SSM keeps the data filed with SSM and this data is to be used for reporting and data analysis. It is considered as a core component of business intelligence. The data warehouse is a central repository of integrated data from one or more disparate source. All the information from business entities can be mined from the data warehouse which integrates with SSM's registration systems.

The information can be collected online, where these services are available 24/7 and accessible anywhere. Each information and data can be purchased with different types of packages according to the requirement by the public. Where every profile and detail of each business entity reside in the database,

it can be requested accordingly from the data warehouse. Whether the information on ownership, shareholder, company secretary, addresses, financial statements, and others.

The cycle of preparing the data will start by receiving applications from stakeholders who are mostly from the public, ministry, government agencies, and foreign investors. Officers in charge will identify the requested application by selecting the table from the data scheme tables. The selected tables will be compiled to become a data set before it is converted to either PDF, Excel, or CSV format. From the data warehouse, SSM can use the data for supply, data analysis or enforcement matter. Stakeholders who request the data from SSM can use it for further analyzing, reporting, mining, and decision making.

We can see an increase in the number of requests, especially for information pertaining to business profile and statistics reporting among the Malaysian business community. The information held by SSM regarding businesses, companies, and LLPs will allow the market players the opportunity to make better decisions and the analytical data will assist in providing information on future business opportunities. The data analytic function of the core system would allow for data analysis, comparisons, forecasting, and creating relationships. Things will be simplified and cost effective for the distribution of data.

Introducing MBRS for annual and financial filings is a major step for SSM, facilitating greater transparency and simpler analysis by helping to gather large quantities of high-quality data. The data contents of Financial Statements, Annual Return, Exemption Application, and Key Financial Indicator are defined as taxonomies.

The elements/taxonomies contained in MBRS in Malaysia through SSM are mainly based on the Malaysian Accounting Standards (MFRS/MPERS) and Companies Act 2016. The data are defined within the XBRL taxonomy and organised based on different file types within the taxonomy architecture. The process includes Organising, Acquiring, Validating, Protecting, Storing, and Processing the data.

3. Results

At the moment, there are many products offered by SSM, one of which is the Corporate and Business Information Data (CBID) which is a customizable data for both companies and businesses that you can personalize. The data includes a list of companies or statistical data. It helps make business decisions more efficiently based on facts and business information. With the new systems which will be introduced soon and also with the introduction of MBRS, SSM will also introduce new Corporate Business Information Data (CBID) packages which will allow stakeholders to choose data schemes according to their requirement and needs at a cheaper cost. The varied combination of data

is countless under the packages and the data stored within SSM. It allows them to pick and choose any type of data they require for reference and decision-making purposes. It is a tool that provides the right information at the right time, and it can be significantly improved business outcomes by helping employees to become more productive and efficient, allocating budgets optimally, or boosting profits.

The information contained within the business registry on the annual and financial filings from MBRS would benefit SSM in terms of data collection and analysis purposes. The data in SSM are dynamic and its flexibility is synchronized with changes that happens every day with registration, submission, and lodgement when received by SSM. It is consistent with data entry by human and automatically generated by the registration systems. Data collection is an important aspect as it will assist SSM at the research end and assist the Government in formulating new policies and measures to enhance the business landscape and community in Malaysia. Hence, helping the government to impose any policies and measurement for the business community, improving their approach towards better nations.

The implementation of MBRS would facilitate Malaysia as a whole in exchanging or sharing company information with other countries which have similar XBRL taxonomies. Furthermore, it is also capable of translating information into various languages. The quality of data that has been gathered informed decision facts, trends, and statistical numbers. Furthermore, the data collected does not change the prevailing accounting standards or regulations.

4. Discussion and Conclusion

The integrity of the source data can be relied upon in a more assured manner since the creation of data would include validity checks, mathematical automated calculations, and elimination of transcriptional errors. In order to provide superior service delivery through operational excellence, the source data has to be reliable and accurate before it can be used.

Only SSM can assure that since it is the custodian of business data in Malaysia. SSM would be able to facilitate the analysis of financial reports to aid investigation efforts (e.g. Anti-money laundering, fraud and other compliance matters) and enhance corporate compliance processes and improves data analysis and quality of information for enforcement purposes. SSM also provides comprehensive information and data which can be aggregated and sold to stakeholders in the form of industry analysis or industrial benchmarking.

Joint modeling of discrete-valued marker process, competing recurrent events process, and discrete-valued health status process

Edsel A. Pen˜a[1], Piaomu Liu[2]
[1]Department of Statistics, University of South Carolina Columbia, SC 29208 USA
[2]Department of Mathematics and Statistics, Bentley University Waltham, MA 02452

Abstract
This talk will describe an integrated class of joint stochastic models for a discrete-valued longitudinal marker process, a competing recurrent event process, and a health status process in studies where subjects or units are dynamically observed over possibly random monitoring periods. This class of models is potentially of high relevance and importance in the context of precision or personalized medicine which aims to utilize complex, unstructured, and big data for the purpose of implementing personalized interventions in this futuristic and fast-developing precision medicine approach.

Keywords
Competing risks; Continuous-time Markov chain; Counting process; Precision medicine; Stochastic process models

1. Motivation

The potential and promise of precision or personalized medicine hinges on the availability of appropriate mathematical and stochastic models, together with the proper statistical inference procedures, of the complex, most possibly unstructured, and big data underpinning the decision-making process in this modern and futuristic approach. In this talk we describe an approach to the joint modeling of three major components that could be synergistically interacting for the subjects or units in studies in biomedical setting, engineering settings, and even in socio-economic settings. These three components pertain to a marker process, recurrent event process, and a 'health' status process, together with the impact of covariates and interventions that are dynamically performed for these units. The class of joint models will be described here. Due to space limitations, statistical inference procedures, in particular, the estimation of model parameters, as well the potential relevance and applications in precision medicine, together with the assessment of the 'quality-oflife' (QOL) of units, will be deferred for the talk in

the World Statistics Congress (WSC) in Kuala Lumpur and in future manuscripts.

2. Proposed Joint Modeling Approach

Consider a subject or unit in a biomedical, engineering, or socio-economic setting which is monitored over time. This unit will be monitored over a period $[0, \tau]$, where τ could either be fixed in advance or it could also be random. Associated with this unit will be a covariate vector, denoted by X, representing relevant demographic features. Of main interest is to determine the health status of this unit over time. This will be represented by a process $V = \{V(s) : s \geq 0\}$ which takes values in a finite state space $\mathbb{V} = \mathbb{V}_1 \cup \mathbb{V}_0$ where elements of $\mathbb{V}_1$ are transient states, whereas those in $\mathbb{V}_0$ are absorbing states. These absorbing states may correspond to different competing terminal events, e.g., deaths due to competing causes. The lifetime of this unit will then be

$$S = inf\{s \geq 0 : V(s) \in \mathbb{V}_0\}.$$

Aside from this health status process, there will also be associated with this unit a longitudinal marker process, the second component, represented by $W = \{W(s) : s \geq 0\}$, which takes values in a finite state space $\mathbb{W}$. This marker process provides information about the health status of the unit and vice-versa. At any given point in time, the unit will be in one of these states in $\mathbb{W}$. The third component in our setting is the presence of several Q types of recurrent events. The occurrences of these recurrent events, which are competing with each other, will be tracked by a multivariate counting process $N = \{N(s) : s \geq 0\}$ which takes values in $\mathbb{Z}_{0,+}^{Q}$, where $Z = \{0,1,2,\dots\}$. Similarly to the marker process, the recurrent event process is also affected by the health status process and vice-versa, and there will also be synergistic interaction between the marker and the recurrent event processes. Another important feature governing such systems is the performance of an intervention at each recurrent event occurrence which impacts the subsequent rate of occurrences of these recurrent events.

A bio-medical situation where this setting occurs is that where the health status $V(s)$ of a patient could be in the state space $\mathbb{V}$ = {v_1 = healthy, v_2 = diseased, v_0 = dead} so that $\mathbb{V}_1$ = {$v_1 v_2$} and $\mathbb{V}_0$ = {, v_0}. Thus, v_0 is an absorbing state. The blood pressure (BP) marker process $W(s)$ could take values in the state space $\mathbb{W}$ = {w_1 = Normal BP, w_2 = Low BP, w_3 = High BP }. The competing recurrent events could be hospitalizations due to different causes or ailments.

We shall denote by $\mathbb{F} = \{\mathfrak{F}_s : s \geq 0\}$ the filtration or history governing this unit. Thus, all the stochastic processes considered, such as V, W, N, etc., will be adapted to this filtration. To mathematically simplify our modeling

approach, we shall assume a Markovian property, an assumption typically made and which is realistic under most situations. We now describe the details of our proposed joint stochastic model.

The health status process V will be a continuous-time Markov Chain (CTMC). Thus, there is a probability mass function (pmf) over $\mathbb{F}$, denoted by $p_v(\cdot)$, governing the state of $V(0)$. This pmf may contain unknown parameters. There is then a *baseline* infinitesimal generator matrix $\eta = (\eta(v,v') : v,v' \in \mathbb{V})$ such that

$$\forall v \neq v' \in \mathbb{V} : \eta(v,v') \geq 0 \quad \text{and} \quad \forall v \in \mathbb{V} : \eta(v,v) = -\sum_{v' \in \mathbb{V}} \eta(v,v').$$

Since $\mathbb{V}_0$ are absorbing states, we have that for all $v \in \mathbb{V}_0, \eta(v,v') = 0$. Similarly, the marker process is also a CTMC, with its initial state, $W(0)$, governed by a pmf $p_w(\cdot)$ over $\mathbb{W}$. This pmf may also have unknown parameters. The *baseline* infinitesimal generate matrix for this marker process will be $\zeta = (\zeta(w,w'): w,w' \in \mathbb{W})$ such that

$$\forall w \neq w' \in \mathbb{W} : \zeta(w,w') \geq 0 \quad \text{and} \quad \forall w \in \mathbb{W} : \eta(w,w) = -\sum_{w' \in \mathbb{W}} \eta(w,w').$$

The holding times and transition probabilities for both of these processes, which will be governed by these infinitesimal generators, will be further modulated by the covariate vector and the effects of the other two components. These will be described below after introducing the model elements for the recurrent event process.

For the recurrent event process, we take into consideration to aspects underlying the monitoring of the occurrences of recurrent events. As articulated in the papers [3, 4], there is a need to take into consideration the impact of performed interventions at each event occurrence as well as the impact of the accumulating event occurrences. In addition, for more generality and to be more applicable in the bio-medical setting, the modeling approach is usually via a semi-parametric model. Thus, for the qth type among the Q recurrent event types, we introduce an effective age process $\mathcal{E}_q = \{\mathcal{E}_q(s): s \geq 0\}$ which is a dynamically observable nonnegative piecewise continuous $\mathbb{F}$-predictable process, a baseline nonparametric hazard rate function $\lambda_q(\cdot)$, and a nonnegative function $\rho_q(\cdot;\alpha_q)$ defined over $\mathbb{Z}_{0,+}^{Q}$ and dependent on an unknown parameter vector α_q. This function will encode the impact of the accumulating event occurrences on the rate of recurrent event occurrences.

Finally, we will have an at-risk process $Y = \{Y(s) : s \geq 0\}$, where $Y(s) = I\{\tau \geq s, S \geq s\}$, where $I(\cdot)$ is the indicator function. Thus, $Y(s)$ indicates the subject or unit is still under observation at time s. The process Y is a bounded, left-continuous (hence $\mathbb{F}$-predictable) process. We are now in position to describe our proposed joint model. We first introduce the mappings:

$$\iota_{\mathbb{V}}(v) \equiv (I(v' = v), v' \in \mathbb{V}) \quad \text{and} \quad \iota_{\mathbb{W}}(w) \equiv (I(w' = w), w' \in \mathbb{W}).$$

The class of joint stochastic models could now be described as follows. For brevity, we use the alternate notation $A_s \equiv A_{(s)}$ for any process A.

1. The health status process satisfies the condition: For all $v, v' \in \mathbb{V}$ with $v \neq v'$,

$$\lim_{h\downarrow 0} \frac{1}{h} \Pr\{V((s+h)-) = v' | \mathfrak{F}_{s-}; V_{s-} = v; W_{s-} = w; N_{s-} = n\} = Y(s)\eta(v, v') \exp\{X\beta_1 + \iota_{\mathbb{W}}(w)\kappa_1 + n\gamma_1\};$$

and

$$\lim_{h\downarrow 0} \frac{1}{h} [\Pr\{V((s+h)-) = v | \mathfrak{F}_{s-}; V_{s-} = v; W_{s-} = w; N_{s-} = n\} - 1] = Y(s)\eta(v, v) \exp\{X\beta_1 + \iota_{\mathbb{W}}(w)\kappa_1 + n\gamma_1\}.$$

We recall here that $\eta(v, v) = -\sum_{v' \in \mathbb{V}} \eta(v, v')$.

2. The longitudinal marker process satisfies the condition: For all $w, w' \in \mathbb{W}$ with $w \neq w'$,

$$\lim_{h\downarrow 0} \frac{1}{h} \Pr\{W((s+h)-) = w' | \mathfrak{F}_{s-}; V_{s-} = v; W_{s-} = w; N_{s-} = n\} = Y(s)\zeta(w, w') \exp\{X\beta_2 + \iota_{\mathbb{V}}(v)\delta_2 + n\gamma_2\};$$

and

$$\lim_{h\downarrow 0} \frac{1}{h} [\Pr\{W((s+h)-) = w | \mathfrak{F}_{s-}; V_{s-} = v; W_{s-} = w; N_{s-} = n\} - 1] = Y(s)\zeta(w, w) \exp\{X\beta_2 + \iota_{\mathbb{V}}(v)\delta_2 + n\gamma_2\}.$$

We recall here that $\zeta(w, w) = -\sum_{v' \in \mathbb{V}} \eta(w, w')$.

3. The multivariate counting process describing the recurrent events occurrences satisfies: For infinitesimal positive h, with

$$dn \in \{(0,0,\ldots,0), (1,0,\ldots,0), (0,1,\ldots,0), \ldots, (0,0,\ldots,1)\},$$

we have

$$Pr\{N((s+h)-) - N(s-) = dn | \mathfrak{F}_{s-}; V_{s-} = v; W_{s-} = w; N_{s-} = n\} = Y(s) \prod_{q=1}^{Q} \left[\{\lambda_{0q}[\mathcal{E}_q(s)]\rho_q[n; \alpha_q] exp\{X\beta_3 + \iota_{\mathbb{V}}(v)\kappa_3 + \iota_{\mathbb{V}}(v)\delta_3\}h\}^{dnq} \times \{1 - \lambda_{0q}[\mathcal{E}_q(s)]\rho_q[n; \alpha_q] \exp\{X\beta_3 + \iota_{\mathbb{V}}(v)\kappa_3 + \iota_{\mathbb{V}}(v)\delta_3\}h\}^{1-dnq}\right].$$

The simplest effective age process could either be of form

$$\mathcal{E}_q(s) = s \quad \text{or} \quad \mathcal{E}_q(s) = s - S_{Nq(s-)},$$

where the first form results from performing *imperfect* repairs or interventions at event occurrences, while the latter form results from performing a *perfect* repair or intervention after the last event occurrence prior to time s, and S_k is the time of the last event occurrence prior to time s. The effective age processes are determined dynamically since the repairs or interventions performed after each recurrent event occurrence are usually not determined at time zero but decided upon after the event occurrence. The ρ_q functions on the other hand could for instance be of form

$$\rho_q(n{:}\,\alpha_q) = \exp\{[\log(1+n)]\alpha_q\}, n \in \mathbb{Z}_{0,+}^{Q}$$

4. The final requirement to completely specify the joint stochastic model is the assumption that, given $\mathfrak{F}_{s-}$, then $\{V(t), t \geq s\}, \{W(t), t \geq s\}$, and $\{N(t), t \geq s\}$ are conditionally independent. This conditional independence assumption enables the construction, in a dynamic fashion, of the full likelihood function or process.

3. Some Aspects of the Class of Joint Models

The distinctive trait of this joint model is the interplay among the three components: the health status, the marker, and the recurrent events. Each of these affect the others in the sense that the future occurrences of transitions or events, given the present, for each of the components are affected by the current state of the other two components. As such dependencies of the random paths are induced and there is a synergistic dynamicity to the paths of the different processes. Each of them have some baseline behavior which are encoded in the baseline parameters: the infinitesimal generator η for the V-process; the infinitesimal generator ζ for the W-process; and the baseline hazard rate functions $\lambda_{0q}s$ for the N-process. Some form of proportionality is then imposed to model the modulation induced by the other components and the covariate vector through the exponential link functions.

There are many model parameters in this joint model, which implies that in order to perform reasonable inference, a sufficient number of subjects or units over reasonable monitoring periods will be required. The model parameters are:

- Parameter of$p_v(\cdot)$ and parameter of $p_W(\cdot)$.
- Baseline infinitesimal generators $\eta(v, v'), v, v' \in \mathbb{V}$.
- Baseline infinitesimal generators $\zeta(w, w'), w, w' \in \mathbb{W}$.
- Baseline hazard rate functions $\lambda_{0q}(\cdot), q = 1,2,\ldots,Q$, which are specified nonparametrically.

- The regression coefficients β_1, κ_1, and γ_1 in the sub-model for the evolution of the V -process.
- The regression coefficients β_2, δ_2, and γ_2 in the sub-model for the evolution of the W-process.
- The regression coefficients $\beta_3, \kappa_3,$, and δ_3 in the sub-model for the evolution of the N-process.
- The parameters $\alpha_q s$ in the functions $\rho_q(\cdot; \alpha_q) s$ in the N-process sub-model.

In the CTMCs for the V - and W-processes, ordinarily the state holding times and the transition probabilities are completely determined by the infinitesimal generators, but in our model these are affected by the other components through the exponential link functions and the relevant regression coefficients. Since the values of the processes may change at each event or transition time, then the state holding times are governed by piecewise exponential distributions, but where the changes occur are determined by where the transitions or events occur, hence are dynamic in some sense.

In each of the models for the three components, there exists a 'competing risks' aspect. In the V-process, the states are in some sense competing with each other.

This is also the case with the W-process; and also with the N-process. Thus, when the likelihood function is constructed, this competing aspect needs to be incorporated, but this is immediately taken cared of by the likelihood construction using Jacod's [2] (see also [1]) approach.

4. Statistical Inference Issues

Of critical importance is to be able to infer about the model parameters of this class of joint models in order that the model could be used in practice. Such statistical inference will be based on independent observations of n subjects or units that are monitored over their respective monitoring periods. For the ith unit the random observables X_i, V_i, W_i, N_i, and Eqi are observed over $[0, \tau_i]$. The likelihood process is then constructed from their realizations. However, due to space limitations, we do not present the statistical inference approach in this paper, but defer its discussion for the talk during the WSC. Suffice it to say that the first step in performing the statistical inference is the construction of the appropriate likelihood process. This is constructed by exploiting the Markovian structure and also the conditional independence among the three components given the present. Inference for the parameters

within each sub-model could then be performed separately, but note that these sub-model likelihood functions depend on data from the other components. Also, because the nonparametric baseline hazard rate functions are being evaluated at the effective age functions, to obtain the estimates of their associated cumulative hazard functions, a time-change approach implemented in [4] is required. Other inferential aspects will be discussed at the WSC talk by the first author.

References

1. P. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.
2. J. Jacod. Multivariate point processes: predictable projection, radon-nikodym derivatives, representation of martingales. *Z. Wahrsch. verw. Geb.*, 34:225–244, 1975.
3. E. Pen˜a and M. Hollander. *Mathematical Reliability: An Expository Perspective* (eds., R. Soyer, T. Mazzuchi and N. Singpurwalla), chapter 6. Models for Recurrent Events in Reliability and Survival Analysis, pages 105–123. Kluwer Academic Publishers, 2004.
4. Edsel Pen˜a, Elizabeth Slate, and Juan Ramon Gonzalez. Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference*, 137:1727–1747, 2007.

Compilation of Consumer Price Indices at the national level, some experiences from Africa, Asia and the Caribbean

Rameshwar Srivastava
RPS Associates, The Royal Statistical Society (RSS)

Abstract

The paper will cover the compilation of Consumer Price Indices (CPIs) at the national level, using experiences from six countries in Africa (Ghana, Sierra Leone, The Gambia), Asia (India) and the Caribbean (British Virgin Islands and U.S. Virgin Islands). In most countries, the CPI is considered a key indicator of economic performance, as well as an index for the temporal adjustment of wages and social benefits. The coverage of goods and services in the CPI depends on what type of index it is intended to be, and what contractual payments or benefits it is intended to adjust over time. For instance, in the United States, housing makes up one-third of the CPI, because it includes owner-occupied imputed rents, whereas in Europe, only actual rents are included, and so the relative weight of the housing component is about 6 percent of the basket.

Keywords

Indicator of economic performance, adjustment of wages and social benefits, owner-occupied imputed rents, relative weight.

1. Introduction

The author has the unique experience of having worked and lived in 12 countries. With regards to CPI specific experience, the speaker was the Head of Prices Section in the Ghana Statistical Services; was the United Nations Statistical Training Adviser in Sierra Leone and contributed in the revision of CPI; and more recently was UNDP Consultant in The Gambia and one of the task was to analyse national Households Income and Expenditure Survey (HIES) and compute the weights to revise the CPI.

The author was assigned by the Commonwealth Secretariat, London, UK as Price Indices Economist to the British Virgin Islands (BVI) in the Eastern Caribbean. His task was to analyse the HIES and revise the CPI. Then he moved to US Virgin Islands (USVI) and established the first CPI for USVI.

2. Methodology

The purposes of a CPI is to measure changes in consumer prices. The uses are (i) a general measure of inflation, (ii) indexation by government, (iii) prices, wages and salary adjustment in contracts, (iv) current cost accounting, and (v)

national accounting deflation etc. The index is estimated as a weighted average of elementary aggregate indices, preferably defined in terms of a three-dimensional stratification of items, regions and types of outlets.

The scope of the index has two main dimensions: geographical and the reference population. Many countries started with the CPI for the capital city and later extended to other urban and rural regions and the national combined index. By reference population we mean is it for low-income families, middle-income households, industrial workers, agricultural workers, expatriates, etc.

When CPI uses weights which reflect the composition of the aggregate expenditures of the reference population, this is known as plutocratic. When CPI provides equal importance to all households by averaging expenditure proportions of the reference population instead of summing expenditure amounts, this is known as democratic. Plutocratic weighting is more appropriate if the index is used as a general measure of inflation or for national accounts and other deflations. Sometimes the population data is used to weight different regions (urban/rural) in the calculation of the overall index. Thus the regional population shares when multiplied by regional average expenditure may provide reasonable weights.

A Household Income & Expenditure Survey (HIES) or Household Budget Survey (HBS) is the main source of data for weighting. Some developed countries have annual surveys, some countries conduct every 5th or 10th year. Small changes in weight have little effect on the index, so if an annual survey is not possible due to resources, a five or ten-year frequency should be maintained. The HIES data are also used by a wide range of public and private organizations for economic analysis and planning purposes.

All weights need to be rescaled to sum to 100 or 1000. The selection of items for the consumption basket is based on annual average consumption expenditure proportion to that of the total, if it is above 1/10,000 or 1/1,000 from HIES data. The number of items in the basket are in the range 200¬500 depending upon the size of the country and the reference population.

3. Results

The Table 1 below provides some basic indicators to indicate that the six countries were very diversified in terms of population, gross domestic product (GDP) and inflation.

Table 1: Some Indicators of six countries focused in the paper

Some Indicators	Ghana	Sierra Leone	Gambia	BVI	USVI	India
Population in ('000)	28,800	7,400	2,100	32	107	1,135,000
GDP per head ($) 2017	2,046	510	709	34,200	37,000	1,050
Annual inflation rate (%)	9.0 (Jan-19)	17.1 (Jan-19)	6.4 (Dec-18)	2.1	1.0	2.05 (Jan-19)

Ghana

The data collection and compilation of CPI in Ghana is done by the Ghana Statistical Service (GSS). The Central Bureau of Statistics (CBS), the predecessor of GSS was responsible for it since 1963. The CPI covers the whole country, both urban and rural areas of all ten regions. There are 242 goods and services in the basket. In 2003 on the advice of an IMF Consultant, the data collection was done twice a month instead of once per month before. Forty markets, made up of 9 urban and 31 rural markets have been sampled in the country. For each item, six price quotations are taken in the urban market and three in the rural market. Ghana concluded in a paper presented at UNECE and ILO CPI conference that there is not much to gain from the two readings vis-à-vis cost and quality of data. Collecting prices once per month would be 50% less and supervision will be more effective.

Sierra Leone

Efforts to construct a CPI in Sierra Leone date as far back as 1941, however, a systematic attempt on scientific lines started with the quarterly CPI for Freetown. This CPI that was led by the Ministry of Labour covered low income earning families with the weight base as 1951. Later the Ministry revised the CPI with 1961 as the base. This index covered African families of two or more persons with no earner receiving more than £20.00 (or Leone 40.00) per month. In 1973, the work of compilations of CPI was transferred to the Central Statistics Office (CSO). The revision of CPI for Freetown with base 1978=100 was published in August 1984.

According to Sierra Leone CPI Report 2012 by Statistics Sierra Leone (SSL, CSO renamed) published in September 2013, inflation was being monitored on a monthly and annual basis. The composite CPI was computed as a weighted average of the different center sub-indices. The expenditure weights were obtained from the 2003/04 Sierra Leone Integrated Household Survey (SLIHS) Income and Expenditure Module. The CPI basket covered a total of 400 items; and the CPI was estimated as a weighted aggregate of a fixed basket of

these 400 goods and services popularly consumed in Sierra Leone. The index covered sampled outlets from five urban towns representing the four geographic regions of the country. The CPI reference year was 2007. All prices collected were the prevailing retail market prices from six (6) markets in Freetown, three (3) Markets in Bo Town, three (3) markets in Kenema Town, three markets (3) in Koidu and three (3) markets in Makeni Town for weekly prices, making a total of 18 markets as data collection centers for the CPI exercise in Sierra Leone.

These aggregates were determined using the Classification of Individual Consumption by Purpose (COICOP) for Household Final Consumption Expenditure with 2007 as the new base year. The COICOP classification which is an international standard to disaggregate the CPI has 12 functions excluding the function of "All Items" which is the aggregate index of all functions. Geometric mean formula was used to compute the mean of prices observed for each product by elementary aggregate.

The Gambia

The Gambia Bureau of Statistics (GBoS) collected monthly prices of 250 items. The data collection was spread all over the country comprising of 28 different market centres consisting of 7 in the Greater Banjul, 11 in South Bank and 10 in North Bank spread from Banjul to Basse. The prices were collected once every month for each item, unlike the general practice of collecting prices on a weekly basis.

The CPI had a base year 2004 (2004 = 100) and was the second series which was derived from the household expenditure survey conducted in 2003-2004 with national coverage. The indexes were estimated from the cost of the basket in the current period's prices by multiplying the quantities of items purchased in the base period. A price index was obtained from the ratio of the revalue basket to the total price of the basket in the base period. This was referred to as a Laspeyre's price index.

Table 2 : The weighting composition by COICOP classification.

Item (COICOP	Ghana (2002)	Ghana (2012)	Sierra Leone (1978)	Sierra Leone (2007)	Gambia (2004)	Gambia (2010)
Combined	100	100	100	100	100	100
01. Food and non-alcoholic beverages	44.91	43.9	63.1	41.86	56.33	47.94
02.Alcoholic beverages, tobacco and narcotics	2.23	1.7	3.8	1.71	0.71	0.59
03.Clothing and foot wear	11.29	9.0	5.4	7.34	11.44	9.46
04.Housing, Water, Other Fuels	6.98	8.6	18.2	13.7	3.46	9.41
05.Furnishings, Household Routine Maintenance	7.83	4.7		5.86	5.32	7.42
06.Health	4.33	2.4		11.36	1.24	0.79
07.Transport	6.21	7.3		7.75	4.52	**11.32**
08.Information and communication	0.31	2.7		2.04	4.6	5.97
09.Recreation, sport and culture	3.04	2.6		1.47	1.58	0.97
10.Education services	1.6	3.9		2.88	4.4	3.37
11.Restaurants and accommodation services	8.28	6.1		0.92	0.37	0.01
12.Personal care, social protection and miscellaneous goods and services	2.99	7.1	9.5	3.13	6.02	2.66

British Virgin Islands (BVI)

BVI's first CPI was initiated in May 1972. The May 1972 base was revised with January 1979. The subsequent revision was March 1985 followed by March 1995. The 1994-95 was a first comprehensive Household Income and Expenditure Survey in BVI. The table below indicates 1984/85 and 1994/95 weights and number of items in the basket. BVI economy was growing and we could see the changes in expenditure pattern during a decade as shown in Table 3 below.

Table 3: Change in expenditure weights in one decade.

Major groups	1984/85		1994/95	
	No of items	Weights	No of items	Weights
Food, beverages & tobacco	102	40.0	85	26.8
Housing	11	20.6	11	19.5
Furniture & household	24	6.2	28	11.8
Clothing & footwear	22	11.5	29	15.1
Transportation	12	11.4	11	11.7
Services	9	8.2	16	9.9
Miscellaneous	15	2.1	10	5.2
Total	205	100.0	190	100.0

US Virgin Islands

In January 2001, the Bureau of Economic Research (BER) of the Government of the USVI signed a professional contract with the author as the consultant to establish a CPI for USVI. Prior to that, USVI was using US southern states's CPI, such as Florida. USVI does not have a Statistical Office. BER and the Eastern Caribbean Center (ECC) at the University of VI (UVI) are responsible for surveys, census and statistics. ECC had conducted HIES and published results, which were used to develop the first CPI.

India

The construction of CPI dates back to the period following the First World War. The enactment of Minimum Wages Act, 1948 was an important milestone in the history of the compilation of CPI. This act requires fixation as well as revision of minimum wages from time to time. The compilation of CPI for Urban Non-Manual Employees (UNME) (also referred to as middle class population) has been undertaken by the Central Statistical Organization (CSO) since 1961 whereas the CPI for Agricultural Labourers and CPI for Industrial Workers are compiled by the Labour Bureau since September 1964 and October 1946 respectively. The Labour Bureau also started compiling the CPI for Rural Labourers since November 1995.

Table 4: A Comparative table showing the salient features of CPIs compiled at National Level follows :

	Salient Feature	CPI(UNME)	CPI(IW)	CPI(AL)	CPI(RL)
1.	Source of weights: Family Living Survey (FLS)/ Consumer Expenditure Survey	FLS 1982-83	FLS 1999-2000	1983 (NSS 38th Round)	
2.	Base year of the series	1984-85	2001	1986-87	1986-87
3.	No. of centres/ villages	59 urban centres	78 centres	600 villages	600 villages
4.	No. of markets/ quotations	1022	289	1461	1461
5.	No. of items in the consumption basket	146-345	175-200	260	260
6.	Index released for	59 centres & all-India	78 centres & all-India	20 centres & all-India	20 centres & all-India

The periodicity of all Indices are monthly and time lag is within one month.

4. Discussion and Conclusion

Typically, the number of items in the consumption basket of the CPI varies between 100-400. The sample of markets used was 2-3 in case of small island countries, 20-40 in case of African countries, and hundreds to thousands in case of big countries like India.

Elementary aggregate index were calculated using arithmetic (AM) or geometric mean (GM). It should be noted that AM is always greater than or equal to a GM. The formulae most commonly used in compiling price indexes are the fixed base-weighted Laspeyres.

Some experts have urged that there was no need to conduct expensive HIES frequently to revise the weights of the CPI consumption basket. Scott3 showed in his paper that in case of Ghana by keeping nine group weights equal (worst possible weighting), annual inflation (January 1989 over January 1988) was 26.4% compared with 26.2% with the published weighting.

The Economist publishes Big Mac Index based on one item and also computes Purchasing Power Parity (PPP) based on the index, so why should we worry for collecting prices on hundreds of items.

One CPI does not fit for all. The consumption basket is different for urban and rural consumers; it is not the same for industrial, agricultural and rural workers as in case of India, and hence there are three different CPIs. We can

not use the local CPI to adjust salaries and allowances of international staff and hence International Civil Service Commission (ICSC) compiles indices which are used to adjust salaries and allowances of United Nations and other international agencies staff.

International organizations and national statistical offices in recent years have devoted much attention to the following topics, which needs further discussions:

(i) New sources of data such as electronic scanner data; web-scraped data;
(ii) Need for constructing and publishing more than one CPI that will meet specific requirements;
(iii) Compilation of housing indices, which could be internationally comparable;
(iv) Measuring the online purchases.

One of the disadvantage of online prices is that it covers a small number of retailers and quantities sold is lacking. The advantages of online data is low cost, more frequency, and one could price all products available at retailer web. The advantages of traditional CPI data is more retailers and product categories are covered. The quantities or expenditure weights derived from the Consumer Expenditure Surveys are available. The disadvantage is the higher cost and lower frequency. Both online and offline data should be used and complement each other.

In case of a smaller country one CPI could be enough for example BVI or USVI. Many countries are compiling more than one CPI, as in case of India shown in Table 4 above to meet the specific needs, because one CPI cannot serve all.

References

1. Sierra Leone Government (1984). Technical Report on Revision of Consumer Price Index for Freetown with Base 1978=100 (CSO Freetown, August 1984)
2. Sierra Leone CPI Report 2012 (2013). Published by Statistics Sierra Leone in September 2013
3. Scott, C. (1983). A note on weights for consumer price indexes. INTER-STAT no.8, pp.51-55, March 1993
4. Srivastava, R. (1998). A note on weights for consumer price indexes. INTER-STAT no.17, pp 89¬96, September 1998
5. Turvey, R. et al. (1989). Consumer price indices: an ILO manual, International Labour Office, 1989
6. UNECE/ILO/IMF/OECD/Eurostat/The World Bank (2009). Practical Guide to producing Consumer Price Indices, United Nations, New York and Geneva, 2009

7. Government of India, Ministry of Statistics and Programme Implementation, Central Statistics Office, New Delhi: Manual on Consumer Price Index 2010, December 2010

Issues and challenges in the measurement of cost of living for the adjustment of salaries of expatriate officials of international organizations

Ibrahim S Yansaneh[1]
International Civil Service Commission, United Nations, New York, NY 10017, USA

Abstract
The remuneration of the expatriate officials of most international organizations is often based on the principle of equalization of the purchasing power of salaries paid at various locations of assignment relative to a reference duty station. This principle is actualized by the adjustment of salaries to account for the relative cost of living, measured through cost-of-living surveys. This paper highlights the issues and challenges associated with the measurement of cost of living in the context of salary adjustments of international organizations, with specific focus on the United Nation's Post Adjustment System. The paper also discusses measures taken to address and resolve the specified challenges, as well as related emerging issues and future directions in this field of statistical endeavour.

Keywords
consumer price index, cost-of-living index, expenditure surveys, price surveys

1. Introduction

This paper highlights some issues and challenges associated with the measurement of cost of living for purposes of adjusting the salaries of expatriate officials of international organizations. In the special case of the United Nations (UN), the measurement of cost of living is regulated by the Post Adjustment System (PAS), the organization's salary adjustment system. The output of the measurement is a "Post Adjustment Index" or PAI. The remuneration that is derived from the PAI includes a component reflecting compensation for the cost of living at a duty station relative to New York, the base of the PAS. Thus, the PAI ensures purchasing power parity of salaries of expatriate officials in locations around the world, relative to their counterparts in New York. Challenges are encountered throughout the entire spectrum of activities related to the survey process associated with the cost-of-living measurement, as well as the application of the results to the setting and

[1] yansaneh@un.org
The views expressed in this paper are those of the author and do not necessarily reflect the views of the United Nations

adjustment of salaries. These challenges are highlighted below, along with strategies adopted by the PAS to address them.

2. Measurement challenges

The PAI is designed to reflect the international character of the UN staff population and to be robust enough to be applicable to 200-odd locations with widely varying levels of general economic development, stability of such economic indicators as inflation and local currency exchange rates; availability of goods and services; and differences in the number, composition, expenditure patterns, and turnover of staff. The wide disparity in these characteristics, across the covered locations, presents major measurement challenges for data collection and processing.

i. Data collection

Four main types of data are collected in PAS cost-of-living surveys: (a) price data on about 300 items in PAS's market basket, from retail outlets at the various locations; (b) expenditure data from eligible staff via a web-based survey questionnaire; (c) consumer price indices (CPIs) obtained from national statistics offices, and (d) currency exchange rates. Additionally, where available, market rent data are available are obtained from external sources. The issues with data collection range from the limitations of certain markets, and problems with comparability of retail outlets and products across widely different markets, which make it difficult to ensure like-to-like price comparisons. In the absence of the requisite market research, the impact of these problems is mitigated by more active engagement among stakeholders, the development of detailed and tight item specifications, and an ex post matching of outlets and items at the data processing stage. For electronic and high-technology items, whose specifications change rapidly over time, the real-time price-comparisons (RTPC) approach was developed, with broad specifications to capture enough perfectly matching items at the comparison duty station and the base in real time, thus rendering computationally intensive quality-adjustment methods, such as hedonics, unnecessary.

Expenditure data are collected via self-administered online questionnaires, developed and tested in-house, based on experience acquired in previous survey rounds, but not subjected to rigorous cognitive testing with focus groups. There is therefore the risk of misinterpretation of the survey questions and instructions, leading to reporting errors. This problem is mitigated by strategic engagement with stakeholders, including pre-survey consultations, live demonstrations of survey instruments, and the provision of technical tools to facilitate the administration of the surveys. Data validations and skip patterns are embedded in the web questionnaire to facilitate the survey experience, reduce respondent burden, and minimize reporting errors.

ii. Data processing

The processing of cost-of-living data is conducted through an integrated data management system. The main issue here is that of the treatment of outliers, and missing or erroneous data. The small samples of observations that characterize the datasets from many locations do not comply with the usual assumptions underlying the application of the standard statistical criteria for the identification and exclusion of outliers. To address this problem, the ICSC's standard operating procedures provide for several layers of analysis and quality control checks, in which the analyst is granted discretion to apply experience accumulated from processing price data, as well as to consider various factors, both quantitative and qualitative, to determine whether, or not, a given price should be considered an outlier. In any case, the results of data processing are subjected to a multi-stage peer-reviewed quality control process before they are published.

3. Methodological challenges

The methodological challenges are related to the compilation of a cost-of-living index that reflects a trade-off between pure statistical measurement and compensation policy considerations. They include the determination of expenditure weights needed for compilation of the index, as well as the specification of the index formula and aggregation. Essentially, the PAI is based on a system of bilateral comparisons between each duty station and New York. Yansaneh and Pagan (2011) provides a brief overview of the methodology for compilation of the PAI. More details on the PAI methodology, context, governance, and institutional operational environment, can be found in ICSC (2018). Essentially, the PAI consists of the following five major components:

(i) In-area, excluding housing (IA-H)- relates to living costs incurred locally. It has an internal hierarchical structure, in line with the Classification of Individual Consumption According to Purpose (COICOP). Its index is a weighted geometric average of the cost-of-living relativities of its basic headings, which are estimated as an unweighted geometric average of item price ratios (Jevons), whereas each item ratio is the ratio of unweighted average prices at the duty station relative to New York (Dutot). Its weight is obtained as a residual, after subtracting from the reference net remuneration, the sum of the weights of the other four components; and is pro-rated to all lower level components, down to the basic heading level, in proportion to a set of "common expenditure weights" (average expenditure shares for each component across a subset of locations where about half of eligible staff are assigned, which are considered as representative of the expenditure patterns of the average UN staff).

(ii) Housing (H) - relates to rent and other housing-related costs for the primary dwelling at the location of assignment, including costs for maintenance, utilities, and other housing costs. Its weight is based on staff-reported survey data. Its index is based on market rent data or staff reported data.

(iii) Pension Contribution (PC) - relates to the amount of pension contribution paid by staff, obtained from administrative sources. Its weight is the fixed amount expressed as a percentage of a reference net remuneration (that of an average staff member);

(iv) Medical Insurance (MI)- relates to the amount of insurance premium paid by staff, obtained from administrative sources. Its weight is the average medical insurance premium paid by staff, and the index is the ratio of this average premium at the duty station versus New York;

(v) Out-of-Area (OA)- relates to expenditures outside the country of assignment. Its weight is determined from out-of-area expenditures reported by survey respondents, and the index is estimated as a weighted arithmetic average of US dollarized CPIs of 26 selected countries.

The first methodological challenge is how to derive the overall weight of the PAI and allocate it to its five major components. The accuracy of the weight depends on the quality of expenditure data reported by staff, and this has challenges of its own. Since the surveys are self-administered and voluntary, with no requirement for diaries, there is a risk of both recall and telescoping errors. In any case, due to a variety of reasons, including response burden, recall problems, vested interest, and confidentiality, it is simply not practicable to obtain good-quality information regarding all sources of household income or all household expenditures. In view of all these considerations, and the compensation context of the PAI, for which the measurement target is the average household, the overall weight of the PAI is set equal to the net remuneration of an average staff member, at the time of the survey.

Another methodological challenge is how to aggregate these five components in a way that produces an index that simultaneously satisfies the requirements of a cost-of-living index, and desirable compensation policy objectives, including the overarching requirement that duty stations not be disadvantaged in a uniform or systematic way. These policy requirements effectively eliminate consideration of most of the superlative indices (see Diewart (1976), for details), which, if applied in the spatial context of the PAI, would lead to systematic decreases in PAIs for comparison duty stations, in large part, due to the wide disparity in the weights and sub-indices between the base and most of the comparison locations, as well as a certain lack of substitution of consumption among the highest level components of the PAI.

The aggregation formula for the macro-components of the PAI, as well as the subcomponents of the only two macro-components with an internal

hierarchical structure *(IA-H and H)*, depends on the economic assumptions regarding elasticity of demand or consumer preferences or behavior. At the highest level of aggregation, the five components of the PAI are aggregated arithmetically, as are the seven basic headings comprising the H component are also aggregated arithmetically. However, under the assumption of constant elasticity of substitution among the sub-components of the *IA-H* component, a weighted geometric aggregation is used. In summary, the PAI is defined by the following formula:

$$PAI = RF * COLI,$$

where RF denotes the *"Rebasing Factor"*, the theoretical pay level at the base location (PAI of New York) at the beginning of a survey round, and *COLI* is the cost-of-living index defined by:

$$COLI = w_{IA-H} GEOMEAN(R_i , \omega_i) + w_H I_H + w_{PC} I_{PC} + w_{MI} I_{MI} + w_{OA} I_{OA}$$

That is, COLI is a weighted arithmetic average of the indices of the five macro-components, where the weights and indices are as described above (R_i *and* ω_i are, respectively, the ratios and common expenditure weights corresponding to the *IA-H* basic headings). It can be seen that the PAI has the functional form of a multilateral Walsh formula for the *IA-H* component, modified in two ways: use of expenditure shares derived from a subset of locations participating in the comparison, rather than all locations, and the appending of the four additive components, with all components assigned weights for the duty station, not the base location.

4. Operational rules for salary setting

The PAI is applied for purposes of salary setting and adjustment through a system of operational rules reflecting compensation policy considerations, such as stability and predictability of salaries, as well as the moderation of the disparity of salaries between high- and low-cost locations. Once purchasing power parity (PPP)[2] of salaries is established between the comparison duty station and New York based on the PAI calculated from a cost-of-living survey, a pay index is derived to determine salary. Between surveys, PPP is approximated by updating both indices through different mechanisms: the PAI is adjusted for movements in local currency exchange rates relative to the US$; inflation as measured by CPIs (for the *IA-H* component); and movements in the other component indices (*H, PC, MI*, and *OA*), which are adjusted by other mechanisms. The pay index is adjusted in accordance with compensation policy priorities.

[2] The use of the term PPP here does not conceptually coincide with its use in other contexts such as that of the ICP. For example, the PPP measure includes an OA component, the scope of which includes non-consumption commitments, not usually included in PPPs, but relevant in the context of compensation for UN staff.

The challenge here is that while the updating of the PAI is done with the use of objective data and mechanisms, which maintain approximate PPP of salaries over time, that of the pay index may produce unintended consequences, emanating from various competing compensation policy objectives. For example, an exclusive focus on stability of salaries may have the impact of creating an excessive disparity between the evolutions of the two indices over time, which must then be reconciled upon implementation of the results of the next survey, possibly leading to significant salary reductions, and hence undermining the objective of predictability of salary adjustments. Significant reductions in salary, regardless of whatever mitigation measures may be in place to ensure a gradual transition to lower pay levels, invariably undermine confidence in the methodology underpinning the salary adjustment system and may make it difficult for staff to accept the results generated, thereby failing to achieve a key criterion of data quality.

5. Concluding remarks

This paper highlights some of the challenges that arise in the measurement of cost of living in the context of the PAS, as well as information on how they are dealt with. The challenges stem from the inevitable trade-off between the statistical and economic requirements of a cost-of-living index and desirable compensation policy goals. At the time of writing, a comprehensive review of the PAS is in progress. Both the statistical methodology underpinning the compilation of the PAI and the operational rules for salary setting are being reviewed. The review presents an opportunity to address long-term strategic issues, including the challenges highlighted in this paper.

One of the issues being examined is the feasibility of moving from the current system of bilateral comparisons with New York as the base, using the formula shown in Section III; to a system of multilateral indices that is transitive and based on superlative building blocks, such as a Törnqvist index-based bilateral star system with New York at the center; or a multilateral version, with transitivized parities. As such alternative schemes might result in a star centre different from New York, there is a need to carefully examine the extent to which they are compatible with bedrock UN compensation principles. Also, there is a special focus on the measurement of the housing component, in recognition of its large relative weight. Issues under examination include the trade-off between representativity of market rent data with respect to staff residential patterns, and comparability of the data across locations; as well as the degree of consistency with the evolution of spatial rent indices for a given location relative to New York, with temporal rent indices for that location. Finally, there is the conceptual issue of whether strict adherence to international statistical standards is compatible with the realities of the PAI measurement objective. For instance, the definition of *OA* expenditures,

currently in accord with international statistical standards, is being re-examined in view of the increasing incidence of UN staff residing outside the national borders of their countries of assignment. In view of the PAS compensation context, all proposals will be rigorously tested before being incorporated into the PAS methodology.

References

1. Yansaneh, I.S. and Pagan, R (2011), *Constructing and updating a cost-of-living index for the purpose of ensuring purchasing power parity of the remuneration of expatriate officials*. Proceedings of the 58th Congress of the International Statistical Institute; page 5485-5490
2. *The United Nations Post Adjustment System* (2018): https://icsc.un.org/Resources/COLD/Booklets/PAbooklet.pdf
3. Diewert, W. E (1976), Exact and Superlative Index Numbers, Journal of Econometrics 4, 115-145

Challenges and opportunities for cost of living measurement in the Big Data era – Perspectives from Brazil

Pedro Luis do Nascimento Silva[1], Vladimir Gonçalves Miranda[2]
Instituto Brasileiro de Geografia e Estatística - IBGE, Rio de Janeiro, Brazil

Abstract

The so-called 'big data era' is a time of substantial change to the data landscape, with potential access to a variety of sources that could be useful for enhancing CPI measurement. In Brazil, experiments have begun to collect some prices using web-scraping techniques. IBGE has also started negotiations with tax authorities to obtain access to electronic transaction receipts (e-records) now issued routinely across much of the trade and service sectors. These and other administrative sources have also become potential alternatives to either replace or supplement current data collection operations towards producing the Brazilian CPI. Our paper discusses these challenges and opportunities while reporting on developments in Brazil towards improved CPI measurement.

Keywords

CPI; Big data; Web scraping; e-records

1. Introduction

Cost of Living (COL) measurement has never been easy. The main indicator used for this purpose is the Consumer Price Index (CPI) [ILO, 2004, Stoevska, 2018]. Due to its fixed basket approach, the CPI provides an approximation to variations in COL. Measurement of the CPI presents its own challenges, such as the definition and maintenance of a representative basket of goods and services consumed and corresponding weights, and the regular collection of a representative sample of prices for items in such a basket. In countries like Brazil, the challenges also include proper coverage of the population, as well as the large set of socio-economic situations to be considered when designing price data collection operations.

The so-called 'digital era' promoted profound changes in several aspects of Society. The advent of new kinds of goods and services, the creation of e-commerce and new transaction platforms, and the massive spread and use of mobile communications devices with high-quality internet connections (giving

[1] pedro-luis.silva@ibge.gov.br; [2] vladimir.miranda@ibge.gov.br
The views expressed in this paper are those of the authors and do not necessarily reflect the views of IBGE.

rise to m-commerce) have revolutionized consumer habits. An important by-product of these changes is the large number of new data sources and vast amounts of in-formation on transactions now potentially available to support CPI measurement and estimation.

National Statistics Offices (NSOs) aim to provide timely and accurate statistics to portray the societies that they serve. To remain relevant and achieve their goals they need to keep track of change and to adapt to the evolving data landscape. Typical NSOs also face challenges to produce more while constrained by decreasing budgets, weak legal frameworks regarding their right to access some of the new data sources (private or even governmental), and facing pressures from new private statistics producers that are taking advantage from emerging 'big data' sources [Cavallo et al, 2016].

CPIs are some of the most important economic statistics produced by NSOs around the world [Stoevska, 2018]. The original goal of a CPI was to provide an approximate measure of the cost of living for a given target population. In recent times, CPIs have been used as deflators for Na-tional Accounts, as well as core macroeconomic indicators that Central Banks consider when defining monetary policy [ILO, 2004, Eurostat, 2018].

In this scenario, the adoption of alternative and potentially cheaper data sources for CPI estimation by the NSOs seems attractive, if not mandatory. However, this is far from straightforward. Such data sources are not designed for statistical purposes, and hence their incorporation in CPI estimation might require development and adoption of new methodologies. In some cases, they may even be unusable given CPI quality measurement standards currently in place. In addition, access to such new data sources might not be easy, because NSOs lack adequate legal frame-works supporting rights of access.

The traditional approach for estimation of CPIs relies on the measurement of the price variations for the goods and services in a 'basket' derived at a given point in time [ILO, 2004]. With the advent of the digital economy, new goods and services are created and others disappear at a fast pace. Therefore, maintenance of a representative 'basket' of goods and services consumed, and the collection of prices for any new items presents major demands for CPI compilers. Aiming to circumvent such difficulties, NSOs all around the world are experimenting with the use of alter-native data sources for CPI compilation. The most relevant sources under study include web data, scanner data and administrative registers [Van Loon et al, 2018, Hov et al, 2018, Mendon-ça et al, 2018, Breton et al., 2016, da Silva et al., 2019].

Here we describe some of the main challenges and opportunities related to the adoption of such new data sources for CPI compilation in Brazil. We focus on the main relevant issues that need to be accounted for the use of web data and 'electronic fiscal records' for CPI compilation in Brazil.

2. Overview of the Brazilian CPI

IBGE compiles the official inflation measure for Brazil (called IPCA), as part of a larger set of consumer price indices for different reference periods and target populations (defined according to income profile and geographical coverage) [Miranda et al., 2019, IBGE, 2013]. The set of indices produced is part of a framework called National System of Consumer Price Indices (NSCPI)[3]. In this system, INPC and IPCA are the most important indices. INPC measures inflation for low income families (1 − 5 Brazilian national minimum wages). IPCA has a broader scope and measures inflation for the Brazilian population with income in the 1 − 40 Brazilian national minimum wages range. IPCA is adopted by the Brazilian Central Bank to monitor tar-gets set for national inflation and to define monetary policy.

Both indices are compiled and published monthly for 16 Brazilian states. The state-level indices are then aggregated to compose the 'national' indices [Miranda et al., 2019, IBGE, 2013]. Each of the 16 states have their own basket of goods and services determined by local consumer habits observed in the Household Budget Survey (HBS). Hence, the national basket contains every element that appears in a given local basket with its respective national weight. The basket is derived according to plutocratic criteria and the indices adopt a domestic scope [Miranda et al., 2019]. In a country of continental dimensions such as Brazil, derivation and use of local baskets is an important ingredient to guarantee that the diversity of habits is represented in the NSCPI indicators. Also, the calculation of indicators for different ranges of income is also an important feature of the system, since the Brazilian society is also characterized by large income inequality.

3. Challenges and opportunities for the use of big data in the NSCPI

a. Web data

The exploration of web data on prices for the compilation of CPIs was triggered by the pioneering work of Cavallo and collaborators within the scope of the MIT Billion Prices Project [Cavallo et al, 2016]. This approach is appealing since apparently the prices and descriptions of products announced by web retailers are of good quality, 'free' and can be downloaded massively in high frequencies by means of automated web scrapers. Following this trend, many NSOs are experimenting with the use of web data in the production of their official CPIs [Van Loon et al, 2018, Breton et al., 2016, da Silva et al., 2019].

For a country with the geographic and socio-demographic peculiarities of Brazil, the adoption of web data in replacement of traditional sources should be considered with caution, since some deficiencies of web data can be intensified here. One important methodological caveat is that so far, the HBS

[3] SNIPC, in Portuguese.

only provides weights for products commercialized in brick-and-mortar stores and those need to be used as proxies for products traded via web. For the 2017 - 2018 HBS being finalized, questions on the kind of store (online or offline) where purchases were performed were introduced. With such information more details on the web shopping process will be known and derivation of weights for online shopping might be possible.

Another issue stems from the fact that 'location' of the web site hosts, the location from where the consumer orders the goods, and the delivery address for the goods may all be different, and therefore there would need to be decisions regarding to which region the particular transaction should be allocated, given our approach to build the national CPI from aggregation of regional indices. In addition, though prices offered online often have no distinction according to different regions of the country, different delivery fees apply depending on delivery address, and such fees may also depend on the amount of goods purchased, thus making it difficult to associate a proper price for a single good. Level and evolution of prices at web and brick-and-mortar stores are not necessarily the same, though in a recent study [Cavallo, 2017] found no evidence of price discrimination between online and offline prices for big retailer chains, with an observation of ≈ 5% difference in Brazil.

In Brazil, only about 75% of the households have some form of internet access [IBGE, 2018]. The average velocity of internet connection in Brazil is three times lower than the global aver-age and the access is much lower for rural households. However, internet access is growing fast due to the use of mobile devices and promoting an important increase of e-commerce in Brazil. According to a recent report [ebit, 2019], 7 out of 10 Brazilians owned a mobile phone in 2018, an increase of 7% over 2017. This increase is strongly correlated with the enhancement of the e-commerce observed in 2018, 12% larger in volume of sales than those observed in 2017 [ebit, 2019]. In 2018, 58 million Brazilians (27% of the country's population) performed at least one online purchase. Of these, 10 million performed their first online purchase in 2018, 64% of which by means of a mobile phone. The growth in the volume of sales of the m-commerce in 2018 over 2017 was 41%, amounting to ≈ 42% of the volume of e-commerce sales in 2018 [ebit, 2019]. The number of online orders was 123 million in 2018 (increase of 11% over 2017), however this led to a modest average of ≈ 2 purchases a year for the online buyers in 2018, mostly concentrated in sectors like make-up, and clothing and accessories which respond to ≈ 40% of the orders.

Given the peculiarities of the e-commerce in Brazil discussed above, the adoption of web data for the NSCPI should proceed with caution. IBGE is experimenting a parsimonious introduction of web data to improve CPI compilation in combination with traditional sources. The first approach that IBGE is experimenting with involves web scraping for prices of products where

manual collection on the web is already the norm, such as for airline ticket prices. The other experiment under way relies on the use of web data to expand the amount of information collected for improvement of the CPIs compiled. A pioneer approach in this area is the collection of web data to support implementation of hedonic quality adjustment methods in the NSCPI. The use of web data to expand the number of elements in the CPI basket and the compilation of an experimental web-only CPI are also under consideration for a future stage of this project.

We briefly discuss only the first case here. For more details we refer to [da Silva et al., 2019]. A pilot project on the use of web data is ongoing for the collection of prices of airfares. Since nowadays airfares are mostly traded online, the traditional collection consists in having price collectors visit the web sites of the main airline companies on pre-established dates, and then recording the prices observed for pre-determined routes. This process is time consuming and subject to human error: data collection takes about 2 – 4 hours per week, per area, and extra time is demanded for the head office team to check if the prices collected are consistent. The pilot project aims for the replacement of the manual collection by automatic web scraping routines. A home-maid web scraper was developed to automate the current collection process: scraping once a week for the same routes on the same airlines.

The pilot is running for almost a year now and the results are promising, as summarized in Figure 1, that compares an experimental weekly index built from the manual and automatic price data collections [da Silva et al., 2019]. Figure 1 shows good agreement between the manual and automatic series. However, before implementation of the automated solution in the monthly production process of the CPIs, some extra care is needed. The main problems observed in the pilot were: dealing with anti-robot policies - one airline has blocked the web scraper for a while; website architectures may change without warning, thus requiring adjustments in the web scraper to ensure that the correct information is extracted; and network instabilities during some special sales events have disturbed the robot's ability to retrieve prices [da Silva et al., 2019].

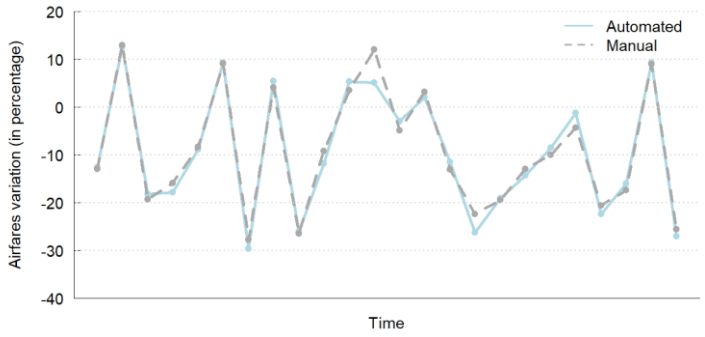


Figure 1: Experimental weekly price variation derived for airfares, with series derived for data collected by manual and automatic tools in the same dates. Reproduced with permission from [da Silva et al., 2019].

Although the access issues and the requirements for maintenance of the web scraper are important limitations to have in mind, a series of benefits can be derived by the adoption of the automatic process. The time and effort required for data capture are substantially reduced, hence freeing time the price collectors can allocate to other tasks. The automatic process enables creation of high-quality data backups, since snapshots are automatically extracted from the screens where the prices were collected, enabling verification of the information collected. Finally, the automatic process also enables increasing the number of price quotes for routes in a straightforward and cheap manner, which in turn may lead to more accurate estimates of the CPI.

In order to guarantee that the IBGE robots are not going to be blocked, IBGE has initiated a conversation with the airline companies to negotiate access agreements. This process is also fundamental to create a relationship between IBGE and the companies, which may help to re-solve technical issues that might arise, for example, when their websites are expected to change.

b. Transactions e-records

Many NSOs are devoting efforts towards the use of scanner data. Scanner data from private retailers often contain very detailed information on the products traded such as: prices and quantities of products sold, the time of the purchase, etc. This source is more attractive than web scraping, because prices recorded correspond to goods and services actually traded. However, access to scanner data is much more difficult in the absence of legislation that grants the NSO legal rights of access to such data, as is the case in Brazil at present. Therefore, individual negotiations with each retailer are needed, and in some cases, may take years until an agreement is reached, which may or may not require the NSO to pay for the data to be provided [ILO, 2004].

IBGE still has no access to scanner data. However, a peculiar feature of the fiscal system of Bra-zil points for another source that can be even more attractive than scanner data. The fiscal legislation in Brazil requires that every legal sales transaction performed by a store (brick and mortar or online) generates an individual electronic record with information that includes: description and code of each product sold; the price paid per unit of each product sold; the total price paid by product; the total amount paid; the date and time of the sale; the media used for payment (cash, credit or debit card); and detailed information on the establishment, including its unique fiscal identification code (the CNPJ). In some cases, information on the buyer is also available, especially for transactions where buyers are enterprises.

The information contained in the e-records is as useful as those contained in scanner data sets and would therefore provide one of the best data sources for CPI compilation. Such information is already provided to government fiscal entities, covering all the legal transactions performed by all retailers in the country. Currently the e-records are collected by the states' tax authorities and consolidated for the whole country by the Federal Revenue Service.

The access to the information contained in the data base of the e-records could provide many improvements for the NSCPI. It should be possible to expand the coverage of the NSCPI to finer geographical levels, and to compile CPIs for different target populations based on consumer baskets constructed for different profiles (finer level of income ranges, for instance). The use of more powerful superlative indices would also be possible since information on price and quantities are available in such data sets. The e-records can potentially provide information or the derivation of more frequent updates of the CPI baskets and the weights of the basket components. The number of elements in the basket could also be increased, since the "collection costs" which scales with the number of items in the baskets and constrains the number of elements to be priced are not present here, though processing and storage costs of the e-records should be considered. The indices could, in principle, be calculated in higher frequencies since the information is updated more frequently.

Important challenges are also posed for the use of the e-records for CPI compilation. The first is the access to such data. The e-records are protected by fiscal confidentiality laws, and so far, IBGE has not managed to get legal rights to access to the transactions e-records. To try to circumvent this restriction initial negotiations were held with the tax authorities for the access but so far have not succeeded.

Another important point to consider is that product descriptions and classifications in the e-records are probably not in straight correspondence with the classification required by the CPI. Such analysis and the integration of the e-records into a classification and unified product description system as

needed for the CPI constitutes a non-trivial task. The development of an automatic classification system based on machine learning techniques will probably be required, such as the ones developed for use with scanner data by some countries [ILO, 2004].

As observed with scanner data sets, high product churn and chain drift issues [ILO, 2004] might also affect price indices compiled using the e-records if traditional CPI formulas based on the matched model method are adopted. To circumvent this problem, unit value formulations or the use of multilateral methods should be considered. Such methodological changes would have impact on the computational routines used to compile the CPI and would require acquisition of knowledge about such methods by the CPI staff.

Caution should also be devoted to the separation of the transactions contained in the e-records between those made by families and those made by businesses, since the target of the CPIs are family's transactions. Some data treatment would be required to restrict the transactions considered to those most likely relating to transactions made by families.

Finally, the IT infrastructure needed to store and process such huge data volumes would have to be very robust and scalable. The fiscal authority has the storage and processing capabilities to satisfy their own needs, but these capabilities cannot be shared with the IBGE. One potential approach to address this issue would be for IBGE to specify sampling procedures to be applied to the e-records database on a regular basis, and to receive only the specified sample data to use in the CPI compilation. Such an approach would, however, depend strongly on gaining some initial access to the e-records in order to be able to develop proper sampling procedures.

4. Conclusion

This paper analysed the challenges and opportunities for the adoption of alternative big data sources for CPI compilation in Brazil. The focus was on the analysis of web data and tax records. The main constraint for the use of web data in Brazil and the reasoning for the initial implementation of web information at the NSCPI were discussed. The pioneering project for web scraping for airfares was briefly discussed showing that the use of automatic collection tools is promising, though care need to be taken for their implementation in the monthly routines of the CPIs.

The tax e-records was also considered, with an assessment of their potential and the main technical problems that need to be overcome for adoption and implementation at the NSCPI. The impacts that such a source would provide for the expansion of the NSCPI coverage were also considered.

References

1. Breton, R. et al. (2016). Research indices using web scraped data. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices.
2. Cavallo, A. (2017). Are online and offline prices similar? evidence from large multi-channel retailers. American Economic Review, 107.
3. Cavallo, A. et al (2016). The billion prices project: Using online prices for measurement and research. Journal of Economic Perspectives, 30:151–78.
4. da Silva, L. T. et al. (2019). Studies of new data sources and techniques to improve CPI compilation in Brazil. Paper submitted to the 16th meeting of the Ottawa Group, Rio de Janeiro, Brazil.
5. ebit. (2019). 39th webshoppers report.
6. Eurostat (2018). Harmonized index of consumer prices (HICP): Methodological manual.
7. Hov, K. N. et al. (2018). Using scanner data for sports equipment. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices.
8. IBGE (2013). Sistema nacional de índices de preços ao consumidor: métodos de cálculo. Série relatórios metodológicos, 14. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv65477.pdf.
9. IBGE (2018). Continuous national household sample survey: results published for the 4th quarter of 2017. URL https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101631.
10. ILO (2004). Consumer price index manual: Theory and practice. Revised chapters available for download at: https://www.imf.org/en/Data/Statistics/cpi-manual.
11. Mendonça, V. et al. (2018). Exploring new administrative data sources for the development of the Consumer Price Index: The Portuguese experience with actual rentals for housing. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices.
12. Miranda, V. G. et al. (2019). Consumer price indices at IBGE: 40 years and counting. Paper submitted to the 16th meeting of the Ottawa Group, Rio de Janeiro, Brazil.
13. Stoevska, V. (2018). Official consumer price indices - historical perspective. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices.
14. Van Loon, K. et al. (2018). Integrating big data in the Belgian CPI. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices.

Modeling and optimization of the SPRT control chart for individual observations

Chenglong Li
School of Management, Northwestern Polytechnical University, Xi'an-710072, China

Abstract

In many practical situations, it may not be feasible to take samples larger than one for various reasons, so control charts must be based on individual observations (= 1) rather than larger samples of > 1. The sequential probability ratio test (SPRT) chart is particularly suitable for monitoring individual observations. The existing research conducts the design or optimization algorithm for a SPRT chart from a statistical point of view only. This paper uses a Markov chain approach and proposes an economic model for designing the SPRT chart. The results based on an extensive performance comparison, indicate that under various scenarios the SPRT chart uniformly outperforms some other competing charts on a cost basis.

Keywords

Control chart; Economic modeling; Individual observation; Optimization; SPRT; Markov chain.

1. Introduction

The use of most control charts requires sample sizes larger than one (known as rational subgroups). However, in some applications, only a single item can be sampled at each point in time. This usually happens when sampling and testing are time-consuming, expensive and even destructive. Many traditional (Shewhart-type) control charts perform poorly in such situations. Instead, memory-type control charts, say, the CUSUM chart or the EWMA chart, are often recommended in order to make efficient use of all the information in the sequence of individual data points.

Actually the SPRT chart is also an alternative choice for monitoring processes with individual observations. The SPRT chart is normally defined as a sequence of SPRT's, each separated by a fixed sampling interval. By inspecting sequentially one observation at a time, the SPRT chart allows the sampling rate used at each sampling point or SPRT to vary based on the data observed at the current SPRT, with the possibility of a decision about the process after each observation.

Stoumbos and Reynolds (1996, 1997) made the early efforts on the research of SPRT charts. In recent years, a number of research has been devoted to the development and improvement of SPRT-based schemes; see

Li *et al.* (2009), Ou *et al.* (2011), Haridy *et al.* (2013), Ou *et al.* (2015), among others. These studies have greatly promoted the theoretical development and practical application of SPRT charts.

However, all of them carry out the design of a SPRT chart from a statistical point of view only. It is well known that purely statistical model is not certainly optimal from the economic point of view. To bridge the research gap, in this paper, we employ a Markov chain approach and render an economic model for designing the SPRT chart (for a long-run production).

The remainder of the paper is structured as follows. Section 2 introduces the operation of a SPRT chart. Section 3 is devoted to the model development for the SPRT chart. The performance comparison is conducted with several competing control charts in Section 4. Finally, Section 5 ends with conclusions.

2. SPRT control scheme

A SPRT control scheme has four parameters: the sampling interval (m), the lower limit (g), and the upper limit (h), the reference value (γ). A sample point or a SPRT is taken at the end of each fixed sampling interval of m items, which are produced in succession from the production line. Within a sample, when the i th observation xi has been taken, it is used to update the test statistic ω . For an upper one-sided SPRT chart (where $-\infty <$ $< h < \infty$),

$$\omega i = \omega i-1 + zi - \gamma \tag{1}$$

$$z_i = \frac{x_i-\mu_0}{\sigma_0} \tag{2}$$

where $\omega 0 = 0$ denotes the starting value of the control statistic; $\mu 0$ and $\sigma 0$ are the incontrol (IC) values of the mean and standard deviation of the quality characteristic. The lower one-sided SPRT chart for detecting a decrease in the mean operates in a similar manner. In this work, we focus on the upper one-sided SPRT chart for detecting an increase in the mean. Inside a sample point or SPRT, if $\omega i > h$, the process is considered to be out-of-control (OC) and an alert signal is triggered. If $g \leq \omega i \leq h$, sampling is continued sequentially. Finally, if $\omega i < g$, the process is considered to be IC and the current SPRT is terminated.

3. Model development

3.1 Probabilistic model

Assume that the process failure mechanism can be described by a geometric distribution with the parameter p and the assignable cause results in a shift in the mean from $\mu 0$ to $\mu_1 = \mu_0 + \delta\sigma_0$ ($\delta > 0$) but with no change in the standard deviation σ_0. Once the process is operating in the OC state, it remains so until stopped for investigation following a signal on the control chart. Then, the process resumes in the IC state after the negative effect is

removed. At any point in time, the process operates in either IC or OC state, which is only partially observable through 3 sampling or inspection. Let $\boldsymbol{\theta_r} = [\theta_{ic;}\ \theta_{oc;r}]$ be the state probability vector of the process at the beginning of the r th monitoring cycle. Then, the transition probability matrix $\boldsymbol{T}$ has the basic form of

$$\boldsymbol{T} = \begin{bmatrix} \mathrm{t}_{ic\to ic} & \mathrm{t}_{ic\to oc} \\ \mathrm{t}_{\mathrm{oc}\to \mathrm{ic}} & \mathrm{t}_{\mathrm{oc}\to oc} \end{bmatrix} \quad (3)$$

where $t_{A\to B}$ refers to the probability of the process status at the beginning of the current monitoring cycle to be in the state of B conditioned on the previous one in the state of A. The stationary distribution of a Markov chain is able to be derived if we obtain the elements of $\boldsymbol{T}$.

In order to characterize the operation of a SPRT chart within a sample, suppose, the interval between the lower limit g and upper limit h is portioned into s subintervals of equal length. The width of each subinterval is given by $\Delta = (h - g)/s$. Thus, the test statistic ω_i experiences s different transient states, i.e., $W_1 = [g, g + \Delta]$, $W2 = [g + \Delta, g + 2\Delta]$, ,, $Wj = [g + (j - 1)\Delta, g + j\Delta]$, ,, $Ws = [g + (s - 1)\Delta, h]$, within the control limits before going to an "IC state" ($\omega_i < g$) or an "OC state" ($\omega_i > h$). Assume that the probability density within the j^{th} subinterval $[g + (j - 1)\Delta, g + j\Delta]$ for $j = 1,2, \ldots s$, is concentrated as a probability mass at the center, $Oj = g + (j - 0.5)\Delta$. We further define the two absorbing states, $W_0 = [-\infty, g]$ and $W_{s+1} = [h, +\infty]$, such that there is a total of $s + 2$ subintervals. Let us adopt a simplified notation scheme to label the states by transforming the real value ωi to an integer between 0 and $s + 1$ in the following manner, for $\omega_i \in W_j \longrightarrow w_i = j$.

Thus, the SPRT inspection system can be modelled as a Markov chain and the set of states space is described by a pair of integers (u, w). A variable, u, relates to the status of the process: $u = 0$ for a process is IC, and $u = 1$ for an OC process. The index $w \in \{0,1, \ldots j, \ldots s, s + 1\}$ indicates which subinterval the test statistic ωi is falling into, as defined before. In this way, the Markov chain (u, w) has $2 \times (s + 2)$ states. Note that, a SPRT inspection stops only when the process is in one of the four states: (0,0), (0, $s + 1$), (1,0), (1, $s + 1$). We let $q^{ic}_{(0,0)}$, $q^{ic}_{(0,s+1)}$, $q^{ic}_{(1,0)}$ and $q^{ic}_{(1,s+1)}$, respectively, denote the corresponding expected probabilities when a monitoring cycle originates at IC state and ends at the four states, and likewise, we have $q^{ic}_{(0,0)}$, $q^{ic}_{(0,s+1)}$, $q^{ic}_{(1,0)}$ and $q^{ic}_{(1,s+1)}$. These probabilities are relevant to and are used for computing the transition probability matrix $\boldsymbol{T}$ as shown in Eq. (3), after arrangement resulting in

$$\boldsymbol{T} = \begin{bmatrix} q^{ic}_{(0,0)} + q^{ic}_{(0,s+1)} + q^{ic}_{(1,s+1)} & q^{ic}_{(1,0)} \\ q^{oc}_{(0,0)} + q^{oc}_{(0,s+1)} + q^{oc}_{(1,s+1)} & q^{oc}_{(1,0)} \end{bmatrix} \quad (4)$$

Applying the property of Markov chain, we can obtain these probabilities. The technical details are omitted here for brevity. Then, the stationary distribution $\boldsymbol{\theta} *$ can be derived as follows,

$$\boldsymbol{\Theta}^* = \begin{bmatrix} \frac{1-q_{(1,0)}^{oc}}{1-q_{(1,0)}^{oc}+q_{(1,0)}^{ic}} & \frac{q_{(1,0)}^{ic}}{1-q_{(1,0)}^{oc}+q_{(1,0)}^{ic}} \end{bmatrix} \quad (5)$$

3.2 Cost model

Several types of costs need to be considered: fixed cost of sampling c_1, variable cost of sampling c_2, cost of scrapping inspected items (destructive inspection) c_3, cost of releasing non-conforming items c_4, cost of false alarms c_5, and cost of finding the assignable cause c_6. The expected total cost incurred in a monitoring cycle under IC set-up:

$$C_{ic} = c_1 + (c_2 + c_3) \cdot ASN_{ic} + (m_{ic}\varphi_{ic} + m_{oc}\varphi_{oc})_4 + q_{(0,s+1)}^{ic}c_5 + q_{(1,s+1)}^{ic}c_6 \quad (6)$$

and similarly, the expected total cost incurred in a monitoring cycle under OC set-up:

$$C_{oc} = c_1 + (c_2 + c_3) \cdot ASN_{oc} + m\varphi_{oc}c_4 + q_{(0,s+1)}^{oc}c_5 + q_{(1,s+1)}^{oc}c_6 \quad (7)$$

where $m_{ic} = m \cdot (1 - p)^m + \sum_{i=1}^{m}(m-i) \cdot p(1-p)^{m-i}$ and $m_{oc} = \sum_{i=1}^{m} i \cdot p\,(1-p)^{m-i}$ represent the average number of items produced under IC and OC, respectively, within the fixed sampling interval when a monitoring cycle originates at IC state; φ_{ic} and φ_{oc} denote the corresponding defective rate under IC and OC, respectively. Therefore, the expected cost per item can be expressed as

$$EC = \frac{\Theta*[C_{ic}\ C_{oc}]^{\mathrm{T}}}{\Theta*[(ASN_{ic}+m)\ (ASN_{oc}+m)]^{\mathrm{T}}} \quad (8)$$

3.3 Optimization

As the full expression of Eq. (8) is mathematically complicated, closed-form analytical results are difficult to derive. On this occasion, numerical search methods are recommended. Previous studies of SPRT chart reveal that γ should be optimally chosen to be $\delta/2$ (see, e.g., Stoumbos and Reynolds, 1997). The optimal values of the remaining three parameters, i.e., m^0, g^0 *and* h^0 have to be derived based on the optimization model formulated as

$$(m^0, g^0, h^0) = \arg\min_{(m,g,h)} EC \quad (9)$$

4. Numerical investigation

To demonstrate the economic superiority of the SPRT chart, we compare it with two economically designed individuals CUSUM charts—the fixed-parameter (Fp) CUSUM chart and the variable sampling interval (VSI) CUSUM chart. The reference values of the two CUSUM charts are optimally set to $\delta/2$ as well. The numerical investigation was carried out based on a comprehensive set of 72 process scenarios.

Figure 1 shows the advantage of the SPRT chart over the CUSUM charts. We find that the VSI CUSUM chart has averaged 9.7 percent increase from the optimal expected cost of the SPRT chart over the 72 process scenarios for $N(0,1)$, averaged 8.0 percent increase for $t(4)$, and averaged 16.5 percent increase for $G(4, 1)$. The economic performance of the Fp CUSUM chart is even

worse in many cases. From the above, the SPRT chart is highly competitive, compared to those classical recommendations for individual observations.

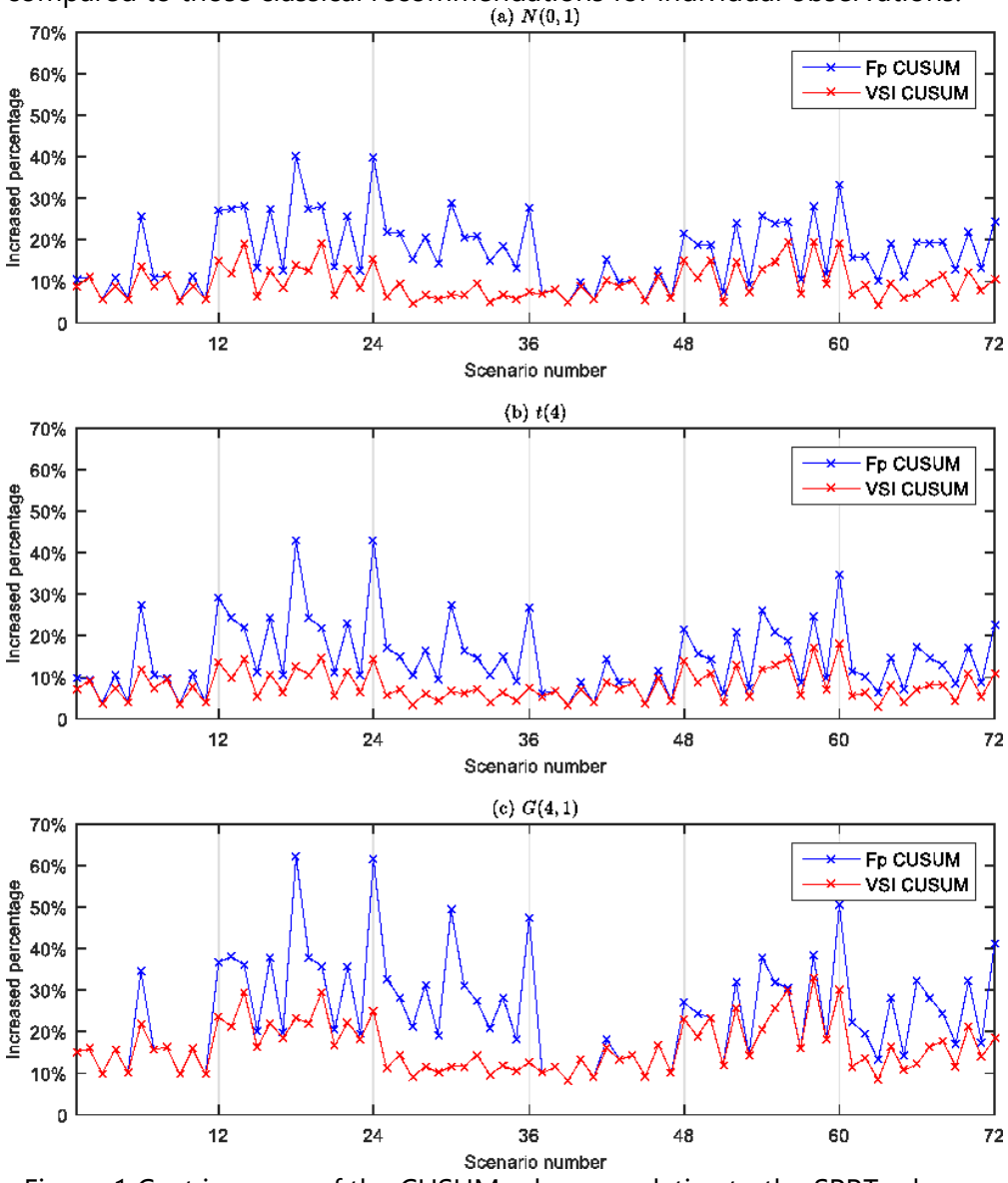


Figure 1 Cost increase of the CUSUM schemes relative to the SPRT scheme

5. Conclusion

In this paper, we construct an economic model for optimally designing the SPRT chart. The results show that the SPRT chart maintains preponderant economic superiority under various process scenarios and is supposed to be attractive in practice.

References

1. Haridy, S., Wu, Z., Lee, K. M. & Bhuiyan, N. 2013. Optimal average sample number of the SPRT chart for monitoring fraction nonconforming. European Journal of Operational Research, 229, 411-421.
2. Li, Y., Pu, X. & Tsung, F. 2009. Adaptive charting schemes based on double sequential probability ratio tests. 6 Quality and Reliability Engineering International, 25, 21-39.
3. Ou, Y., Wu, Z., Khoo, M. B. C. & Chen, N. 2015. A rational sequential probability ratio test control chart for monitoring process shifts in mean and variance. Journal of Statistical Computation and Simulation, 85, 1765-1781.
4. Ou, Y., Wu, Z., Yu, F.-J. & Shamsuzzaman, M. 2011. An SPRT control chart with variable sampling intervals. The International Journal of Advanced Manufacturing Technology, 56, 1149-1158.
5. Stoumbos, Z. G. & Reynolds, M. R., Jr. 1996. Control charts applying a general sequential test at each sampling point. Sequential Analysis, 15, 159-183.
6. Stoumbos, Z. G. & Reynolds, M. R., Jr. 1997. Control charts applying a sequential test at fixed sampling intervals. Journal of Quality Technology, 29, 21-40.

A new nonparametric homogeneously weighted moving average control chart for monitoring the process location

Muhammad Abid[1], Hafiz Zafar Nazir[2], Muhammad Riaz[3]
[1]Government College University, Faisalabad, Pakistan
[2]University of Sargodha, Pakistan
[3]King Fahad University of Petroleum and Minerals, Saudi Arabia

Abstract

The main advantage of homogenously weighted moving average (HWMA) control chart in comparison to the exponentially weighted moving average (EWMA) control chart is that the plotting statistic of HWMA chart assigns specific weight to the current observations and the remaining weights are equally distributed between the previous observations. This study suggests a new non-parametric HWMA chart using an arcsine transformation for monitoring the process target. To compute the average run lengths profile a Monte Carlo simulations are used. The dominance of the proposed chart is constructed against its competitors such as nonparametric EWMA sign, EWMA arcsine, CUSUM sign and mixed EWMA-CUSUM arcsine charts. The study found that the proposed chart performs efficiently for detecting small and as well as larger shifts in process target.

1. Introduction

Statistical process monitoring (SPM) consists of several tools which are used to monitor, control and improve the quality of a product. From these SPM tools control chart is an important tool to simplify process control. Control charts are categorized into memoryless and memory control charts. Memoryless control charts only utilize the up-to-date information in the statistic but the memory control charts develop on the basis of past information along with up-to-date information. Shewhart[1] control chart belongs to the category of memoryless charts and performs efficiently to detect large shifts in process parameters. The cumulative sum (CUSUM) by Page[2] and the exponentially weighted moving average (EWMA) by Roberts[3] charts belongs to the type of memory charts and useful to detect smaller shifts in the process location or/and variation.

In the literature, several kinds of modifications of the charting strategies have been proposed under various setups. Lucas and Crosier[4] and Lucas and Saccucci[5] applied the fast initial response (FIR) on CUSUM and EWMA charts, respectively. Abbas et al.[6] and Zaman et al.[7] introduced mixed EWMA-CUSUM and mixed CUSUM-EWMA charts, respectively. These charts showed quicker small shifts detection ability against the EWMA and CUSUM charts. Abid et

al.[8-9] introduced the nonparametric EWMA and nonparametric CUSUM charts, respectively, by employing the ranked set sampling (RSS) procedure.

The main disadvantage associated with the plotting statistic of an EWMA chart is that it gives larger weights to the current observations and lesser weights to the previous observations. To overcome this deficiency, Abbas[10] proposed a new chart named as homogeneously weighted moving average (HWMA) chart. The plotting statistic of HWMA chart give a particular weight to the current sample and the remaining weights are equally distributed between the previous samples. To the best of our knowledge, there is no work done in the literature on nonparametric HWMA chart. In this study, we develop a nonparametric HWMA arcsine chart to fill this gap in the literature.

2. Methodology

2.1 Existing non-parametric control charts

In this section, structures of some existing non-parametric control charts such as: $\mathrm{EWMA_{NS}}$, $\mathrm{EWMA_{NAS}}$, $\mathrm{CUSUM_{NS}}$ and $\mathrm{MEC_{NAS}}$ are presented.

2.2 Non-parametric sign and arcsine EWMA control charts

Let Y_j, $j = 1, 2, \dots, n$ be a randon sample of size n obtained from a process and has a process mean μ. Outline $X_j = Y_j - \mu$ and then $p = P(X_j > 0)$ 'process proportion'. For in-control process we assume, $p = 0.5$. The sign statistic is defined as:

$$S^+ = \sum_{j=1}^{n} I_j, \tag{1}$$

where

$$I_j = \begin{cases} 1 & \text{if } X_j > 0 \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, 2, \dots, n.$$

Then the distribution of S^+ would follow binomial distribution with parameters $(n, 0.5)$ for an in-control process. Therefore $E(S^+) = n/2$ and $Var(S^+) = n/4$. Based on (1) the statistic of non-parametric sign EWMA ($\mathrm{EWMA_{NS}}$) chart proposed by Yang et al.[14] is defined as:

$$EWMA_{S_j^+} = \lambda S_j^+ + (1 - \lambda) EWMA_{S_{j-1}^+}, \tag{2}$$

where $0 < \lambda \leq 1$ and S_j^+ indicates the value of S^+ in the j^{th} sample. Initially the value of $EWMA_{S_0^+}$ is set equal to the mean of S^+ i.e., $EWMA_{S_0^+} = n/2$. The mean and asymptotic variance of (2) are defined as: $E(EWMA_{S_j^+}) = n/2$ and $Var(EWMA_{S_j^+}) = \frac{\lambda}{2-\lambda}(\frac{n}{4})$.

The control limits of the $\mathrm{EWMA_{NS}}$ chart based on the asymptotic standard deviation of the statistic given in (2) are:

$$UCL_{EWMA_{S_i^+}} = \frac{n}{2} + C\sqrt{\frac{\lambda}{2-\lambda}\left(\frac{n}{4}\right)}, \quad CL_{EWMA_{S_i^+}} = \frac{n}{2}, \quad LCL_{EWMA_{S_i^+}} = \frac{n}{2} - C\sqrt{\frac{\lambda}{2-\lambda}\left(\frac{n}{4}\right)}.$$

where λ and C are selected to obtain the desire in-control average run length (ARL_o).

Yang et al.[14] observed that the in-control ARL results of $\mathrm{EWMA_{NS}}$ chart is not consistent when p differs from 0.5. So, to overcome this deficiency, Yang et al.[11] also developed the non-parametric arcsine EWMA ($\mathrm{EWMA_{NAS}}$) chart using the arcsine transformation i.e., $T = sin^{-1}\left(\sqrt{\frac{S^+}{n}}\right)$. The distribution of T follows the normal distribution with $\text{mean} = sin^{-1}\left(\sqrt{p}\right)$ and $\text{variance} = \left(\frac{1}{4n}\right)$ (cf. Yang et al.[11]).

The plotting statistic of $\mathrm{EWMA_{NAS}}$ chart is written as:

$$EWMA_{T_i} = \lambda T_i + (1-\lambda) EWMA_{T_i-1} \tag{3}$$

Firstly, $EWMA_{T_0} = sin^{-1}\left(\sqrt{p}\right)$ and the $E(EWMA_{T_j}) = sin^{-1}\left(\sqrt{p}\right)$ and the $Var(EWMA_{T_j}) = \frac{\lambda}{2-\lambda}\left(\frac{1}{4n}\right)$, respectively (cf. Yang et al.[11]).

So, the control limits of the $\mathrm{EWMA_{NAS}}$ chart are:

$$UCL_{EWMA_T} = sin^{-1}\left(\sqrt{p}\right) + L\sqrt{\frac{\lambda}{2-\lambda}\left(\frac{1}{4n}\right)},$$

$$CL_{EWMA_T} = sin^{-1}\left(\sqrt{p}\right),$$

$$LCL_{EWMA_T} = sin^{-1}\left(\sqrt{p}\right) - L\sqrt{\frac{\lambda}{2-\lambda}\left(\frac{1}{4n}\right)}.$$

2.3 Non-parametric sign CUSUM control chart

Using the statistic given in (1) Yang and Cheng[12] developed the two plotting statistic i.e., C_t^+ and C_t^- of the non-parametric CUSUM sign ($\mathrm{CUSUM_{NS}}$) chart as follows:

$$\left.\begin{aligned} C_t^+ &= max\left(0, C_{t-1}^+ + S_t^+ - (np_0 + k)\right) \\ C_t^- &= min(0, C_{t-1}^- - (np_0 - k) + S_t^+) \end{aligned}\right\} \tag{4}$$

where $t = 1, 2, \ldots$ and initially, $C_t^+ = 0$ and $C_t^- = 0$. The statistics given in (6) are plotted against their control limits h and $-h$, respectively. The process is considered to be out-of-control if $C_t^+ \geq h$ or $C_t^- \leq -h$, else, it is in-control.

2.4 Non-parametric mixed control chart

Abbasi et al.[13] proposed a mixed nonparametric EWMA and CUSUM ($\mathrm{MEC_{NAS}}$) control charts using the arcsine transformation. The statistic of $\mathrm{MEC_{NAS}}$ control chart is given below:

$$\left.\begin{aligned} MEC_t^+ &= \max[0, \left(Y_j - p_0\right) - k_j + MEC_{t-1}^+ \\ MEC_{\bar{t}} &= \max[0, -\left(Y_j - p_0\right) - a_i + + MEC_{i-1}^-] \end{aligned}\right\} \tag{9}$$

2.5 Proposed arcsine HWMA control chart

Hunter[14] observed that the plotting statistic of EWMA chart proposed by Page[2] is given more weights to the current observations and fewer weights to the previous observations. To overcome this deficiency, Abbas[32] proposed a HWMA chart and its statistic is defined as:

$$H_j = \theta \overline{X}_j + (1-\theta)\overline{\overline{X}}_{j-1} \tag{5}$$

where $\overline{X}_j$ represent the sample mean for jth sample, $\overline{\overline{X}}_{j-1}$ is the mean of the means of previous $j-1$ samples and θ is the sensitivity parameter lies between zero and one i.e., $0 < \theta \leq 1$. The $\overline{\overline{X}}_{j-1}$ is also defined as:

$$\overline{\overline{X}}_{j-1} = \frac{\sum_{l=1}^{j-1} \overline{X}_l}{j-1} \tag{6}$$

The value of $\overline{X}_0$ equal to the target mean μ_0. The statistic given in (1) can also be rewritten as (cf. Abbas[32]):

$$H_j = \theta \overline{X}_j + \left[\left(\frac{1-\theta}{j-1} \right) \left\{ \overline{X}_{j-1} + \overline{X}_{j-2} + \ldots + \overline{X}_1 \right\} \right] \tag{7}$$

The statistic represented in (5) is to give weight θ to the current sample and to adopt homogeneity in weight rest of all the samples are given weights $1-\theta$ (cf. Abbas[32]). Abbas[32] defined the control limits of HWMA chart as follows:

$$\left.\begin{aligned} LCL_j &= \begin{cases} \mu_0 - \mathrm{K}\sqrt{\frac{\theta^2}{n}\sigma_0^2}, & \text{if } j = 1 \\ \mu_0 - \mathrm{K}\sqrt{\frac{\sigma_0^2}{n}\left(\theta^2 + \frac{(1-\theta)^2}{(j-1)}\right)} & \text{if } j > 1 \end{cases} \\ CL &= \mu_0 \\ UCL_j &= \begin{cases} \mu_0 + \mathrm{K}\sqrt{\frac{\theta^2}{n}\sigma_0^2}, & \text{if } j = 1 \\ \mu_0 + \mathrm{K}\sqrt{\frac{\sigma_0^2}{n}\left(\theta^2 + \frac{(1-\theta)^2}{(j-1)}\right)} & \text{if } j > 1 \end{cases} \end{aligned}\right] \tag{8}$$

Based on the arcsine transformation the proposed non-parametric arcsine HWMA ($\mathrm{HWMA_{NAS}}$) chart is defined as:

$$HAS_j = \theta T_j + (1-\theta)\overline{T}_{j-1} \tag{9}$$

where $\overline{T}_{j-1} = \frac{\sum_{l=1}^{j-1} T}{j-1}$ and the value of $T_0 = sin^{-1}\left(\sqrt{p}\right)$. The statistic given in (9) can also be rewritten as:

$$HAS_j = \theta T_j + \left[\left(\frac{1-\theta}{j-1} \right) \left\{ \overline{T}_{j-1} + \overline{T}_{j-2} + \ldots + \overline{T}_1 \right\} \right] \tag{10}$$

The control limits of $\mathrm{HWMA_{NAS}}$ chart are defined as:

$$\left.\begin{aligned} LCL_j &= \begin{cases} sin^{-1}\left(\sqrt{p}\right) - \mathrm{K}\sqrt{\frac{\theta^2}{4n}\sigma_0^2}, & \text{if } j = 1 \\ sin^{-1}\left(\sqrt{p}\right) - \mathrm{K}\sqrt{\frac{\sigma_0^2}{n}\left(\theta^2 + \frac{(1-\theta)^2}{(j-1)}\right)} & \text{if } j > 1 \end{cases} \\ CL &= sin^{-1}\left(\sqrt{p}\right) \\ UCL_j &= \begin{cases} sin^{-1}\left(\sqrt{p}\right) + \mathrm{K}\sqrt{\frac{\theta^2}{4n}\sigma_0^2}, & \text{if } j = 1 \\ sin^{-1}\left(\sqrt{p}\right) + \mathrm{K}\sqrt{\frac{\sigma_0^2}{n}\left(\theta^2 + \frac{(1-\theta)^2}{(j-1)}\right)} & \text{if } j > 1 \end{cases} \end{aligned}\right] \tag{11}$$

where K controls the width of the control limits and it is selected based on the desired in-control average run length.

3. Result

Table 1 The average run length profiles of the existing charts.

p_1	$n = 10$				$n = 12$			
	$\mathrm{EWMA_{NS}}$	$\mathrm{EWMA_{NAS}}$	$\mathrm{CUSUM_{NS}}$	$\mathrm{MEC_{NAS}}$	$\mathrm{EWMA_{NS}}$	$\mathrm{EWMA_{NAS}}$	$\mathrm{CUSUM_{NS}}$	$\mathrm{MEC_{NAS}}$
0.05	4.19	3.44	2.71	2.23	4.01	3.10	2.36	2.11
0.15	5.17	4.44	3.94	3.21	4.56	4.05	3.60	2.90
0.25	7.02	6.29	6.01	4.93	6.26	5.73	5.48	4.36
0.3	8.87	8.08	7.90	6.69	7.92	7.33	7.15	5.80
0.35	12.35	11.37	11.26	10.40	10.88	10.20	10.13	9.00
0.4	21.39	19.10	19.12	22.79	18.65	16.92	16.93	18.46
0.45	64.55	51.47	52.54	88.58	58.13	45.29	45.50	74.48
0.5	369.20	369.53	372.33	369.96	372.00	371.25	370.00	369.56
0.55	64.30	51.37	52.56	86.69	59.07	44.74	44.93	73.73
0.6	21.33	19.04	19.23	22.54	18.58	17.07	16.91	18.46
0.65	12.36	11.40	11.23	10.59	10.96	10.29	10.10	8.94
0.7	8.87	8.11	7.88	6.66	7.91	7.36	7.14	5.80
0.75	7.05	6.33	6.02	4.93	6.28	5.74	5.47	4.38
0.85	5.17	4.44	3.93	3.20	4.57	4.04	3.59	2.90
0.95	4.19	3.44	2.71	2.23	4.01	3.10	2.36	2.10

Table 2 The average run length profiles of the proposed chart.

n		p_1																		
		0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
10	ARL	1.17	1.54	2.04	2.64	3.49	4.79	7.53	14.08	42.93	371.75	42.3	14.3	7.49	4.77	3.48	2.64	2.03	1.54	1.17
	SDRL	0.56	0.92	1.18	1.45	1.85	2.67	4.7	9.94	33.31	308.67	32.49	10.33	4.68	2.65	1.87	1.45	1.19	0.92	0.56
	P5	1	1	1	1	1	1	1	3	4	6	4	3	1	1	1	1	1	1	1
	P25	1	1	1	1	3	3	4	6	17	114	17	6	4	3	3	1	1	1	1
	P50	1	1	1	3	3	4	6	12	36	310	36	12	6	4	3	3	1	1	1
	P75	1	3	3	4	5	6	10	19	60	559	60	20	10	6	5	4	3	3	1
	P95	3	3	4	5	7	10	17	34	106	955.05	104	35	16	10	7	5	4	3	3
12	ARL	1.14	1.45	1.92	2.47	3.18	4.3	6.54	12.34	37.78	369	36.97	12.54	6.53	4.33	3.19	2.47	1.92	1.47	1.13
	SDRL	0.41	0.7	0.96	1.18	1.5	2.22	3.91	8.29	28.61	297.87	27.92	8.45	3.88	2.19	1.47	1.17	0.96	0.72	0.39
	P5	1	1	1	1	1	1	2	3	4	8	4	3	2	1	1	1	1	1	1
	P25	1	1	1	1	2	3	4	6	16	122	16	6	4	3	2	1	1	1	1
	P50	1	1	2	3	3	4	6	11	32	311	31	11	6	4	3	2	2	1	1
	P75	1	2	3	3	4	5	9	17	53	556	51	17	8	6	4	3	3	2	1
	P95	2	3	4	4	6	8	14	29	94	935.05	91	29	14	8	6	4	4	3	2
15	ARL	1.18	1.51	1.89	2.39	3.06	4.28	6.71	13.3	43.14	370.61	43.43	13.23	6.76	4.31	3.11	2.38	1.9	1.5	1.17
	SDRL	0.39	0.6	0.75	0.94	1.26	1.99	3.65	8.39	34.51	405.49	34.4	8.31	3.65	2.01	1.32	0.93	0.74	0.6	0.38
	P5	1	1	1	1	1	2	2	3	7	20	7	3	2	2	1	1	1	1	1
	P25	1	1	1	2	2	3	4	7	18	110	19	7	4	3	2	2	1	1	1
	P50	1	1	2	2	3	4	6	11	34	248.5	34	11	6	4	3	2	2	1	1
	P75	1	2	2	3	4	5	8	18	58	485	59	17	9	5	4	3	2	2	1
	P95	2	3	3	4	5	8	14	29	111	1086	110	29	14	8	6	4	3	3	2

4. Discussion and Conclusion

The performance comparison between proposed and existing charts are done on the basis of a well-known measure the average run length (ARL). We compared the $\mathrm{HWMA_{NAS}}$ with the $\mathrm{EWMA_{NS}}$, $\mathrm{EWMA_{NAS}}$, $\mathrm{CUSUM_{NS}}$ and $\mathrm{MEC_{NAS}}$. It is turned out that the proposed chart outperforms in detecting the small and moderate shifts in the process location than the existing charts (cf. Table 1 and 2). So we recommend the use of the proposed chart for the

practitioners in order to detect earlier shifts when the characteristic of interest does not follow the normal distribution.

References

1. Shewhart WA (1931) Economic control of quality of manufactured product. Bells labs Techn J 9(2):364–389
2. Page ES (1954) Continuous inspection schemes. Biometrika 41: 100–115
3. Roberts SW (1959) Control chart tests based on geometric moving averages. Technometrics 1(3):239–250
4. Lucas JM, Crosier RB (1982) Fast initial response for CUSUM quality control schemes. Technometrics 24:199–205 18.
5. Lucas JM, Saccucci MS (1990) Exponentially weighted moving average control schemes properties and enhancement. Technometrics 32(1):23–26
6. Abbas N, Riaz M, Does RJMM (2012) Mixed exponentially weighted moving average control CUSUM-charts for process monitoring. Qual Reliab Eng Int 29(3):345–356
7. Zaman B, Riaz M, Abbas N, Does RJMM (2015) Mixed cumulative sum-exponentially weighted moving average control charts: an efficient way of monitoring process location. Qual Reliab Eng Int 31(8):1407–1421
8. Abid M, Nazir HZ, Tahir M, Lin Z (2017) An efficient nonparametric EWMA Wilcoxon signed-rank chart for monitoring location. Qual Reliab Eng Int 33(3): 669-685.
9. Abid M, Nazir HZ, Tahir M, Riaz M (2018) On designing a new cumulative sum Wilcoxon signed for monitoring process location. PLOS ONE 13(4): e0195762
10. Abbas N (2019) Homogeneously weighted moving average control chart with an application in substrate manufacturing process. Comp Indust Eng 120(6): 460-470.
11. Yang SF, Cheng SW (2011) A new nonparametric CUSUM sign control charts. Qual Reliab Eng Int 27(7):867–875
12. Hunter JS (1986). The exponentially weighted moving average. J Qual Techn 18: 203–210

Optimal design of some distribution-free EWMA schemes for simultaneous monitoring of location and scale with dynamic fast initial response feature

Zhi Song[1,3], Amitava Mukherjee[2], Jiujun Zhang[3]
[1]Shenyang Agricultural University, Shenyang, China
[2]XLRI-Xavier School of Management, XLRI Jamshedpur, India
[3]Liaoning University, Shenyang, China

Abstract

This paper proposes a new model to optimally design an exponentially weighted moving average (EWMA) scheme with dynamic fast initial response (FIR) feature. An EWMA scheme with the FIR usually detects an initial out-of-control situation much quicker as compared to a standard EWMA scheme. Almost all studies related to the effects of FIR features on monitoring schemes focused on their out-of-control performance. However, when the process actually operates in control set-up, the FIR feature may increase the number of false alarms at the early stage of monitoring. In practice, it is important to restrict the probability of an early false alarm along with improving the out-of-control performance of a monitoring scheme. Noting this, we propose a method to optimally design an EWMA scheme with the FIR feature that guarantees both objectives simultaneously. The proposed method is an improvement over the classical and popular statistical design approach. A data-dependent estimation approach based on Kernel density estimation for the optimal parameter is evaluated and discussed. We apply the proposed optimization model to some distribution-free FIR-based EWMA schemes for joint monitoring of location and scale. Simulation results exhibit that our proposed procedure generally performs well under various continuous distributions.

Keywords

Statistical process control; Exponentially weighted moving average; Fast initial response; Nonparametric; False alarms

1. Introduction

Statistical process monitoring (SPM) schemes play a transformative role for the improvement of product and service quality. Exponentially weighted moving average (EWMA) monitoring schemes are widely used for the detection of small shifts in process parameters. Sometimes it is important to implement a monitoring scheme that is more sensitive than traditional EWMA schemes and allows even quicker detection of a shift, see for example, Lucas and Saccucci [1]. One can easily achieve this by incorporating the fast initial

response (FIR) feature in an EWMA scheme. Lucas and Crosier [2] first proposed the FIR feature for CUSUM schemes. They established that this additional feature helps in detecting early changes more quickly by assigning some non-zero constant to the starting values of the plotting statistic of the CUSUM scheme. Lucas and Saccucci [1] introduced a similar FIR feature for the EWMA schemes. In recent years, many researchers made significant contributions in literature related to the FIR feature in various SPM schemes. Most of these papers were excellently written but mainly focused on the out-of-control (OOC) performance of the schemes with FIR features. There is very little consideration of the consequences of adding the FIR feature on the early false alarm probabilities. Knoth [3] exclusively outlined that the FIR feature certainly helps in achieving higher sensitivity to early changes, but it also increases the probability of early false alarms. Too many early false alarms may have an impact on production cost and also lead to a state of confusion about the suitability of the charting procedure. On the other hand, setting very low probability of early false alarms reduces the efficiency of the monitoring scheme, the almost invariably causes a delay in detecting a true in shift in the process parameters. To this end, we attempt to develop an efficient approach. We observe that almost all previous works on one-sided monitoring schemes with FIR feature consider a starting value halfway between the target value and the control limit. Unlike previous works of fixing a head start value of 50% for the schemes with FIR features, we recommend an optimal (dynamic) head start value that optimizes the chart performance. Our proposed approach restricts probability of an early false alarm to a prefixed value at any situation, while optimizes the early detection of true signal in presence of the FIR feature.

The rest of the paper is organized as follows. In the next Section, we introduces four different nonparametric EWMA schemes under consideration. Section 3 discusses the methodology of optimization of EWMA scheme with FIR features. We study all these schemes with optimal parameters in Section 4. Several remarks conclude this paper in Section 5.

2. Presentation of the competing distribution-free EWMA schemes

We assume that a reference sample $X = (X_1, X_2, \cdots, X_m)$ collected from an IC process with a continuous cumulative distribution function (cdf) $F(x)$. We also assume that X_i is are independently and identically distributed (i.i.d) random variable each having cdf F. Let $Y_j = (Y_{j1}, Y_{j2}, \cdots, Y_{jn})$, $j = 1, 2, \cdots$ be the jth Phase-II (test) sample of size n mutually independent of the reference sample, from a cdf $G(x) = F(\frac{x-\theta}{\delta})$, $\theta \in \Re$, $\delta > 0$, where the constants θ and δ represent the unknown location and scale parameters, respectively. We further assume that Y_{ji}s are i.i.d for every i ($1 \leq i \leq n$ and j, ($j > 1$). The process is considered to be

IC if during the course of Phase-II monitoring we observe $F = G$, that is when $\theta = 0$ and $\delta = 1$. When the process is OOC, either $\theta \neq 0$ or $\delta \neq 1$ or both.

2.1 EL monitoring shemes

This section provides the structure of traditional EL monitoring sheme and further extends its design by using FIR features. Mukherjee [4] proposed some single distribution-free EWMA schemes for monitoring the location and the scale parameters of an unknown but continuous univariate process at Phase-II. These schemes are based on the well-known Lepage statistic and referred to as the EL procedures. The Lepage statistic is the sum of squares of the standardized Wilcoxon rank-sum (WRS) statistic for location and the standardized Ansari-Bradley (AB) statistic for scale. The WRS statistic, denoted as $T_{W,j}$, is interpreted as the sum of ranks of the jth test sample in the combined sample of size N(=m+n). Further, the AB statistic, denoted as $T_{AB,j}$, is defined by the sum of the absolute deviation of ranks of the jth test sample in the combined sample from the average rank, that is, (N +1)/2. The standardized WRS and AB statistics for the jth inspection state are $S_{W_j} = \frac{T_{W,j}-\mu_W}{\sigma_W}$ and $S_{AB_j} = \frac{T_{AB,j}-\mu_{AB}}{\sigma_{AB}}$, respectively, where (μ_W, σ_W) and (μ_{AB},σ_{AB}) are the means and standard deviations of $T_{W,j}$ and $T_{AB,j}$, under the IC case: $\theta = 0$ and $\delta = 1$. Detailed expressions for (μ_W, σ_W) and (μ_{AB},σ_{AB}) are given in the work of Mukherjee and Chakraborti [5] and hence are omitted here. Thereafter, the Lepage statistic is defined as $S_j^2 = S_{W_j}^2 + S_{AB_j}^2$.

The plotting statistic of the EL procedure is given by $Z_j = \max\{2, \lambda S_j^2 + (1-\lambda) Z_{j-1}\}$, j = 1,2,... and with the starting value $Z_0 = 2$, as $E(S_j^2|IC) = 2$. Here, $0 < \lambda \leq 1$ is the smoothing parameter. Next, we look on some FIR-based EL schemes, which allow to improve the detection performance at early time points.

1. **EL scheme with FIR version of Lucas and Saccucci [1] [EL-fir]** Lucas and Saccucci [1] used an FIR feature for the EWMA to improve its performance at start-up. We propose using the FIR feature with the fixed-width control limit EL scheme (denoted as EL-fir) in the line of Lucas and Saccucci [1], which is formulated as follows:

$$C = inf\left\{ j \in \mathbb{N} | Z_j > 2 + L\sqrt{\frac{4\lambda}{2-\lambda}} \right\}$$

The EL-fir scheme triggers a signal whenever Z_j exceeds the $UCL\ 2 + L\sqrt{\frac{4\lambda}{2-\lambda}}$. The stopping variable C is the number of samples until the scheme first generates a signal. We set the starting value $Z_{oh} = 2 + h \times L\sqrt{\frac{4\lambda}{2-\lambda}}$, where

h ($0 \leq h \leq 1$) is the head-start parameter. When $h = 0$, the EL-fir scheme is the original EL scheme with fixed control limit.

2. EL scheme with FIR version of Rhoads et al. [6] [EL-fvacl] Rhoads et al. [6] proposed an FIR feature for the EWMA in the manner suggested by Lucas and Saccucci [1] but using the time-varying control limits. Following the same idea, the FIR feature with the EL scheme based on the variance-adjusted control limit, denoted as EL-fvacl, can be constructed in the following way:

$$C = inf\{ j \in N \mid Z_j > 2+L\sqrt{\frac{4\lambda}{2-\lambda}(1-(1-\lambda)^{2j}}\},$$

with the starting value $Z_{0h} = 2+h\times L\sqrt{\frac{4\lambda}{2-\lambda}(1-(1-\lambda)^{2})} = 2+h\times L\sqrt{\frac{4\lambda}{2-\lambda}(\lambda(2-\lambda))}$ (cf. Knoth [3]).

2.2 EC monitoring schemes

Mukherjee [7] proposed a single distribution-free EWMA scheme for jointly monitoring the location and the scale parameters, which is based on the Cucconi statistic and referred to as the EC procedure. This section provides the structure of the EC monitoring scheme and further extends its design by using FIR features. Before defining these structures, we first briefly review the Cucconi statistic.

Define the following statistics:

$$T_{1,j} = \sum_{k=1}^{N} kI_k \; and \; S_{1,j} = \sum_{k=1}^{N} k^2 I_k,$$

where I_k is an indicator variable, $I_k = 0$ or 1 according as the kth order statistic of the combined sample is an X observation or a Y_j observation. $T_{1,j}$ is the WRS statistic for location, and similarly $S_{1,j}$ represents the sum of the squares of the ranks of the jth test sample in the combined sample. Further, the sum of the squares of anti-ranks of the jth test sample in the combined sample, say $S_{2,j}$, is given by

$$S_{2,j} = \sum_{k=1}^{N} (N+1-k)^2 I_k = n(N+1)^2 - 2(N+1)T_{1,j} + S_{1,j}.$$

Define the standardized statistics: $U_j = \frac{S_{1,j}-\mu_1}{\sigma_1}, V_j = \frac{S_{2,j}-\mu_2}{\sigma_2}$, and $\rho = Corr(U_j, V_j | IC)$, where (μ_1, μ_2) and σ_1, σ_2 are the respective means and standard deviations of $S_{1,j}$ and $S_{2,j}, \rho$ denotes the correlation coefficient between U_j and V_j. Detailed expressions for $(\mu_1, \mu_2), \sigma_1, \sigma_2$ and ρ are given in the work of Chowdhury et al. [8] and hence are omitted here. Consequently, the Cucconi statistic is defined by

$$C_j = \frac{U_j^2 + V_j^2 - 2\rho U_j V_j}{2(1-\rho^2)}.$$

Note that $E(C_j|IC) = 1, Var(C_j|IC) = 1$. Then the EC scheme is given by $Z_j = \max\{1, \lambda C_j + (1-\lambda)Z_{j-1}\}, j = 1,2,\ldots$ and with the starting value $Z_0 = 1$. Next, we incorporate the similar FIR features as in Section 2.1 in the EC scheme for quicker detection of an initial OOC situation.

1. **EC scheme with FIR version of Lucas and Saccucci [1] [EC-fir]** We develop using the FIR feature with the fixed control limit EC scheme (denoted as EC-fir) in the line of Lucas and Saccucci [1], which can be constructed in the following way:

$$C = inf\left\{j \in \mathbb{N} | Z_j > 1 + L\sqrt{\frac{\lambda}{2-\lambda}}\right\}.$$

We set the starting value $Z_{0h} = 1 + h \times L\sqrt{\frac{\lambda}{2-\lambda}}$. When $h = 0$, the EC-fir scheme is the fixed control limit EC scheme without FIR features.

2. **EC scheme with FIR version of Rhoads et al. [6] [EC-fvacl]** The FIR feature with the EC scheme based on the variance-adjusted control limit, denoted as EC-fvacl, is formulated as follows:

$$C = inf\left\{j \in \mathbb{N} | Z_j > 1 + L\sqrt{\frac{\lambda}{2-\lambda}((1-(1-\lambda)^{2j})}\right\}.$$

with the starting value $Z_{0h} = 1 + h \times L\sqrt{\frac{\lambda}{2-\lambda}(\lambda(2-\lambda))}$ (cf. Knoth [3]).

3. Optimization of EWMA schemes with FIR

In this section, we focus on optimally designing an FIR-EWMA scheme that restricts early false alarm probabilities and facilitates quick detection. As mentioned earlier, use of the FIR feature in a monitoring scheme has gained a great deal of attention among the researchers over the years. Nevertheless, the majority of the existing schemes with the FIR feature set the initial value of the monitoring statistic at the halfway between the process target value and the control limits. That is, they are based on a 50% advanced head start. The OOC performance of such scheme is often better, specially for detecting shifts at an early stage. Nevertheless, the probability of a few early false alarms also increases. Therefore, we prefer to design an optimal FIR-based EWMA scheme selecting the head start or starting value Z0h that ideally increases sensitivity without compromising with the early false alarm probabilities.

Generally, the ARL is the widely used criterion for assessing the performance of monitoring schemes. However, since the FIR feature is more meaningful for early detection, we use two other design criteria rather than depending solely on the ARL performance measure. One design criterion is cumulative unconditional false alarm probability (CUFAP), which is denoted by CUFAP(x)=P($RL \leq x\,|IC$). Another is cumulative unconditional true signal probability (CUTSP), denoted as CUTSP(x)=P($RL \leq x\,|OOC$). Thereafter, we may use a simple algorithm to optimize the design of FIR-based EWMA schemes via the following steps:

Step-1: Given m, n, λ, and a target IC ARL (ARL_0), say τ, compute L for a given EWMA-LS scheme without the FIR feature. The abbreviation LS stands for the Location-Scale statistic, namely, Lepage or Cucconi, with $Z_0 = E$(LS|IC).

Step-2: Let $\xi \in [0,1]$ is a small positive proper fraction chosen a-priori by the practitioner, denotes the maximum allowable CUFAP till W-th test sample. Compute CUFAP(W) for a pre-specified $W \in$ N. If CUFAP(W) $> \xi$, increase L by a margin of 0.01 and re-compute ARL_0 and CUFAP(W). Note that, in general, the CUFAP(W) is a decreasing function and the ARL_0 is an increasing function of L given other charting parameters, namely, m, n, λ, remain same.

Step-3: Repeat the Step-2, unless, CUFAP(W) $\leq \xi$.

Step-4: Set the CUFAP(W) $= \xi$, and use the L obtained at the end of the Step-3, say L_0, to determine dynamic FIR quotient. Using Z_0 and L_0 determine the distance, say dLS between Z_0 and UCL at the first test sample. Set FIR feature $h_{LS} \in [0,1]$, where $h_{LS} = h$ implies that the starting value will be reset at $Z_{0h} = Z_0 + h \times d_{LS}$.

Step-5: We see, in general, for fixed ARL_0, if Z_0 increases to $Z_{0h} h > 0,$, CUFAP(W) increases. Therefore, to maintain the specified preapproved level of CUFAP(W), we need to increase the L to L_h and consequently, the ARL_0, will also increase. Compute L_h for each $Z_{0h} h = 0.01(0.01)1.00$.

Step-6: Given m, n, λ, W and target shift $\theta = \theta_s, \delta = \delta_s$ compute CUTS$P_{h,\theta_s,\delta_s}(W)$ for all $h = 0.00(0.01)1.00$.

Step-7: Select dynamic FIR as $h_{LS} = h$ for which CUTSP$_{h,\theta_s,\delta_s}(W)$ is maximum.

Because of the distribution-free nature of the proposed class of monitoring schemes when the process is IC, determination of L_h value based on a fixed CUFAP(W) will be the same for all continuous distributions. We may generate

m observations from a standard normal distribution for the reference sample and n observations from the same distribution for each test sample. When several combinations of parameters yield the pre-specified CUFAP(W) value, the CUTSP(W) criterion is used to select the optimal monitoring scheme parameters. The combination (h, L_h) that yields the maximum CUTSP(W) is selected to be the optimal combination. However, OOC run length distribution depends on the underlying density and shift sizes. In other words, CUTSP(W) values may vary with different densities and shift sizes. The exact value of CUTSP(W) cannot be obtained since the underlying process distribution $F(x)$ is unknown in practice. Therefore, computation of CUTSP(W) requires special attention and to this end we propose a Kernel density estimation (KDE) approach. KDE is a nonparametric way to estimate the probability density function (pdf) of a random variable, based on a pilot sample. We can further draw the simulated reference sample of size m and the simulated test sample of size n from the fitted density and use the obtained L_h to compute the estimates of CUFAP(W) and CUTSP(W). Then the optimal estimate of h, say $\hat{h}$, is selected for which the estimated CUTSP(W) is maximum. Inevitably, some error may creep in during estimate of h. Consequently, the estimated $\hat{h}$ differs slightly from the theoretical h, or called true h. Note that the pure theoretical h is practically unobservable as the underlying process distribution $F(x)$ is unknown. We consider this as a benchmark. The efficacy of the optimization model depends on the closeness between the $\hat{h}$ and h. In other words, we expect the estimated optimal design scheme to be closer to the theoretical optimal scheme.

4. Performance comparison

In the previous section, we frame the optimization model of nonparametric FIR-based EWMA schemes to satisfy a desired value of CUFAP(W) and to minimize CUTSP(W). We note that the exact underlying distribution is unknown. Therefore, we propose a nonparametric approach, namely KDE for estimating CUFAP and CUTSP. Clearly, the actual performance of the scheme is affected by the accuracy of estimate of CUTSP. We here apply the optimization model to the above FIR-based EL and EC schemes as described in Section 2. In order to conduct a thorough investigation, our study includes three typical distributions and considers thin-tailed, heavy-tailed, symmetric and skewed distributions. Specifically, the distributions considered in the study are: (a) the thin-tailed symmetric normal distribution abbreviated as $N(\theta, \delta)$, the IC sample is from $N(0,1)$, but the test samples are from a $N(\theta, \delta)$; (b) the heavy-tailed symmetric Cauchy distribution, denoted by Cauchy(θ, δ). The IC sample is taken from Cauchy(0,1), with the test samples coming from a Cauchy (θ, δ) distribution with pdf $f(x) = \frac{\delta}{\pi(\delta^2+(x-\theta)^2)}, x \in (-\infty, \infty)$; (c) the

shifted exponential distribution (denoted by SE(θ, δ)) represents the skewed distribution. The IC sample is from SE(0,1), but the test samples are from a SE(θ, δ) distribution having pdf $f(x) = e^{-\frac{1}{\delta}(x-\theta)}, x \in (\theta, \infty)$ with mean=$\theta + \delta$ and variance= δ^2. To examine the effect of shifts in location and scale parameters, we consider the quartile deviation (QD) for each of the three distributions. 12 combinations of θ and δ values are considered, that is, θ =0, QD/4 and QD/2 along with δ =1, 1+QD/4, 1+QD/2 and 1+QD. For brevity, we only represent $m = 100$, $n = 5$ and $\lambda = 0.1$ case for all schemes for comparison purposes. A similar conclusion holds for other parameter conditions. Two situations, namely "Ideal case (True case)" and "Practical case (Estimated case)" are considered for numerical studies. As their names imply, these cases are briefly explained in Section 3. The ideal case is unobservable because the process distribution is unknown. The practical case is the optimization procedure of what we are facing. Here we choose $W = 10$ and $\xi = 0.04$, i.e., CUFAP(10)=0.04. Employing constraints on CUFAP(10)=0.04, we obtain the optimal chart schemes (h, L_h) and $(\hat{h}, L_{\hat{h}})$, respectively for the ideal case and the practical case, to achieve a minimum of CUTSP(10). Usually, exact amount of possible shift is unknown and therefore, the practitioners will prefer to choose a combination of $(\hat{h}, L_{\hat{h}})$ that has overall good performance irrespective of the exact size of shift. To this end, we further introduce the mean of $\hat{h}$, recorded as $\bar{\hat{h}}$. We compute this by calculating the mean of $\hat{h}$ over shifts (θ, δ) under consideration. We also consider the mean of h, say $\bar{h}$ as a benchmark and for comparison. The comparative results for the two cases are summarized in Table 1. From Table 1, we find that the optimization model performs very efficiently for all the three distributions for most of the 11 OOC scenarios in terms of the proximity between h and $\hat{h}$. Furthermore, the last row under each distribution in Table 1 is $\bar{h}$ $(L_{\bar{h}})$ and $(\bar{\hat{h}}, L_{\bar{\hat{h}}})$ of each scheme. It is observed that they are extremely close.

5. Concluding remarks

In this paper, we present an optimal designing strategy for the nonparametric EWMA schemes with FIR features to facilitate early detection of shift with a restriction on the early false alarm probability. Then we apply this design method to the wellknown EL and EC monitoring schemes for implementation. It is worth mentioning that the distribution-free characteristic of the plotting statistic of a nonparametric scheme is, in general, not valid under a process shift. As a consequence, the pure theoretical optimal charting scheme is practically unobservable. Noting that the underlying process distribution is often unknown, we propose a data-dependent estimation procedure based on KDE for evaluation of the optimal design parameters. Simulation results show that overall performance of the estimation procedure

based on KDE is highly encouraging. Moreover, results are robust for various continuous distributions. Thus, the proposed optimal designing strategy can be very useful in practical applications.

References

1. J.M. Lucas, M.S. Saccucci, Exponentially weighted moving average control schemes: properties and enhancements, Technometrics. 32 (1990) 1-12.
2. J.M. Lucas, R.B. Crosier, Fast initial response for CUSUM quality-control schemes: give your CUSUM a head start, Technometrics. 24 (1982) 199-205.
3. S. Knoth, Fast initial response features for EWMA control charts, Stat. Pap. 46 (2005) 47-64. doi:10.1007/bf02762034
4. A. Mukherjee, Distribution-free phase-II exponentially weighted moving average schemes for joint monitoring of location and scale based on subgroup samples, Int. J. Adv. Manuf. Tech. 92 (2017) 101-116.
5. A. Mukherjee, S. Chakraborti, A distribution-free control chart for the joint monitoring of location and scale, Qual. Reliab. Engng. Int. 28 (2012) 335-352.
6. T.R. Rhoads, D.C. Montgomery, C.M. Mastrangelo, A fast initial response scheme for the exponentially weighted moving average control chart, Qual. Eng. 9 (1996) 317-327. doi:10.1080/08982119608919048.
7. A. Mukherjee, Recent development in phase-II monitoring of location and scale - an overview and some new results, 61st ISI World Statistics Congress, Marrakesh Morocco, 2017.
8. S. Chowdhury, A. Mukherjee, S. Chakraborti, A new distribution-free control chart for joint monitoring of unknown location and scale parameters of continuous distributions, Qual. Reliab. Engng. Int. 30 (2013) 191-204.

Table 1: Optimal $h(L_h)$ $\hat{h}(L_{\hat{h}})$ of various EL and EC schemes with the FIR feature for $N(\theta,\delta)$, Cauchy (θ,δ) and SE (θ,δ) distributions with $m=100, n=5$ when $\lambda=0.1$ and CUFAP(10)=0.04.

$N(\theta,\delta)$					
θ	δ	EL-fir	EL-fvacl	EC-fir	EC-fvacl
0	1+QD/4	0.14(2.65),0.13(2.65)	0.04(3.32),0.05(3.33)	0.16(2.85),0.15(2.85)	0.03(3.55),0.03(3.55)
0	1+QD/2	0.14(2.65),0.12(2.63)	0.05(3.33),0.05(3.33)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
0	1+QD	0.18(2.69),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.16(2.85)	0.04(3.57),0.03(3.55)
QD/4	1	0.14(2.65),0.14(2.65)	0.04(3.32),0.07(3.36)	0.24(2.96),0.20(2.91)	0.03(3.55),0.04(3.57)
QD/4	1+QD/4	0.16(2.67),0.12(2.63)	0.05(3.33),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
QD/4	1+QD/2	0.18(2.69),0.20(2.72)	0.05(3.33),0.05(3.33)	0.16(2.85),0.20(2.91)	0.03(3.55),0.03(3.55)
QD/4	1+QD	0.18(2.69),0.18(2.69)	0.05(3.33),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
QD/2	1	0.16(2.67),0.14(2.69)	0.04(3.32),0.04(3.32)	0.21(2.92),0.18(2.89)	0.03(3.55),0.06(3.59)
QD/2	1+QD/4	0.14(2.65),0.18(2.69)	0.05(3.33),0.03(3.32)	0.24(2.96),0.24(2.96)	0.03(3.55),0.03(3.55)
QD/2	1+QD/2	0.18(2.69),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
QD/2	1+QD	0.18(2.69),0.18(2.69)	0.04(3.32),0.05(3.33)	0.16(2.85),0.16(2.85)	0.03(3.55),0.01(3.54)
$\bar{h}(L_{\bar{h}}),(\bar{\hat{h}},L_{\bar{\hat{h}}})$		0.16(2.67)0.16(2.67)	0.05(3.33),0.05(3.33)	0.18(2.89),0.18(2.89)	0.03(3.55),0.03(3.55)
Cauchy(θ,δ)					
θ	δ	EL-fir	EL-fvacl	EC-fir	EC-fvacl
0	1+QD/4	0.16(2.67),0.16(2.67)	0.04(3.32),0.05(3.33)	0.29(3.05),0.24(2.96)	0.03(3.55),0.03(3.55)
0	1+QD/2	0.18(2.69),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
0	1+QD	0.18(2.69),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
QD/4	1	0.12(2.63),0.10(2.62)	0.05(3.33),0.04(3.32)	0.16(2.85),0.13(2.83)	0.03(3.55),0.05(3.59)
QD/4	1+QD/4	0.14(2.65),0.16(2.67)	0.04(3.32),0.04(3.32)	0.24(2.96),0.21(2.92)	0.03(3.55),0.02(3.55)
QD/4	1+QD/2	0.16(2.67),0.18(2.69)	0.04(3.32),0.05(3.33)	0.16(2.85),0.16(2.85)	0.03(3.55),0.02(3.55)
QD/4	1+QD	0.21(2.73),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.21(2.92)	0.03(3.55),0.03(3.55)
QD/2	1	0.14(2.65),0.16(2.67)	0.04(3.32),0.04(3.32)	0.24(2.96),0.24(2.96)	0.03(3.55),0.03(3.55)
QD/2	1+QD/4	0.16(2.67),0.16(2.67)	0.04(3.32),0.04(3.32)	0.16(2.85),0.16(2.85)	0.03(3.55),0.03(3.55)
QD/2	1+QD/2	0.18(2.69),0.18(2.69)	0.04(3.32),0.04(3.32)	0.24(2.96),0.21(2.92)	0.04(3.57),0.03(3.55)
QD/2	1+QD	0.28(2.83),0.25(2.79)	0.04(3.32),0.05(3.33)	0.24(2.96),0.21(2.92)	0.03(3.55),0.03(3.55)
$\bar{h}(L_{\bar{h}}),(\bar{\hat{h}},L_{\bar{\hat{h}}})$		0.17(2.69),0.17(2.69)	0.04(3.32),0.04(3.32)	0.19(2.90),0.19(2.90)	0.03(3.55),0.03(3.55)
SE(θ,δ)					
θ	δ	EL-fir	EL-fvacl	EC-fir	EC-fvacl
0	1+QD/4	0.18(2.69),0.13(2.65)	0.04(3.32),0.05(3.33)	0.16(2.85),0.18(2.89)	0.03(3.55),0.03(3.55)
0	1+QD/2	0.16(2.67),0.18(2.69)	0.04(3.32),0.04(3.32)	0.16(2.85),0.18(2.89)	0.03(3.55),0.03(3.55)
0	1+QD	0.12(2.63),0.16(2.67)	0.04(3.32),0.05(3.33)	0.16(2.85),0.16(2.85)	0.01(3.54),0.03(3.55)
QD/4	1	0.29(2.85),0.28(2.83)	0.06(3.35),0.08(3.37)	0.41(3.30),0.41(3.30)	0.06(3.59),0.06(3.59)
QD/4	1+QD/4	0.16(2.67),0.18(2.69)	0.04(3.32),0.04(3.32)	0.24(2.96),0.24(2.96)	0.02(3.55),0.03(3.55)
QD/4	1+QD/2	0.12(2.63),0.14(2.65)	0.04(3.32),0.04(3.32)	0.24(2.96),0.24(2.96)	0.03(3.55),0.03(3.55)
QD/4	1+QD	0.21(2.73),0.18(2.67)	0.04(3.32),0.05(3.33)	0.16(2.85),0.21(2.92)	0.03(3.55),0.03(3.55)
QD/2	1	0.30(2.86),0.30(2.86)	0.04(3.32),0.05(3.33)	0.24(2.96),0.24(2.96)	0.08(3.61),0.04(3.57)
QD/2	1+QD/4	0.18(2.69),0.14(2.65)	0.05(3.33),0.04(3.32)	0.29(3.05),0.24(2.96)	0.03(3.55),0.02(3.55)
QD/2	1+QD/2	0.30(2.86),0.28(2.83)	0.05(3.33),0.04(3.32)	0.24(2.96),0.24(2.96)	0.03(3.55),0.03(3.55)
QD/2	1+QD	0.22(2.75),0.21(2.73)	0.05(3.33),0.05(3.33)	0.16(2.85),0.16(2.85)	0.03(3.55),0.02(3.55)
$\bar{h}(L_{\bar{h}}),(\bar{\hat{h}},L_{\bar{\hat{h}}})$		0.20(2.72),0.20(2.72)	0.05(3.33),0.05(3.33)	0.22(2.95),0.23(2.96)	0.03(3.55),0.03(3.55)

A comparative study between the standard deviation of the time to signal (SDTS) performance of the GR and SSGR schemes

Zhi Lin Chong[1], Michael BC Khoo[2], Huay Woon You[3], Wei Lin Teoh[4], Sin Yin Teh[5]

[1]Universiti Tunku Abdul Rahman, Perak, Malaysia.
[2,5]Universiti Sains Malaysia, Penang, Malaysia
[3]Universiti Kebangsaan Malaysia, Selangor, Malaysia
[4]Heriot-Watt University Malaysia, Putrajaya, Malaysia.

Abstract

The Group Runs (GR) scheme is introduced as an improved version of the synthetic scheme; whereas the Side-Sensitive Group Runs (SSGR) scheme is introduced as an enhancement of the GR scheme with the side-sensitive feature. However, the average run length or average time to signal criterion is commonly applied as a sole performance measure for both the GR and SSGR schemes. Hence, this paper aims to reexamine the performances of the GR and SSGR schemes using the performance measure of the standard deviation of the time to signal (*SDTS*), which have not been investigated in the literature. Based on the comprehensive simulation studies, it is discovered that although the SSGR scheme has a better out-of-control *SDTS* performance compared to the GR scheme, its in-control *SDTS* performance is inferior to that of the GR scheme.

Keywords

Control chart, Group Runs (GR), Side-Sensitive Group Runs (SSGR)

1. Introduction

Contemporarily, quality control is vitally viewed as high quality product will boost the reputation and improve the sustainability of any industry. Statistical Process Control (SPC) is commonly implemented to control and enhance the product quality. Among the many tools in SPC, control scheme receives the most attention due to its effectiveness in categorizing the process as in-control (IC) or out-of-control (OC). The first control chart was introduced by W. A. Shewhart in the 1920s and it is called the Shewhart $\bar{X}$ scheme. Since then, numerous extensions on the Shewhart $\bar{X}$ scheme have been proposed. Although the Shewhart $\bar{X}$ scheme is the most widely applied control scheme in industries and it is effective in detecting large and abrupt mean shifts, it receives much criticisms due to its inefficiency in detecting small to moderate and persistent mean shifts. To solve this problem, Roberts (1959) and Page (1954) introduced the Exponentially Weighted Moving Average (EWMA) and Cumulative Sum (CUSUM) schemes, respectively, which are effective in detecting small mean shifts.

In recent years, Wu and Spedding (2000) introduced the synthetic scheme through the combination of the Shewhart $\bar{X}$ sub-chart with the Conforming Run Length (CRL) sub-chart. They showed that the synthetic scheme surpasses the traditional Shewhart $\bar{X}$ scheme under all magnitudes of mean shifts and also outperforms the EWMA $\bar{X}$ scheme for mean shift sizes $\delta > 0.8$. As an effort to improve the synthetic scheme, Gadre and Rattihalli (2004) introduced the Group Runs (GR) scheme by incorporating the Shewhart $\bar{X}$ sub-chart with an improved version of the CRL sub-chart. Moreover, Gadre and Rattihalli (2006) further improved the GR scheme by introducing the Modified GR (MGR) scheme. Then, Gadre and Rattihalli (2007) introduced the Side-Sensitive GR (SSGR) scheme, which is essentially a GR scheme with the side-sensitive feature. The Side-Sensitive MGR (SSMGR) scheme was proposed by (2010), which is essentially a MGR scheme with the side-sensitive feature.

Recently, the GR-type schemes are widely investigated by numerous researchers. You et al. (2015) studied the SSGR scheme in the scenario where the process parameters are unknown and have to be estimated from a Phase-I sample. Lim et al. (2015) scrutinized the economic and economic statistical designs of the SSGR scheme by minimizing the cost. Chong et al. (2017) introduced the GR revised m-of-k runs rules scheme by combining the GR scheme and revised m-of-k runs rules scheme proposed by Antzoulakos and Rakitzis (2008). Chong et al. (2019) investigated the MGR scheme when the process parameters are unknown.

This study is motivated by Yew et al. (2016) who considered the performance comparison of the GR and SSGR schemes using the average time to signal (ATS) criterion. However, as in all the GR-type schemes mentioned above, Yew et al. (2016) did not consider the performance of the standard deviation of the time to signal ($SDTS$) between the GR and SSGR schemes. In the performance comparison using the $SDTS$ criterion, the scheme with the lowest $SDTS$ is desirable as it demonstrates that the variability of time to signal distribution of the scheme is lower and hence its ATS performance is more predictable. Therefore, the aim of this study is to examine the $SDTS$ performance of the GR and SSGR schemes.

2. Operations of the GR and SSGR schemes:

The operations of the GR and SSGR schemes are given in this section.

2.1 The GR Scheme:

The GR scheme is proposed as an extension of the synthetic scheme. Similar to the synthetic scheme, for the GR scheme, a point plotted outside the control limits of the $\bar{X}$ sub-chart is not immediately treated as an OC signal, but just a nonconforming sample, awaiting the decision from the *CRL* sub-chart. We define the *CRL* value as the number of conforming samples between the previous (excluded in the count) and current (included in the count) nonconforming

samples. The GR scheme issues an OC signal for the first occurrence of $CRL_1 < LG$ or two successive $CRL_i < L_G$ and $CRL_{i+1} < LG$, for i = 2, 3, ... Here, CRL_i is the i-th *CRL* value and *LG* is the *LCL* of the extended *CRL* sub-chart. For this reason, the GR scheme can signal at any sample except the second sample, i.e. i = 2. The performance of the GR scheme can be measured by using the *ATS* criterion, given by (Gadre and Rattihalli, 2004)

$$ATS(\delta) = \frac{n}{P(\delta)} \times \frac{1}{[1-(1-P(\delta))]^2} \quad (1)$$

where $P(\delta)$ is the probability for nonconforming sample to happen under shift size δ, i.e.

$$P(\delta) = 1 - \Phi\left(k - \delta\sqrt{n}\right) + \Phi\left(-k - \delta\sqrt{n}\right) \quad (2)$$

2.2 The SSGR scheme

The SSGR scheme is proposed as an extension of the GR scheme with side-sensitivity. Similar to the GR scheme, the SSGR scheme also issues an OC signal if $CRL_1 < L_{SSG}$ or two successive $CRL_i < L_{SSG}$ and $CRL_{i+1} < L_{ssG}$, for i = 2, 3, ..., but the difference is that the SSGR scheme needs to satisfy another condition, i.e. the two consecutive $\bar{X}$ samples contributing to two successive *CRLs* have to be plotted on the same side of the target mean. The *ATS* of the SSGR scheme is given by (Gadre and Rattihalli, 2007)

$$ATS(\delta) = \frac{n}{P(\delta)} \times \frac{1-\beta(1-\beta)A^2}{[1+\beta(1-\beta)(A-2)]A^2} \quad (3)$$

where $P(\delta)$ is given in (2), $\beta = \frac{1-\Phi\left(k-\delta\sqrt{n}\right)}{P(\delta)}$ and $A = 1 - [1 - P(\delta)]^L$.

3. Optimization design of the GR and SSGR schemes

In this section, we present the optimization design of the GR and SSGR schemes. For the optimization design, we need to find the optimal parameters (k, L_G) and (k, L_{ssG}) of the GR and SSGR schemes, respectively, such that the OC *ATS* (ATS_1) is minimized, where the IC average run length (ARL_0) is attained at 370 so that a low false alarm rate is achieved. We present the optimization procedure written in the ScicosLab software (www.scicoslab.org) to search for the optimal parameters (k, L_G) of the GR scheme as follows:

Step 1. Specify the desired n, δ and ARL_0.

Step 2. Initialize L_G as 1.

Step 3. Use nonlinear equation solver to solve (1) by searching for k such that $ATS(\delta) = ATS_0$ is satisfied, where $ATS_0 = n \times ARL_0$ is the desired IC ATS value. Then, calculate the ATS_1 value using the present parameters (k, L_G).

Step 4. Increase L_G by 1 and go back to Step 3 if $L_G = 1$ or there is a reduction in the ATS_1 value. Otherwise, go to the next step.

Step 5. Choose the parameters (k, L_G) that result in the minimum ATS_1

value as the optimal parameters.

Note that the optimal parameters (k, L_{ssG}) of the SSGR scheme can also be obtained using the above optimization procedure by replacing L_G with L_{ssG} and (1) with (3).

4. A Comparison between the GR and SSGR schemes based on SDTS

We compare the performance between the optimal GR and SSGR schemes in terms of *SDTS* in this section. Note that we use the Statistical Analysis System (SAS) software to compute the SDTSs based on the Monte-Carlo simulation procedure. We utilize the same simulation procedure as shown in Yew et al. (2016) to compute the ATS for the GR and SSGR schemes. Here, we repeat the simulation for 10000 trials and the average of the 10000 trials is the ATS value. Similarly, the SDTS for the GR and SSGR schemes can be computed based on the standard deviation of these 10000 trials.

To have a comprehensive comparison between the GR and SSGR schemes, we considered the input parameters combination of $n \in \{3, 5, 7\}$, $\delta_{opt} \in \{0.5, 1.0, 1.5\}$ and $ARL_0 \in [370, 500)$. Note that δ_{opt} represents the optimal mean shift size where a prompt detection is desired. In practical situations, small to moderate sample sizes are generally recommended to reduce the cost of sampling. Here, the combinations of δ_{opt} are considered as in Lee et al. (2013). To have a correct implementation of the GR and SSGR schemes, the practitioners are encouraged to investigate the input parameters combination based on their respective needs.

Then, based on the combination of $(n, \delta_{opt}, ARL_0$ input parameters, we determine the optimal parameters, i.e. (k, L_G) and (k, L_{ssG}) of the GR and SSGR schemes, respectively, using the optimization procedure in Section 3. We present these optimal parameters in Table 1. Note that these optimal parameters are chosen to minimize the ATS_1 for respective n and δ_{opt}, such that the desired ARL_0 is attained. From Table 1, we observe that the optimal parameters (k, L_G) and (k, L_{ssG}) of the GR and SSGR schemes, respectively, generally decrease or remain the same as n and δ_{opt} increase. However, when ARL_0 increases from 370 to 500, the optimal parameter k will increase; whereas the optimal parameter L_G or L_{ssG} will either increase or remain the same.

The optimal parameters (k, L_G) and (k, L_{ssG}) presented in Table 1 are applied to compute the SDTSs of the GR and SSGR schemes, respectively, and the results are shown in Tables 2 to 4. We compare and study the SDTSs of the optimal GR and SSGR schemes for different mean shifts sizes, i.e. $\delta \in \{0, 0.25, 0.50, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0\}$. We say that the process is IC when $S = 0$, whereas the process is OC when $\delta > 0$. Note that in Tables 2 to 4, the $SDTS_G$ and $SDTS_{ssG}$, respectively, denote the SDTSs of the GR and SSGR schemes.

4.1 Performance comparison between the GR and SSGR schemes (IC SDTS):

In this subsection, we compare the performance of the IC SDTS ($SDTS_0$) between the GR and SSGR schemes, i.e. case $\delta = 0$ in Tables 2 to 4. In these tables, we observe that the $SDTS_0$ value increases as the ARL_0 value increases. Although the IC ATS (ATS_0) values of the GR and SSGR schemes are not shown in these tables, they can be easily computed using the formula $ATS_0 = n \times ARL_0$. For example, consider the case of $n = 3$, $\delta_{opt} = 0.5$ and $ARL_0 = 370$ in Table 2, the $ATS_0 = n \times ARL_0 = 1110$. Still considering the same case, however, the $SDTS_0$s of the GR and SSGR schemes are 1540.00 and 1570.26, respectively, which is around 40% greater than the ATS_0 value. This result is in contrast with the optimal EWMA and CUSUM schemes considered in Lee et al. (2013), where they showed that the ARL0 and IC standard deviation of the run length ($SDRL_0$) are approximately equal. This implies that the GR and SSGR schemes have a larger variability in the IC performance compared to the EWMA and CUSUM schemes. This may lead to an inconsistent IC performance for the GR and SSGR schemes. The SDTS0 comparison between the GR and SSGR schemes shows that the SSGR scheme has a slightly larger $SDTS_0$ compared to the GR scheme. This indicates that the GR scheme has a slight advantage in the IC performance compared to the SSGR scheme due to a lower variability in the time to signal distribution.

Table 1. Optimal parameters for the GR and SSGR schemes, when n = {3, 5, 7}, δ_{opt} = {0.5, 1.0, 1.5} and ARL_0 = {370, 500}

n	δ	$ARL_0 = 370$		$ARL_0 = 500$	
		GR	SSGR	GR	SSGR
		(k, L_G)	(k, L_{SSG})	(k, L_G)	(k, L_{SSG})
3	0.5	(2.2974, 20)	(2.1573, 15)	(2.3724, 23)	(2.2308, 17)
	1.0	(1.9510, 5)	(1.8658, 5)	(2.0452, 6)	(1.9114, 5)
	1.5	(1.8105, 3)	(1.5951, 2)	(1.8577, 3)	(1.7658, 3)
5	0.5	(2.1759, 12)	(2.0536, 10)	(2.2568, 14)	(2.1217, 11)
	1.0	(1.8105, 3)	(1.7183, 3)	(1.8577, 3)	(1.7658, 3)
	1.5	(1.6928, 2)	(1.5951, 2)	(1.7414, 2)	(1.6444, 2)
7	0.5	(2.1045, 9)	(1.9587, 7)	(2.1749, 10)	(2.0388, 8)
	1.0	(1.8105, 3)	(1.5951, 2)	(1.8577, 3)	(1.6444, 2)
	1.5	(1.6928, 2)	(1.5951, 2)	(1.7414, 2)	(1.6444, 2)

Table 2. *SDTSs* for the GR and SSGR schemes, when n = 3, δ_{opt} = {0.5, 1.0, 1.5} and ARL_0 = {370, 500}

δ	(δ_{opt}, ARL_0)											
	{0.5, 370}		{1.0, 370}		{1.5, 370}		{0.5, 500}		{1.0, 500}		{1.5, 500}	
	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$
0.00	1540.00	1570.26	1375.10	1476.97	1377.98	1390.78	2057.31	2113.99	1833.54	1925.66	1788.88	1853.95
0.25	510.11	412.88	546.54	432.30	560.93	477.01	640.15	510.75	680.58	540.84	714.11	560.17
0.50	94.24	73.14	109.95	81.30	123.81	98.63	108.46	86.32	126.28	96.19	146.33	105.61
0.75	22.58	19.03	28.26	22.33	32.23	28.88	25.95	20.86	30.81	25.46	37.66	28.84
1.00	9.31	7.82	9.83	8.20	11.55	10.87	10.30	8.60	10.61	9.01	12.77	10.39
1.50	2.97	2.58	2.07	1.90	2.28	2.40	3.26	2.78	2.27	1.99	2.43	2.07
2.00	1.17	1.00	0.81	0.74	0.73	0.65	1.27	1.08	0.90	0.79	0.76	0.68
2.50	0.43	0.36	0.28	0.24	0.22	0.16	0.48	0.40	0.32	0.27	0.24	0.21
3.00	0.12	0.09	0.06	0.05	0.04	0.04	0.14	0.11	0.07	0.05	0.05	0.04

Table 3. *SDTSs* for the GR and SSGR schemes, when n = 5, δ_{opt} = {0.5, 1.0, 1.5} and ARL_0 = {370, 500}

δ	(δ_{opt}, ARL_0)											
	{0.5, 370}		{1.0, 370}		{1.5, 370}		{0.5, 500}		{1.0, 500}		{1.5, 500}	
	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$
0.00	2405.43	2536.49	2270.14	2337.24	2209.19	2276.44	3266.60	3348.17	3000.57	3075.42	2956.43	3050.31
0.25	515.33	403.13	635.08	475.42	669.32	499.80	633.17	484.73	775.54	585.32	825.54	623.09
0.50	67.59	53.90	89.97	69.82	103.67	80.22	76.37	62.43	105.40	79.20	120.47	92.19
0.75	15.01	12.91	20.60	17.05	24.45	20.24	16.62	14.32	22.77	18.83	27.39	22.29
1.00	6.58	5.71	7.20	5.95	8.50	7.22	7.20	6.21	7.89	6.65	9.28	7.86
1.50	1.98	1.76	1.38	1.27	1.51	1.21	2.10	1.88	1.44	1.32	1.69	1.32
2.00	0.49	0.39	0.26	0.21	0.20	0.19	0.56	0.44	0.28	0.24	0.23	0.19
2.50	0.07	0.07	0.05	0.05	0.00	0.00	0.09	0.07	0.05	0.05	0.05	0.00
3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. *SDTSs* for the GR and SSGR schemes, when n = 7, δ_{opt} = {0.5, 1.0, 1.5} and ARL_0 = {370, 500}

δ	(δ_{opt}, ARL_0)											
	{0.5, 370}		{1.0, 370}		{1.5, 370}		{0.5, 500}		{1.0, 500}		{1.5, 500}	
	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$	$SDTS_G$	$SDTS_{SSG}$
0.00	3329.10	3477.86	3105.92	3171.58	3106.55	3171.58	4417.75	4588.55	4119.44	4236.85	4102.93	4236.85
0.25	505.21	388.23	601.91	486.61	652.50	486.61	609.30	465.20	735.85	589.43	811.63	589.43
0.50	53.73	46.29	69.25	63.02	81.84	63.02	58.41	50.30	79.38	72.08	94.03	72.08
0.75	11.84	10.56	16.17	15.45	18.58	15.45	13.02	11.22	17.90	16.93	20.20	16.93
1.00	5.39	4.72	5.04	5.21	6.24	5.21	5.76	2.06	5.36	5.72	6.61	5.72
1.50	1.24	1.04	0.85	0.74	0.81	0.74	1.34	1.15	0.91	0.75	0.91	0.75
2.00	0.16	0.14	0.07	0.07	0.07	0.07	0.17	0.16	0.07	0.07	0.07	0.07
2.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

4.2 Performance comparison between the GR and SSGR schemes (OC $SDTS$)

In this subsection, we compare the performance of the OC $SDTS$ ($SDTS_1$) between the GR and SSGR schemes, i.e. case δ>0 in Tables 2 to 4. From Tables 2 to 4, we notice that when δ < 2.0, the SSGR scheme has consistently lower $SDTS_1$ compared to the GR scheme except for two cases, i.e. (n, δ_{opt}, ARL_0) = (7, 1.0, 370) and (7, 1.0, 500) when δ = 1.0; whereas when δ > 2.0, the SSGR scheme has either slightly lower or almost the same $SDTS_1$ compared to the GR scheme. Therefore, we conclude that the SSGR scheme has a superior $SDTS_1$ performance compared to GR scheme. The $SDTS_1$ result, complemented with the ATS_1 result in Yew et al. (2016), showed that the SSGR scheme prevails over the GR scheme in terms of ATS_1 and $SDTS_1$.

5. Conclusions

In this paper, we perform an in-depth SDTS performance comparison between the GR and SSGR schemes. We utilize the ScicosLab software to compute the optimal parameters of the GR and SSGR schemes by minimizing the ATS_1. Then, we use the SAS software to calculate the SDTS of the GR and SSGR schemes for various mean shift sizes δ. From the performance comparison, we

conclude that the GR scheme slightly surpasses the SSGR scheme in terms of $SDTS_0$ performance. However, the opposite is true for the $SDTS_1$, where the SSGR scheme outperforms the GR scheme. This indicates that although practitioners are encouraged to use the SSGR scheme due to its superior $SDTS_1$ performance, they should not ignore the fact that the performance of the SSGR scheme is slightly inferior to the GR scheme in terms of $SDTS_0$ and it may lead to a higher early false alarm than the GR scheme. Future research can compare the performance of the GR and SSGR schemes in terms of SDTS by minimizing the OC median time to signal (MTS_1) instead of the ATS_1, since the median is less affected by outliers compared to the average.

References

1. Antzoulakos, D.L. & Rakitzis, A.C. (2008). The revised m-of-k runs rule. Qual. Eng., 20, 75–81. Chong, Z.L., Khoo, M.B.C., Lee, M.H. & Chen, C.H. (2017). Group runs revised m-of-k runs rule control chart. Commun. Stat. Theory Methods, 46, 6916–6935.
2. Chong, Z.L., Khoo, M.B.C., Teoh, W.L., Yeong, W.C. & Lim, S.L. (2019). Optimal design of the modified group runs (MGR) X chart when process parameters are estimated. Commun. Stat. Simul. Comput., to be published.
3. Gadre, M.P. & Rattihalli, R.N. (2004). A group runs control chart for detecting shifts in the process mean. Econ. Qual. Contr., 19, 29−43.

4. Gadre, M.P. & Rattihalli, R.N. (2006). Modified group runs control charts to detect increases in fraction non conforming and shifts in the process mean. Commun. Stat. Simul. Comput., 35, 225– 240.
5. Gadre, M.P. & Rattihalli, R.N. (2007). A side sensitive group runs control chart for detecting shifts in the process mean. Stat. Methods Appl., 16, 27–37.
6. Gadre, M.P., Joshi, K.A. & Rattihalli, R.N. (2010). A side sensitive modified group runs control chart to detect shifts in the process mean. J. Appl. Stat., 37, 2073–2087.
7. Lee, L.Y., Khoo, M.B.C. & Yap, E.Y. (2013). A comparison between the standard deviation of the run length (SDRL) performance of optimal EWMA and optimal CUSUM charts. J. Qual. Meas. Anal., 9, 1–8.
8. Lim, S.L., Khoo, M.B.C, Yeong, W.C. & Lee, M.H. (2015). Economic and economic-statisticaldesigns of the side sensitive group runs chart. Comput. Ind. Eng., 90, 314–325.
9. Page, E.S. (1954). Continuous inspection schemes. Biometrika, 41, 100–115.
10. Roberts, S.W. (1959). Control chart tests based on geometric moving averages. Technometrics, 1, 239–250.
11. Wu, Z. & Spedding, T.A. (2000). A synthetic control chart for detecting small shifts in the process mean. J. Qual. Technol., 32, 32−38.
12. Yew, S.Y., Khoo, M.B.C., Teoh, W.L., The, S.Y. & Yeong, W.C. (2016). A comparative study of the group runs and side sensitive group runs control charts. Pertanika J. Sci. Technol., 24, 177–189.
13. You, H.W., Khoo, M.B.C., Castagliola, P. & Ou, Y. (2015). Side sensitive group runs X chart with estimated process parameters. Comput. Stat., 30, 1245–1278.

Shewhart monitoring schemes with supplementary side-sensitive runs-rules for the Burr-type XII distribution

J.-C. Malela-Majika, S.K. Malandala, M.A. Graham
University of Pretoria, South Africa

Abstract

Nonparametric or distribution-free control charts are highly desirable since a minimal set of modeling assumptions are necessary for their implementation. The traditional Shewhart monitoring scheme has been improved upon using several techniques which include, amongst other, adding side-sensitive (SS) and non-side-sensitive (NSS) runs-rules to them. It has been shown, in the literature, that SS runs-rules outperform NSS schemes. Accordingly, here a SS Shewhart-type $\bar{X}$ monitoring scheme, supplemented with the 2-*of*-$(h+1)$ standard and improved runs-rules (where h is a non-zero positive integer) for non-normal data is proposed. A Markov chain approach is used to investigate the zero-and steady-state performances. It is found that the proposed schemes outperform many existing schemes. A summary and some concluding remarks are given.

Keywords

Markov chain approach; side-sensitive schemes; steady-state performance; zero-state performance

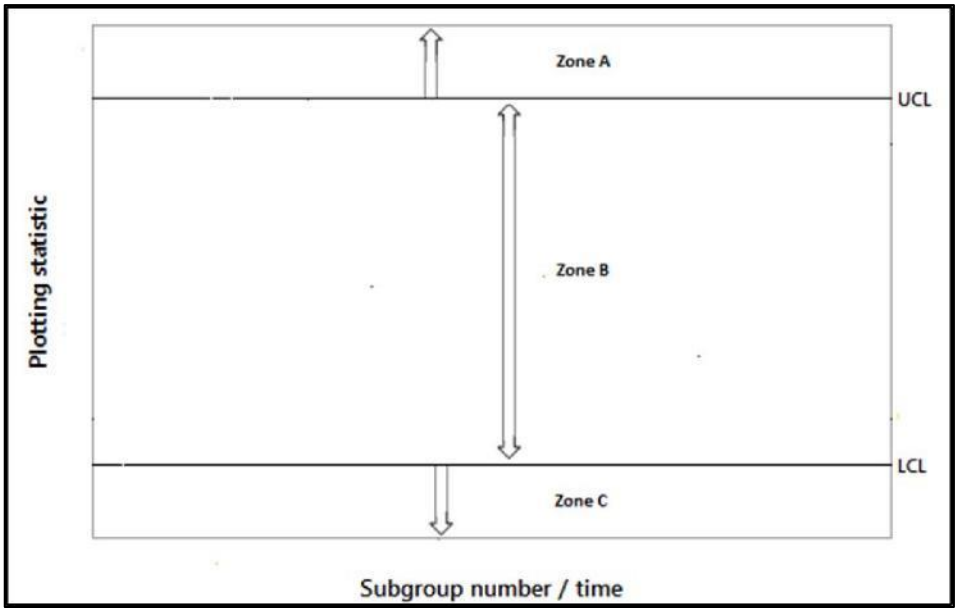
1. Introduction

In this paper it is assumed that the reader is familiar with the basics of control charting, e.g. the construction of the basic Shewhart control chart, the choice of which statistic to be plotted on the control chart (i.e. the choice of charting or plotting statistic), the size of the shift to be detected, the sample size, the frequency of sampling, the monitoring of location and/or spread, metrics used to evaluate control chart performance, an in-control process vs. and out-of-control process etc. All these issues are important as they need to be addressed before a control chart can be implemented. In the statistical process control and monitoring (SPCM) literature it is well-known that the basic 1-*of*-1 scheme (denoted RR1-*of*-1 for our purposes) has control limits $UCL/LCL = \mu_0 \pm k\sigma_0$ where k is the distance between the centerline (CL=μ_0) and the control limits in standard deviation units, LCL and UCL are the lower and upper control limits, respectively, and μ_0 and σ_0 are some known in-control (IC) process mean and standard deviation, respectively. This scheme signals when one plotting statistic (sample mean, $\bar{X}$) plots on or

beyond either of the control limits (see Figure 1). Typically, k is found such that some nominal in-control average run-length (ARL_0) is attained.

Figure 1. Zones for the basic scheme and the 2-*of*-(*h*+1) scheme

To increase the sensitivity of control charts, warning limits have been introduced (see, for example, Champ and Woodall (1987), which has control and warning limits



$$UCL/LCL = \mu_0 \pm k_2\sigma_0$$

and

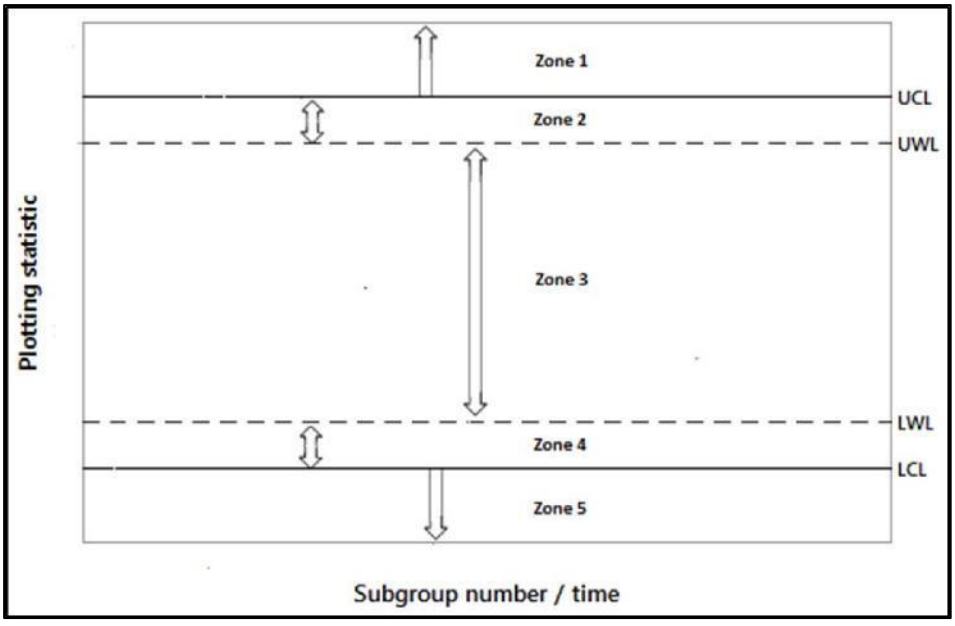
$$UWL/LWL = \mu_0 \pm k_1\sigma_0,$$

respectively, where LWL and UWL are the lower and upper warning limits, respectively, k_1 and k_2 are the distances between the centerline ($CL = \mu_0$) and the warning and control limits, respectively, in standard deviation units, with $k_2 > k_1 > 0$. Typically, k_1 and k_2 are found such that some nominal in-control average run-length (ARL_0) is attained.

Khoo (2003) proposed the $m - of - k$ standard runs-rules (hereafter $SRR_{m\text{-}of\text{-}k}$) schemes which signal when m out of k ($m > 1$ and $k \geq m$) consecutive plotting statistics plot on or beyond the control limits shown in Figure 1. Improving the work of Khoo (2003), Khoo and Ariffin (2006) and Acosta-Mejia (2007) created improved runs-rules (IRR), denoted $IRR_{m\text{-}of\text{-}k}$ here, by combining the basic scheme and the $SRR_{m\text{-}of\text{-}k}$ scheme. This scheme signals when either one plotting statistic falls beyond the control limits (Zone A or E) or when m out of k consecutive plotting statistics plot between the

warning and control limits, regardless of whether some (or all) of the m samples fall in Zone B and others (or all) fall in Zone D (see Figure 2).

Figure 2. Zones for 1-*of*-1 scheme or 2-*of*-(*h*+1) scheme



More recent contributions to the literature, regarding runs-rules, are summarized in Table 1.

Table 1. Recent runs-rules contributions to the literature

Authors	**Year**	**NSS or SS**	**Normal or non-normal**	**Detail**
Shongwe & Graham	2016	NSS & SS	Normal	Investigated NSS and various SS runs-rules and synthetic $\bar{X}$ schemes for normally distributed data
Malela-Majika, Kanyama & Rapoo	2017	NSS	Non-normal	Proposed NSS $SRR_{2\text{-}of\text{-}(h+1)}$ and $IRR_{2\text{-}of\text{-}(h+1)}$ $\bar{X}$ schemes for non-normal data using the Burr-type XII distribution
Malela-Majika, Malandala & Graham	2019	SS	Non-normal	Proposed SS $SRR_{2\text{-}of\text{-}(h+1)}$ and $IRR_{2\text{-}of\text{-}(h+1)}$ Shewhart-type $\bar{X}$ schemes for non-normal data

To differentiate between non-side-sensitive (NSS) and side-sensitive (SS) chart, let's start with the latter. The SS $SRR_{2\text{-}of\text{-}(h+1)}$ schemes signal when two (out of $h+1$) consecutive plotting statistics plot in Zone 1 (or Zone 3), which are separated by at most $h-1$ plotting statistics that plot in Zone 2, whereas for NSS schemes signal whether some (or both) plotting statistics fall in Zone 1 and others (or both) in Zone 3 (see Figure 1). The probabilities of the plotting statistic plotting in Zones 1, 2 and 3, respectively, can easily be computed, but the details are omitted here for conciseness. Only focusing on SS from this point forward, the $IRR_{2\text{-}of\text{-}(h+1)}$ schemes signal when either a single plotting statistics plots in Zone A (or Zone E) or when 2 (out of $h+1$) consecutive plotting statistics plot in Zone B (or Zone D), which are separated by at most $h-1$ plotting statistics that plot in Zone C (see Figure 2). The probabilities of a plotting statistic plotting in Zones A, B, C, D and E, denoted, respectively, can easily be computed, but the details are omitted here for conciseness.

From Table 1 it can be seen that the contribution of this paper is that side-sensitive $SRR_{2\text{-}of\text{-}(h+1)}$ and $IRR_{2\text{-}of\text{-}(h+1)}$ Shewhart-type $\bar{X}$ schemes for non-normal data are proposed (when the normality assumption fails to hold) as alternative to the traditional (parametric) side-sensitive $SRR_{2\text{-}of\text{-}(h+1)}$ and $IRR_{2\text{-}of\text{-}(h+1)}$ Shewhart-type $\bar{X}$ schemes.

The remainder of this paper is structured as follows. Section 2 the design of the proposed schemes under the Burr-type XII distribution is given. In Section 3 the zero-state and steady-state ARL and average extra quadratic loss ($AEQL$) performance measures are thoroughly investigated using the Markov chain approach and the BTXII $\bar{X}$control schemes are compared with traditional $\bar{X}$ control schemes under zero-state and steady-state modes. A discussion is provided in Section 4 along with some concluding remarks.

2. Design of the proposed schemes under the Burr-type XII distribution

The reader is referred to Burr (1942) for details on the Burr-type XII (BTXII) distribution; details are omitted here to conserve space. Here we simply mention the advantages of using the BTXII distribution. Note that the BTXII distribution is used to describe the non-normal probability density function of the IC process. Advantages of this distribution include the simplicity of its cumulative distribution function as well as the option of representing a number of different unimodal distributions. As a result, calculating Type I and Type II errors are easy and the closed-form of the run-length distribution, of control charting techniques designed under the BTXII distribution, are easy to obtain. This paper considers the SS 2-of-$(h+1)$ and 1-of-1 or 2-of-$(h+1)$ schemes to expand the Shewhart-type $\bar{X}$ scheme for non-normal distributed data using the BTXII distribution under the assumption that the process parameters are known (Case K).

The zero-state and steady-state performances are investigated using the Markov chain approach. A control chart is typically evaluated using either the zero-state or the steady-state run-length properties. The former is used to characterize short term run-length properties of a monitoring scheme as the zero-state run-length is the number of plotting statistics at which the chart first signals given it begins in some specific initial state and it is assumed that the mean shift always takes place at the beginning of the process (Zhang and Wu, 2005). The steady-state mode is used to characterize long term run-length properties of a monitoring scheme as the steady-state run-length is the number of plotting statistics at which the chart first signals given that the process begins and stays IC for a long time, then at some random time, a signal is observed (Zhang and Wu, 2005). Although a Markov chain approach is used here, the details, such as setting up the transition probability matrices etc., are omitted here to conserve space. The reader is referred to Fu and Lou (2003) and Shongwe and Graham (2016) for details on the Markov chain approach and SPCM.

3. Results

The zero-state and steady-state ARL and $AEQL$ performance measures are thoroughly investigated using the Markov chain approach. Both these measures have been widely used in the SPCM literature (see Human and Graham (2007) for a discussion on the ARL and Ou, Chen and Khoo (2015) for a discussion on the $AEQL$). Both were considered as the ARL measures the control chart's performance for specific shifts, whereas the $AEQL$ measures the overall performance of the control chart. Only the latter is shown here to conserve space. The proposed methods are compared to the traditional methods and the chart with the lower $AEQL$ performs best. It was found that for the $SRR_{2\text{-}of\text{-}(h+1)}$ scheme, the proposed control schemes perform better than the traditional control schemes for any h value regardless of the sample size for both zero-state and steady-state modes. For the $IRR_{2\text{-}of\text{-}(h+1)}$ schemes, for both zero-state and steady-state modes, the proposed BTXII control schemes are slightly more sensitive than the $IRR_{2\text{-}of\text{-}(h+1)}$ traditional control schemes (see Figures 3 and 4).

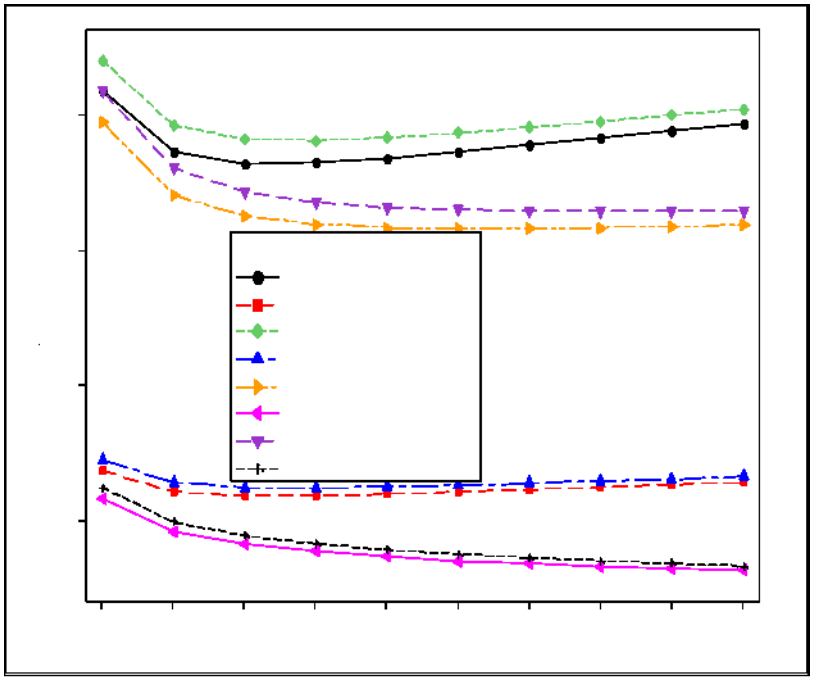


Figure 3. The BTXII X control schemes versus the traditional X control schemes ($SRR_{2\text{-}of\text{-}(h+1)}$ schemes for different values of h) under zero-state and steady-state modes when n = 5 and 10

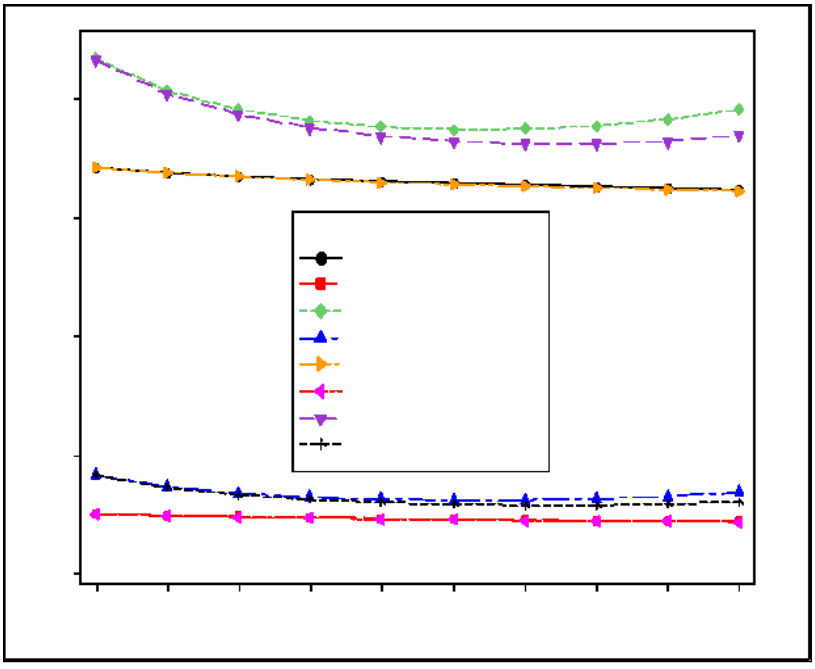


Figure 4. The BTXII X control schemes versus the traditional X control schemes ($IRR_{2\text{-}of\text{-}(h+1)}$ schemes for different values of h) under zero-state and steady-state modes when n = 5 and 10

4. Discussion and Conclusion

We considered the SS 2-*of*-$(h+1)$ and 1-*of*-1 or 2-*of*-$(h+1)$ schemes to expand the Shewhart-type $\bar{X}$ scheme for non-normal distributed data using the BTXII distribution under the assumption that the process parameters are known (Case K). The zero-state and steady-state performances are investigated using the Markov chain approach.

It was observed that the proposed schemes outperform the traditional ones, and present very interesting run-length characteristics under the normal and non-normal distributions. Moreover, the proposed side-sensitive schemes outperform the NSS schemes proposed by Malela-Majika, Kanyama and Rapoo (2017). For small to moderate shift we recommend using the SS $SRR_{2\text{-}of\text{-}(h+1)}$ schemes and for large shifts we recommend using the SS $IRR_{2\text{-}of\text{-}(h+1)}$ schemes. The case where the process parameters are unknown and need to be estimated are currently under investigation.

5. Acknowledgement

Marien Graham's research was funded by the National Research Foundation (NRF) [reference: PR_IFR190111407337, UID: 114814]

References

1. Acosta-Mejia, C.A. (2007). "Two sets of runs rules for the X chart." *Quality Engineering*, *19*(2), 129–136.
2. Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, *13*(2), 215-232.
3. Champ, C.W. and Woodall, W.H. (1987). "Exact results for Shewhart control charts with supplementary runs rules." *Technometrics*, *29*(4), 393-399.
4. Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and Its Applications: A Finite Markov Chain Imbedding Approach.* Singapore: World Scientific Publishing.
5. Human, S.W. and Graham, M.A. (2007). "Average run lengths and operating characteristic curves." *Encyclopedia of Statistics in Quality and Reliability,* 1, 159-168, John Wiley, New York.
6. Khoo, M.B.C. (2003). "Design of runs rules schemes." *Quality Engineering*, *16*(1), 27-43.
7. Khoo, M.B.C. and Ariffin, K.N. (2006). "Two improved runs rules for the Shewhart $\bar{X}$ control chart." *Quality Engineering*, *18*(2), 173-178.
8. Malela-Majika J.C., Kanyama, B.J. and Rapoo, E.M. (2017). "Improved Shewhart-type $\bar{X}$ control schemes under non-normality assumption: A Markov chain approach." *International Journal of Quality Research*, *12*(1), 17-42.
9. Malela-Majika, J.C., Malandala, S.K. and Graham, M.A. (2019). "Shewhart $\bar{X}$ control schemes with supplementary 2-*of*-$(h+1)$ side-sensitive runs-rules under the Burr-type XII distribution." To appear in *Quality and Reliability Engineering International*
10. Ou, Y., Chen, N. and Khoo, M.B.C. (2015). "An efficient multivariate control charting mechanism based on SPRT." *International Journal of Production Research*, *53*(7), 1937-1949.
11. Shongwe, S.C. and Graham, M.A. (2016). "On the performance of Shewhart-type synthetic and runs-rules charts combined with an $\bar{X}$ chart." *Quality and Reliability Engineering International, 32*(4), 1357-1379.
12. Zhang, S. & Wu, Z. (2005). Designs of control charts with supplementary runs rules. *Computers & Industrial Engineering*, *49*(1), 76-97.

Response of the National Institute of Statistics and Geography (INEGI) to natural disasters in Mexico

Arturo Blancas Espejo*
Instituto Nacional de Estadística y Geografía (INEGI)
Ciudad de México, México

Abstract

The National Institute of Statistics and Geography (INEGI-Mexico), in compliance with the Law of the National System of Statistical and Geographic Information, has undertaken several actions related to the prevention and attention of emergencies or catastrophes caused by natural disasters. The actions carried out by INEGI in statistical matters, have focused mainly on designing and generating information on companies, in order to contribute to making timely decisions when a natural contingency occurs. In this context, the following three strategies stand out: *a) Design and update of the master sample of enterprises:* Generated and updated permanently from the Statistical Business Register of Mexico, conformed by statistical designs with diverse cuts: sectorial, company size and geographic, whose primary objective is the immediate obtaining of samples with specific characteristics for the realization of specific surveys that provide pertinent and expedited statistics to give attention to natural disasters that occur in our country. *b) Conducting Specific Surveys:* If required, the surveys are designed based on the master sample, with the aim of generating information on the possible effects that enterprises have suffered from a natural disaster. As an example of the scope of INEGI, we have the Survey on the Affectations of the earthquakes of September 2017, which was carried out as an immediate response to the earthquakes of September 7 and 19 of that year to quantify the affectations that originated this natural phenomenon, thus contributing to the attention actions in favor of the companies whose economic activity was affected. *c) Collaborative Site for Attention to Disasters (by its acronym in Spanish SICADE):* Allows users to consult, download and / or add statistical and geographic information for the analysis and development of strategies in the prevention and response to natural disasters. This site shows direct links to different dependencies in Mexico that have information related to natural disasters in different areas. The document called: "The response of the National Institute of Statistics and Geography (INEGI) to natural disasters in Mexico", will describe more ampleness the actions in statistical and geographic matters related to the prevention and attention of natural disasters that INEGI has headed. The foregoing, with the objective of sharing the Mexican experience with the National Statistical Offices of the world, contributing to a permanent exchange of good practices in this matter, as well

as for the incorporation of this important topic into the global statistical agenda.

Keywords
Affectations, strategies, prevention, risks, accidents.

1. Introduction

The National Institute of Statistics and Geography (INEGI-Mexico), has its legal framework in the Law of the National System of Statistical and Geographical Information, which empowers it to carry out many actions related to the prevention and attention of emergencies or catastrophes caused by natural disasters.

The actions carried out by INEGI have focused mainly on designing and generating information on companies, highlighting three strategies: i) Design and update of the master sample of companies; ii) Conducting specific surveys, and iii) Instrumentation of the Collaborative Site for Disaster Assistance.

In the development of this great statistical effort, many ministries and administrative units have a relevant participation related to the subject, corresponding to INEGI the important task of coordinating and regulating the required statistical and geographic activities.

In this document, details of the actions in statistical matters related to the prevention and attention of natural disasters that INEGI has headed are shown. Its objective is to share the Mexican experience with the National Statistical Offices of the world, to favor a permanent exchange of good practices that allow the authorities of the countries to better face the challenges posed by the risks derived from natural disasters.

2. Legal framework that governs statistical and geographical activities

In April of 2018, the Law of the National System of Statistical and Geographic Information (by its acronym in Spanish LSNIEG)[1] was published, through which the National System of Statistical and Geographic Information (by its acronym in Spanish SNIEG) was created, whose main objective is to supply society and to the State information of quality, pertinent, truthful and opportune, in order to contribute to the national development. Its guiding principles are those of accessibility, transparency, objectivity and independence.

INEGI, within the framework of the LSNIEG, is an autonomous public body of the Mexican State, responsible for regulating and coordinating the SNIEG. Also, in accordance with the principles governing the System, the INEGI carries

[1] https://www.snieg.mx/contenidos/espanol/normatividad/marcojuridico/LSNIEG.pdf

out the actions tending to achieve i) the conceptual adaptation of the Information of National Interest to the needs that the economic and social development of the country imposes, ii) that the information is comparable in time and space, and iii) the adaptation of statistical and geographical procedures to international standards, to facilitate comparison.

The Law itself recognizes the importance of statistical and geographical activities related to natural disasters. In particular, section II of article 64 establishes that "*after complying with the corresponding legal and administrative formalities and favorable agreement of its Governing Board, the Institute shall provide the support requested by the Federal Executive in case information is required to prevent and, where appropriate, respond to emergencies or catastrophes caused by natural disasters*".

For its part, in the last paragraph of Article 78 of the LSNIEG, an explicit mention is made of the nature of such information, since it indicates that *"... it may also be considered as Information of National Interest that is necessary to prevent and, in his case, attend emergencies or catastrophes caused by natural disasters ... "*.

In Mexico, there is a specific legal framework that empowers INEGI to develop statistical and geographic activities that contribute to the prevention and attention of contingencies caused by natural disasters.

3. Actions in statistical matter

INEGI's efforts have focused on the design and implementation of strategies related to the generation of statistical and geographic information, whose fundamental purpose is to contribute to timely decision-making (before, during and after natural contingencies have occurred). In this sense, the actions in statistical matter carried out so far, focus mainly on the design, generation and dissemination of economic information of companies in our country.

It should be noted that these actions could not be carried out by any other Institution, since it implies taking advantage of the extensive statistical infrastructure available to the INEGI, which serve as a basis to offer an immediate response to natural disasters that may occur throughout the year in all the national territory.

Among the most important strategies, the following stand out.

a) Design and update of the master sample of companies.

The main purpose of the master sample is to integrate a set of company's representative of the national context, as well as to carry out the permanent updating of the location and location data of the companies based on the Statistical Business Register of Mexico, in order to obtain timely samples that serve for the realization of specific surveys whose theme focuses on the effects caused by natural disasters.

In particular, the master sample integrates the set of companies that have a telephone number and / or web site, which according to the civil protection authorities are located in the most vulnerable geographical regions affected by a loss of character natural, whether earthquakes, hurricanes and / or floods. The conformation of this sample is based on a general design that starts from two levels of aggregation in which six particular statistical cuts are considered.
A. At the regional level:

- By company size
- By state
- By large sector
- By sector

B. At the federative level:

- By company size
- By large sector

For each possible scenario, the most vulnerable federal entities are analyzed in the face of the onslaught of natural phenomena, which will allow to capture the information of the entire affected group at a regional or federative level, considering the statistical cuts of interest of between the set immersed in the design and conformation of the master sample.

For the groupings by large sector and sector, the North American Industrial Classification System (NAICS) 2013 is used.

The sampling scheme is probabilistic and stratified. Stratification is carried out using the variable of total occupied personnel, according to the segments published by the competent national authority.

Table number 1

Stratification according to the total number of occupied personnel

Company size	Occupied personnel range		
	Industry (Construction and Manufactures)	Trade	Services
Large	251 and more	101 and more	101 and more
Medium	51 - 250	31 - 100	51 - 100
Small	11 - 50	11 - 30	11 - 50
Micro	0 - 10	0 - 10	0 - 10

It is worth mentioning that the inclusion of companies in the master sample is done by including the set of establishments that comprise them; this is due to the fact that the specific studies require addressing the affected regions and not the regions where the corporate or corporate matrices are located.

The size of the master sample is proposed independently for each statistical cut, using the expression to estimate a proportion of 60.0%, 65.0% and 70.0%, with a confidence level of 95.0%, relative error of 10.0% and a rate of No response of 15.0 percent.

Having such statistical infrastructure allows INEGI to be permanently prepared to respond immediately to the demand for basic statistical information of a qualitative nature, with the aim of contributing to the attention of the damages caused by natural disasters that occur in Mexico.

b) Conducting specific surveys.

The carrying out of specific surveys (mainly of a qualitative nature) represents one of the main strategies that contributes to decision-making in the presence of natural disasters and is closely linked to the design and updating of the master sample.

As part of this strategy, it is expected that all studies will be captured through the Computer Aided Telephone Interview Center (CATI), which has been in operation since 2011.

This strategy was implemented for the first time as part of the attention of the emergency caused by the earthquakes that occurred in our country in September 2017.

INEGI carried out the design, collection, processing and dissemination of the Survey of Seismic Impairments of September 2017, just 10 days after the earthquake of September 19 of that year, and this is the timeliest survey in companies that has been made in our country.

This survey offered qualitative information, generated based on the opinions of businessmen, with the purpose of contributing to the decision making in favor of the companies that were affected by the earthquakes that occurred in Mexico on the 7th and 19th of September 2017.

The unit of observation was the establishment and measurement focused on those that according to the Industrial Classification System of North America (NAICS) 2013 are classified in the sectors of Manufacturing Industries, Trade (both wholesale and retail minor) and private non-financial services.
The opinion of the businessmen was captured regarding: i) the effects on their facilities and the services they provide, ii) the supports or aids received, iii) the suspension of activities and the number of days of suspension, iv) the actions carried out by the establishment to collaborate in the attention of the emergency, and v) the expectation of the businessmen on the economic activity for the last quarter of that year.

The sampling scheme was probabilistic. The sampling frame was supplied by the master sample. It was integrated by little more than 2 million establishments of all sizes, which were distributed in the eight states with the greatest impact according to the civil protection authorities.

The sample was integrated by 2,350 establishments of the three sectors under study and proportional to the size of the economic units. Two replacement samples were instrumented to reduce the non-response of those establishments difficult to contact.

He considered a confidence level of 95.0%, a relative error of 10.4%, and an expected non-response rate of 20.0%. The results obtained from this sample were representative of the set of establishments that made up the sampling frame.

The telephone calls to capture the information were made on September 25, 26 and 27 of 2017, reaching 81.2% of the total sample collection.

The society in general recognized the valuable work of INEGI in providing timely statistical information at the time of the emergency. All the information generated by said survey was included in a press release that was disseminated at 8:00 am on September 29, 2017, through the INEGI website.

c) Collaborative Site for Disaster Assistance.

The Collaborative Site for Attention to Disasters (by its acronym in Spanish SICADE)[2] arises from the guiding principles of the LSNIEG. This site was developed by INEGI, who is also responsible for its permanent updating.

The SICADE makes available to user's information about natural disasters that have occurred in the country, as well as relevant data of each federal entity, such as population, number of homes, economic and geographical characteristics, among others.

The information offered by this site is not only provided by the INEGI, since as its name indicates, there are also several government agencies and research centers of the society related to the subject. Among the most prominent participants are the following: Ministry of the Interior; Marine; Energy; Environment and Natural Resources; National defense; Communications and Transportation; Agriculture and Rural Development; the company Mexican Petroleum; and the autonomous organisms National Electoral Institute and Institute of Geography of the National Autonomous University of Mexico.

Besides being a public site for consulting statistical and geographic information related to the subject of natural disasters, it allows the interaction of users who have the possibility to register, not only to download information but to feed this site.

Registered users can share information of the following types: Raster, Technical documents, Vectors, Consultation systems, Web services, Supplementary information, Integrated and tabulated data, among others.

[2] http://geoweb2.inegi.org.mx/sicade/inicio.jsp

The information is grouped or subdivided by type of event: Climatological, Geophysical, Hydrological, Climatic, Man-made, Organic, Biological, Chemical, Technological and Social-Natural.

In the presence of an accident or natural disaster, the affected federal entities are identified, with the purpose of initiating the analysis of the information already available on the site. After, the statistical and geographical information is sent to registered users on the site[3].

Table number 2
Information available in the SICADE that is provided to registered users when required

Road communication	Economic	Sociodemographic	Geographic and cartographic
▪ National road network.	▪ National dictionary of economic units, (by its acronym in Spanish DENUE).	▪ Information on censuses and population and housing surveys. ▪ National housing inventory.	▪ Hydrographic network ▪ Topographic maps in scales 1: 50,000 and ▪ 1: 20,000. ▪ Digital model of elevation. ▪ Geostatistical framework. ▪ Geostatistical and urban cartography

Depending on the type of disaster and weather conditions, different types of satellite data are available:

- Optical: for fires, earthquakes, volcanoes, rainfall and landslides.
- Radar: for rain, landslides and earthquakes, in addition to all those in which the optical image is not available due to poor atmospheric conditions.
- GNSS: to define the movements of the tectonic plates after an earthquake.

During the emergency caused by the earthquakes of September 2017, the actions expected in the SICADE were implemented as part of the solution.

4. What remains in this matter?

The INEGI response to natural disasters in Mexico, has the purpose of promoting an exchange of practices and experiences in the matter among the National Statistical Offices of the world, mainly in those countries where there are greater risks of presenting emergencies originated by natural disasters.

[3] The full statement is available at the following electronic address: http://www.beta.inegi.org.mx/app/saladeprensa/noticia.html?id=3769

This exchange will allow strengthening the statistical and geographic activities developed by the National Statistics Offices, which support the design and implementation of public policies that serve the population and the companies affected by the materialization of natural disaster risks.

Considering the international recommendations of the United Nations on this matter, foreseen in the Sendai Framework for Disaster Risk Reduction 2015-2030, it is vital that this issue be a permanent part of the global statistical agenda in order to create various forums to harness the experiences and recommendations of countries that have already gone down this way.

Responses of the Statistics Bureau of Japan to natural disasters

Masao Takahashi*[1]; Shigeru Kawasaki[2]; Hideo Umezawa[3]
[1] National Statistics Center, Tokyo, Japan
[2] Nihon University, Tokyo, Japan
[3] Statistics Bureau, Tokyo, Japan

Abstract
Natural disasters can occur anywhere in the world. Of course, it is important to prevent such disasters, if possible, or to take proactive measures for mitigating the effects of natural disasters. However, once such a disaster has occurred, the responses for recovery and restoration are the key issues to be considered. This is also applicable to the activities in official statistics. In Japan, we have been experiencing a number of natural disasters every year. In this paper, we present how we responded to natural disasters such as the Great East Japan Earthquake in 2011. The paper includes basic approaches and concepts in the responses to natural disasters, measures for continuation of regular statistical dissemination, provision of statistical information for rescue and restoration, statistical analyses on the effects of the disaster, and lessons learned from the experiences.

Keywords
Great East Japan Earthquake; Statistical Information for Rescue and Restoration; Statistical Analyses on the Effects of Disasters; Lessons Learned

1. Introduction

In the Great East Japan Earthquake, which occurred in March 2011, the total death toll amounted to 15,897, and 2,533 people are still missing (National Police Agency (2019)). It also destroyed more than 400,000 houses and buildings, and many people have been evacuated due to an accident at a nuclear power plant caused by the earthquake.

While the recovery from the Great East Japan Earthquake is still continuing, Japan has been suffering from other large natural disasters such as torrential rains (e.g., in the west part of Japan in 2014, in the Kanto and Tohoku areas in 2015 and in the west part of Japan in 2018) and earthquakes (e.g., in Kumamoto in 2016 and in Hokkaido in 2018) almost every year, and they have taken a large number of lives and forced many people to leave their homes due to damage to houses and utilities.

Efforts to prevent such natural disasters are of course necessary, but once a natural disaster occurs, it is important to respond to it for recovery and restoration. This is also true for the roles of official statistics.

The Statistics Bureau of Japan (SBJ) has been acting in the event of natural disasters by means of various measures. In this context, this paper introduces the basic ideas on countermeasures for the Great East Japan Earthquake, measures taken in the release of regular monthly statistics, compilation and provision of statistical information to support rescue and restoration, statistical analyses on the effects of the earthquake, and lessons learned and issues.

From the next chapter, we will introduce the efforts of the SBJ in response to natural disasters. We will cite the Great East Japan Earthquake in 2011 as an example, which is Japan's biggest disaster in recent years and on which countermeasures, experiences and lessons learned could be applicable to many kinds of other natural disasters.

2. Basic Ideas on Countermeasures for the Natural Disaster

In response to the Great East Japan Earthquake, the SBJ took various measures to produce and provide official statistics so that they could contribute to the restoration and recovery after the earthquake. The basic concepts in producing and providing the official statistics, which were adopted by the SBJ, are as follows:

1) Monthly basic statistics should be released as far as possible.
2) Statistical surveys, which were temporarily discontinued in the disaster area, should be promptly resumed in consideration of the actual conditions of the disaster.
3) Statistical information that is useful for reconstruction should be provided by utilizing existing statistics and new survey results, etc.

At the end of March 2011 after the earthquake, the SBJ was able to publish monthly statistics almost as normal except for the widespread areas damaged by the disaster. As it was difficult to conduct statistical surveys in the disaster-stricken areas, which accounted for about 5% of the total population of Japan, they were excluded from the published results. In that case, the SBJ provided information to users about data loss including the magnitude of the population and economy relative to the whole country. After the surveys in the disaster-stricken areas were resumed and the results for the whole country were published, the SBJ also released the results of retroactive figures for the past period when incomplete results were published to ensure comparability in time series data. In addition, prior to the publication of the results, the SBJ made known the plan in advance about such special handling.

Also, to support recovery and reconstruction operations, the SBJ promptly provided statistical maps of small area statistics based on the 2009 Economic Census for Business Frame and the 2010 Population Census to local governments in the affected areas and also posted them on the SBJ website. In addition, in conjunction with publication of the monthly statistical survey

results, the SBJ analyzed the social and economic trends after the earthquake and published them with commentary.

In the following, we will introduce the measures taken after the disaster in a timeline, i.e., release of regular monthly statistics, the processing of statistical information for rescue and recovery in the damaged areas, and the analyses on the effects of the earthquake.

3. Measures taken in the Monthly Statistics

For the regular monthly statistics, the SBJ devised various measures in estimating and aggregating the statistics keeping the release schedule to avoid inconvenience to the statistical users as far as possible.

Since the earthquake occurred in March 2011, it became extremely difficult to conduct various monthly statistical surveys, especially in the Pacific coastal areas of the Tohoku region. As the influence of the disaster on the statistical results was different for each survey depending on the size and the method of the sampling of the survey, the SBJ produced and disseminated the statistics by adopting an appropriate estimation method for each survey in accordance with the characteristics of the survey.

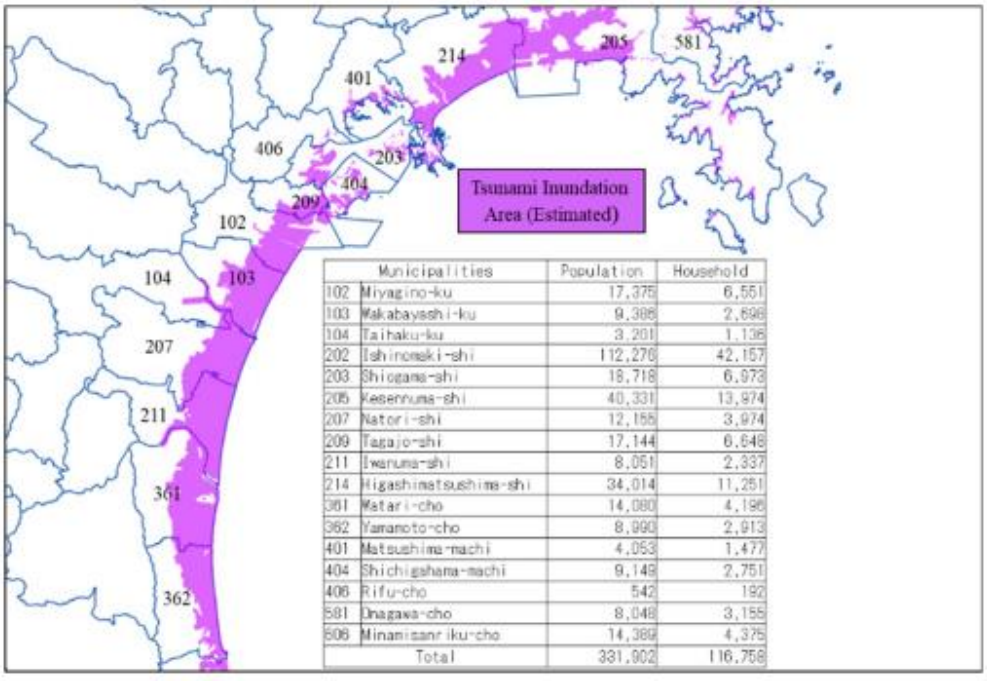
In the Labour Force Survey which was most affected by the earthquake, it was not possible to produce the statistics of “All Japan” as usual, because the survey could not be taken in the Pacific coastal region of Iwate, Miyagi and Fukushima Prefectures (hereinafter referred to as "Tohoku Three Prefectures"). After the earthquake in March 2011, the SBJ began releasing the results of "All Japan excluding Tohoku Three Prefectures" instead of "All Japan." In this connection, the statistics of "All Japan excluding Tohoku Three Prefectures" for the previous year were also produced to enable monthly comparison with the previous year on the basis of the same geographical coverage. The SBJ also published reference information such as changes in the estimation method and the extent of the impact of the earthquake in the three prefectures of Tohoku. Since September 2011, the SBJ was able to resume the survey in the Tohoku Three Prefectures, and the results of "All Japan" came to be published as before.

4. Provision of Statistical Information to Support Rescue and Restoration Operations

Immediately after the disaster, it was necessary for the government to understand the size of the damage, the extent of the affected areas, and the population size and distribution for rescue in the disaster areas. After the immediate aid operations, when the government started the planning for the renovation of the disaster-stricken areas, data describing the social and economic conditions of the region immediately before the disaster was

needed. The SBJ processed existing statistical data and provided appropriate statistics to meet such needs.

Since it was difficult to conduct statistical surveys in the disaster-stricken areas for several months after the disaster, the SBJ processed the existing statistical data in combination with geographical information in the affected areas and provided the results to local governments. For this process, the SBJ was able to obtain information on the estimated range of tsunami-flooded areas based on aerial photographs as additional geographical information. Based on this, the SBJ processed a statistical map of the population and number of households in the small area by the preliminary figures of the 2010 Population Census, which had already been made public, and began offering them to local governments in the affected areas in March 2011. In April, in order to provide them to more people, the SBJ published maps of the wide area of the six prefectures of Aomori Prefecture, Ibaraki Prefecture, and Chiba Prefecture in addition to the Tohoku Three Prefectures on the SBJ website. As an example of such maps, Figure 1 shows the map of Miyagi prefecture.



	Municipalities	Population	Household
102	Miyagino-ku	17,375	6,551
103	Wakabayashi-ku	9,386	2,696
104	Taihaku-ku	3,201	1,136
202	Ishinomaki-shi	112,276	42,157
203	Shiogama-shi	18,718	6,973
205	Kesennuma-shi	40,331	13,974
207	Natori-shi	12,155	3,974
209	Tagajo-shi	17,144	6,648
211	Iwanuma-shi	8,051	2,337
214	Higashimatsushima-shi	34,014	11,251
361	Watari-cho	14,080	4,196
362	Yamamoto-cho	8,990	2,913
401	Matsushima-machi	4,053	1,477
404	Shichigahama-machi	9,149	2,751
406	Rifu-cho	542	192
581	Onagawa-cho	8,048	3,155
606	Minamisanriku-cho	14,389	4,375
	Total	331,902	116,758

Figure 1 Outline of inundation area in Miyagi prefecture and the basic unit area population / number of households in the area

In response to requests from local governments in the affected areas, special approximate tabulation of small areas for Tohoku Three Prefectures was performed using the provisional questionnaire information for the "Basic Complete Tabulation on Population and Households" and "Basic Complete

Tabulation on Industries" for the 2010 Population Census. The results of the tabulation were released from the end of May to mid-July 2011.

In addition, to make it easier to use statistical data on disaster-stricken areas, the SBJ extracted the major statistical indices of the affected areas from the municipal statistical database of the System of Social and Demographic Statistics of Japan, and published a summary data sheet including data on the earthquake damage, on the website of the SBJ.

To make it easy for anyone to find and use this information, the SBJ set up a dedicated webpage on the SBJ website, on which information related to the Great East Japan Earthquake along with various statistical data and information on disaster areas were posted.

5. Statistical Analyses on the Effects of the Earthquake

Statistical results after the earthquake allowed us to understand the effects of the disaster on society and the economy and the situation of recovery. For this reason, the SBJ released commentaries on useful findings whenever new statistics became available. Some examples are as follows.

As the first example, the statistics of the "Report on Internal Migration in Japan" revealed that the earthquake had a significant impact on the flow of the population of the Tohoku Three Prefectures, and published a report that analyzed the characteristics of the population migration during three months and six months after the earthquake. The analysis has been continued for the region (Figure 2).

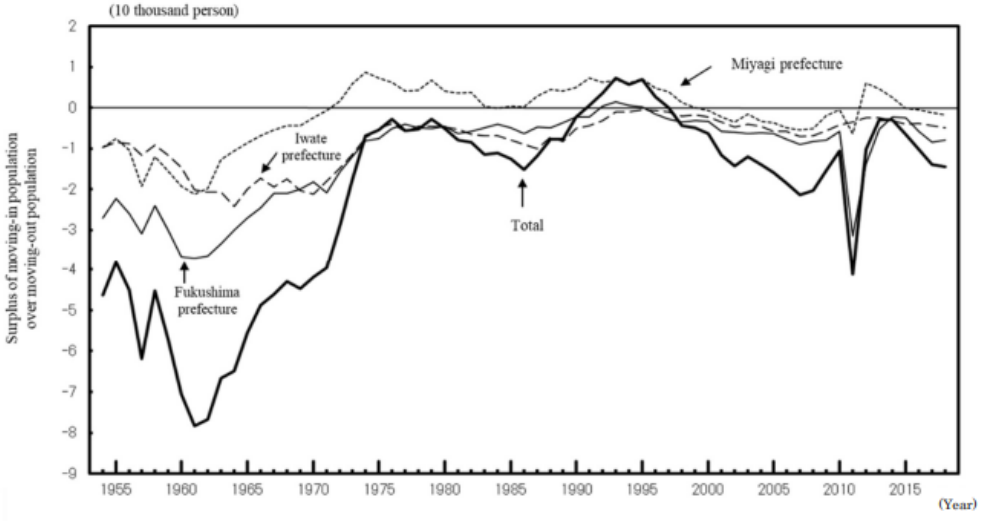


Figure 2 Surplus of incoming population over outgoing population of Iwate, Miyagi and Fukushima prefectures (Japanese migrant) (1954~2018)

The second example is based on the Economic Census conducted in 2009 and 2012. The results revealed how the number of establishments (Figure 3)

and employees (Figure 4) decreased in the municipalities in Tohoku Three Prefectures.

The third one is based on the Housing and Land Survey conducted in 2013, which revealed the influences of the Great East Japan Earthquake on housing. According to the results, the number of households that were relocated because of the earthquake was about 329,000 as of Oct. 1, 2013. Around 40% of them reported that it had become impossible for them to stay living in their original houses (Figure 5).

This kind of analysis and publication was conducted not only by the SBJ but also by the statistical organizations of other Ministries. For detailed information, please refer to the dedicated page on the SBJ website (Statistics Bureau of Japan (2019)).

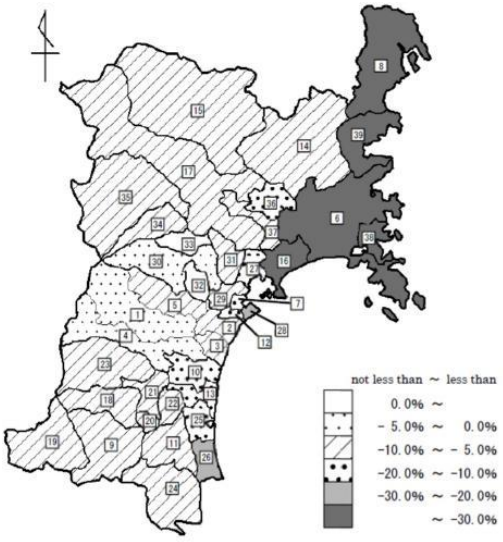


Figure 3 Rate of change in number of establishments by municipality (2009,2012)- Miyagi prefecture

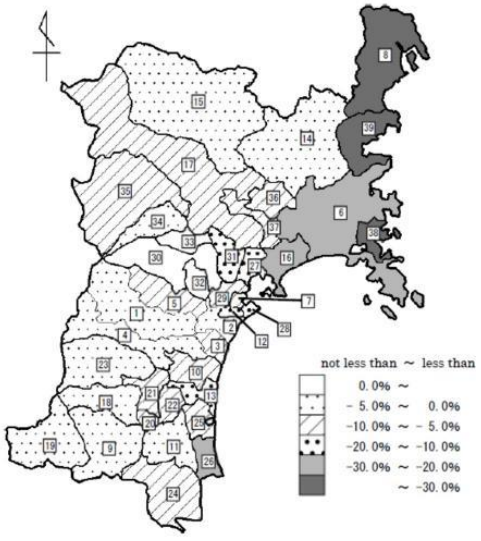


Figure 4 Rate of change in number of employees by municipality (2009,2012)- Miyagi prefecture

1 Aoba-ku	11 Kakuda	21 Murata	31 Osato
2 Miyagino-ku	12 Tagajo	22 Shibata	32 Tomiya
3 Wakabayashi-ku	13 Iwanuma	23 Kawasaki	33 Ohira
4 Taihaku-ku	14 Tome	24 Marumori	34 Shikama
5 Izumi-ku	15 Kurihara	25 Watari	35 Kami
6 Ishinomaki	16 Higashimatsushima	26 Yamamoto	36 Wakuya
7 Shiogama	17 Osaki	27 Matsushima	37 Misato
8 Kesennuma	18 Zao	28 Shichigahama	38 Onagawa
9 Shiroishi	19 Shichikashuku	29 Rifu	39 Minamisanriku
10 Natori	20 Ogawara	30 Taiwa	

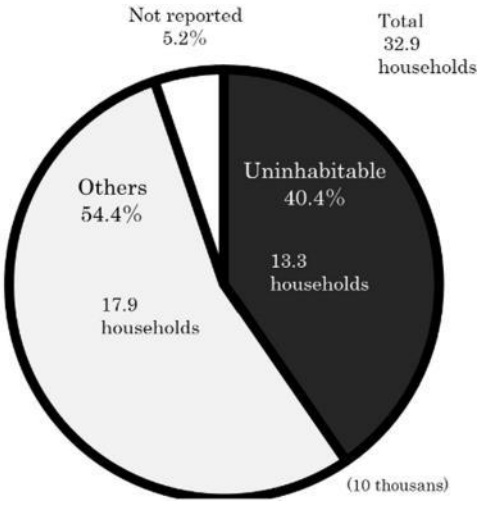


Figure 5 Ratio of ordinary households moving by reason for the moving due to the Great East Japan Earthquake of main earner – Japan (2013)

6. Lessons Learned and Challenges

From the experience of this great earthquake, we believe that there were some issues and lessons learned
about the development of official statistics. We will introduce two main points below.

The first point is that statistics play a fundamental role in decision-making even in the event of a disaster, and statistics must fulfill this mission. After the earthquake, there were many requests for the latest statistics on the situation in the affected areas. On the other hand, the disaster-stricken area is in a severe situation, and it is extremely difficult to conduct a statistical survey to grasp the actual situation. In this context, we believe that the following perspectives need to be kept in mind in order to produce and provide statistics to help support the disaster-stricken areas.

1) Processed statistics should be provided by linking existing statistical data with other relevant information such as geographical information.
2) Statistical surveys in the affected areas should be carefully carried out in consideration of the circumstances and feelings of the victims and the situation of the areas.
3) Existing statistical surveys should be utilized to collect information about the influences of the disaster and the conditions of the people in the area without significantly increasing the response burden.
4) Not only survey statistics but also information such as statistics based on administrative statistics and other information from administrative management should be widely used.

The second point is to save the various statistical data that has been maintained and accumulated in the event of a large-scale disaster as a reference for the future. Some countries seem to have good practices in this regard. For example, the U.S. Census Bureau has a webpage "Emergency Preparedness" on its website. It contains various statistical data related to large-scale disasters, such as hurricanes, earthquakes and wildfires. In Japan, we believe that it is necessary to accumulate and store various data and information collected and maintained in accordance with this so that the lessons of the Great East Japan Earthquake will not be forgotten and kept as records for the future.

7. Conclusion

According to the Statistics Act, official statistics are regarded as "foundations for making rational decisions for the citizens." In order to produce official statistics, it is essential for respondents from households and companies surveyed to respond to statistical surveys. For this reason, we have been working with many enumerators who concentrate their effort on the statistical activities. Official statistics are a valuable source of information for society, which is created by the cooperation of many people, such as those surveyed, enumerators, and related parties in local governments and statistical departments in Ministries. And the importance of these matters remains the same in normal and disaster situations. In the wake of natural disasters such as the Great East Japan Earthquake, we sincerely hope that the producers and users of official statistics will cooperate to ensure that the significance of official statistics is widely recognized and that official statistics will be more useful to the public.

References

1. National Police Agency (2019). Police Countermeasures and Damage Situation associated with 2011Tohoku district - off the Pacific Ocean Earthquake. March 8, 2019. [Online]. Available: https://www.npa.go.jp/news/other/earthquake2011/pdf/higaijokyo_e.pdf. [Accessed April 11, 2019].
2. Statistics Bureau of Japan (2019). Information on the Great East Japan Earthquake. [Online]. Available: https://www.stat.go.jp/info/shinsai/index.html. (In Japanese only). [Accessed April 11, 2019]

The development of disaster statistics in the Philippines: Expenditure accounts for disaster risk reduction[1]

Vivian R. Ilarina[2]
Macroeconomic Accounts Service (MAS),
Sectoral Statistics Office of the Philippine Statistics Authority (PSA).

Abstract

The Philippines is identified as one of the most vulnerable to different types of disasters like typhoons, earthquakes, flood, among others. According to the World Risk Index, the Philippines ranks third among countries with the highest disaster risk levels. The report cites the archipelago's location in the typhoon and earthquake belts, high degree of exposure and vulnerability and the lack of capacities in coping capacities. For disaster risk reduction and management primarily to protect human lives and infrastructure as well as conservation of the environment and natural resources, it is crucial that we generate a comprehensive disaster statistics as guide/tool among planners and policy makers for effective disaster management. However, disaster statistics is seen as relatively new statistical domain. This cut across several disciplines and needs an integrated framework to coordinate and consolidate into one comprehensive system. In the 3rd meeting in 2015, the Asia-Pacific Expert Group on Disaster-related Statistics decided to undertake a pilot test for a provisional outline and summary of core principles on disaster related statistics. This supports the mandate of the Expert Group and the requirements for monitoring progress towards the achievement of the Sendai Framework. In line with this, the Philippine Statistics Authority (PSA), the highest policy making body on statistics, deemed it important to develop an integrated framework for disaster statistics catering the needs of several stakeholders as well as the requirements in the attainment of the Sustainable Development Goals (SDGs). The PSA, as the agency tasked for the compilation of national accounts and the environmental accounts has utilized the existing frameworks of the System of National Accounts (SNA) and System of Environmental-Economic Accounting (SEEA) to supplement the development of the disaster statistics, specifically towards the organization of data for the Disaster Expenditure Accounts – a satellite accounts of the SNA.

[1] Paper presented during the 62nd ISI World Statistics Congress 2019 held in Kuala Lumpur, Malaysia on August 18-23, 2019.

[2] Assistant National Statistician of the Macroeconomic Accounts Service (MAS), Sectoral Statistics Office of the Philippine Statistics Authority (PSA). The author acknowledges the assistance of Mark C. Pascasio of the Expenditure Accounts Division and Polaris C. Bautista of the Environment and Natural Resources Division, both from the PSA.

This paper aims to present the development of expenditure accounts on disaster risk reduction (DRR), mainly the operational framework and assessment of available data vis-à-vis the required data coming from censuses, surveys and administrative-based information to capture expenditures on disaster by the government, corporations and households. This will also discuss some initial results of the key variables/indicators derived from the Expenditure Accounts on DRR of the Philippines. Coordinative mechanisms to ensure the improvement and institutionalization of DRR statistics by the Philippine Statistics Authority will also be discussed

1. Introduction

The Philippines is vulnerable to different types of disasters like typhoons, earthquakes and floods. The World Risk Index for 2016 reported that the Philippines ranked third among the 171 countries with the highest disaster risk levels (IRDR 2016). The report cites the archipelagos' location in the typhoon and earthquake belts, high degree of exposure and vulnerability and the lack of capacities in coping with disasters. For disaster risk reduction and management which is aimed to protect human lives and infrastructure as well as the conservation of the environment and natural resources, it is crucial that we look into a comprehensive disaster information and improved utilization of these information to help planners and policy makers for more effective disaster management.

Disasters affect all elements of society and they threaten sustainable development in many places around the world. However, disaster statistics is seen as relatively new statistical domain. This cut across several disciplines and needs an integrated framework to coordinate the disaster data into a comprehensive data system. In the 3rd meeting of the Asia-Pacific Expert Group on Disaster-related Statistics in 2015, a pilot test of framework on disaster statistics to provide a summary of core principles of disaster related statistics was agreed among Asia Pacific countries. This supports the mandate of the Expert Group and the requirements for monitoring progress towards the achievement of the Sendai Framework.

The Philippine Statistics Authority (PSA), as the highest policy making body on statistics in the Philippines, has aligned its efforts to advance the pilot testing of the framework for the disaster statistics to cater the needs of various stakeholders as well as the requirements of the Sendai Framework on Disaster Statistics and, in a broader framework, the Sustainable Development Goals (SDG). The PSA has build on existing frameworks of the System of National Accounts (SNA) and the System of Environmental-Economic Accounting (SEEA) on the feasibility of highlighting the Expenditure Accounts on Disaster

Risk Reduction and possibly moving forward, as a satellite accounts to provide linkages in the SNA and the SEEA.

This paper aims to present the initial efforts of the Philippine Statistics Authority (PSA) on developing the disaster related statistics for advancing the expenditure accounts on disaster risk reduction (DRR). This will also discuss some initial results of key indicators which can be integrated in the broader framework of the Philippine Expenditure Accounts on Disaster Risk Reduction.

2. Philippine Efforts in Developing the Disaster Statistics for the Expenditure Accounts for Disaster Risk Reduction

Disaster is defined as a serious disruption of the functioning of a community or a society at any scale due to hazardous events interacting with conditions of exposure, vulnerability and capacity, leading to one or more of the following: human, material, economic and environmental losses and impacts. Disaster risk, on the other hand, is the potential loss of life, injury, or destroyed or damaged assets which could occur to a system, society or a community in a specific period of time, determined probabilistically as a function of hazard, exposure, vulnerability and capacity." (United Nations General Assembly 2015).

In most countries, the Sendai Framework serves as their focal point in developing disaster statistics as initiated by the United Nations International Strategy for Disaster Reduction (UNISDR). The Sendai Framework describes disaster risk reduction (DRR) as a scope of work aimed at preventing new disaster risks as well as reducing existing disaster risks and managing residual risk, all of which, contributes to strengthening resilience.

The United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP) in May 2014 through the ESCAP Resolution E/ESCAP/RES/70/2, saw the need of developing a Disaster Related Statistics Framework (DRSF) in the region including the DRSF guidelines to assist countries how to develop and implement disaster statistics. This initiative did a pilot test of the framework in four (4) countries – Philippines, Bangladesh, Fiji, and Indonesia. The framework identified, among others, the compilation Disaster Risk Reduction Expenditure (DRRE) Accounts including the statistics that supports the DRRE for better analysis of the occurrence of disaster that would result in a more effective disaster reduction risk management. Specifically, this will be useful in analyzing the impact of the economic policies within the concern of the environment and disaster.

The pilot testing of the DRSF framework in the Philippines was spearheaded by the Philippine Statistics Authority. Initial results showed that available data on disaster statistics are not adequate to sufficiently support the DSRF framework. However, the pilot testing was able to define the key issues on addressing most of the data problems and to bring up those

recommendations how these data problems can be addressed. The pilot testing was done at the national level which started with the assessment of data describing the five (5) areas of concerns namely: availability, level of disaggregation, source agency, and the use of disaster-related statistics. The available data source agencies were encouraged to accomplish data assessment based on their own way of reporting, recording, collecting, and generating their data. This activity likewise provided an opportunity to build capacities among agencies providing the data support to meet the international and national statistical demand that are disaster-related and to gauge inconsistencies of reporting between data sources.

The development of DSRF is an attempt to provide guidelines for DRR statistics to help with the consistency in the use of different terminologies, clarifications of the scope of measurement demands and provide different methodological advice to improve the quality of the DRSF.

Table 1 below shows the basic disaster related statistics to include the statistics - before, during and after the disaster. The statistics under "before disaster" include risk assessment focusing on exposure, vulnerability, hazard and coping capacity. The statistics under "during and after disaster" refers to disaster impact on human and dwellings.

Table 2 provides a summary of DRSF Tables which describes the Sendai Goals, the description of categories, total number of tables, the total number of tables with data for years 2013, 2014 and 2015 including the percentage of data compiled vis-à-vis the total tables.

Table 1: Basic Statistics of Disaster-related Statistics-Before, During and After a Disaster

BEFORE			
Risk Assessment			
Exposure	Vulerability	Hazard	Coping Capacity
• Population/density • Land cover/use • GDP per capita •Household income	• Age structure • Poverty • Disabled • Gender • Slums • Median income	• Elevation map • Meteorological data • Data from past hazards • Distribution of soil type • Value for surface roughness • Slope and river flow variables (flood) • Annual precipitaton index (drought) • Annual precipitaion and geological data (landslide)	• Preparedness (household level) • Early warning systems • Investment in DRR

DURING AND AFTER	
Disaster Impact	
Human Impact	Dwellings
• Death/Missing • Injured/Ill • Evacuated/Relocated • People affected by dwellings damaged/ destroyed • Loss of jobs • People received aid • Otherwise affected	• Buildings and structures • Machinery • Valuables • Natural Resources • Critical goods and services • Crtitical infrastructure • Cultural heritage • Economic impacts

Data Sources
Population Census/Household Surveys
Household Surveys
Geophysical/ Meteorological organisations
National accounts
Disaster Management Agency Assessment
Ministry of Environment/ Mapping Agency

DRR Activity		
Transfers from Central to local government	Evaluations Relocations People receiving aid	Reconstruction and recovery
International Transfers (ODA) into country		
International Transfers (ODA) outflows		
Total expenditure by DRRCA category		
Total Expenditure on DRRCA by government		
Total Expenditure on DRRCA by non-government		
Beneficiaries on Total Expenditure		
Beneficiaries of Total Transfers Received		

Table 2: Summary of DRSF Tables

Sendai Goals Categories/Description		Total number of tables	No. of tables with data			Percent of data compiled by category per year
			2013	2014	2015	
A	Summary tables of events by hazards types	3	1	1	1	33
B	Selected background statistics by hazard types and administrative units and river sub-catchments	4	2	2	2	50
C	Summary tables of affected population	11	5	5	5	45
D	Summary tables of direct material impacts	3	2	2	2	67
E	Summary tables of direct material impacts in monetary terms	3	0	0	0	0
F	Summary tables of direct cultural impacts	3	0	0	0	0
G	Summary tables of direct environmental impacts	3	0	0	0	0
Transfers	Disaster risk reduction satellite accounting	2	0	0	0	0
TOTAL		**32**	**10**	**10**	**10**	

The Core Set of statistics for the DRSF constructed upon a clearly identified policy demand and existing mandate for the government organization for monitoring. As indicated in the Sendai Framework there are seven (7) targets and by monitoring it, would require good quality statistics on disaster occurrences and direct impacts (See Annex I).

3. The Proposed Expenditure Accounts for Disaster Risk Reduction

In 2016, the Disaster Related Statistics Framework was conducted and pilot-tested in the Philippines, along with other 3 countries - Bangladesh, Fiji, and Indonesia. The summary tables likewise highlighted on the Disaster Risk Reduction Expenditure and Transfers (DRRE). The aim of the DRRE tables is to start developing a special account or a satellite account of the national accounts that are linked with the System of National Accounts (SNA) and the System of Environment-Economic Accounts.

The DRR activities was developed to align with the concepts, standards and formats of the SNA since the information considered in the DRRE are obtained and aligned with the broader framework for the economy. Data and data sources are exactly the data sources used for the SNA. Another task for the DRRE is to be able to separately identify those activity with a primary disaster risk reduction purpose.

Data and Data Sources

In the Philippines, the first attempt to look at the available data in developing the Expenditure Accounts on Disaster Risk Reduction has build on the existing data from the national accounts particularly those data for the compilation of the Public Administration and Defense under the Production Account and the Government Final Expenditure Accounts of the Expenditure Accounts. These data are mainly obtained from the Annual Financial Report (AFR) of the Commission on Audit (COA), the National Expenditure Program (NEP) of the Department of Budget and Management (DBM), Consolidated Report on Official Development Assistance (ODA) Programs and Projects and the ODA Portfolio of the National Economic and Development Authority (NEDA). The Office of Civil Defense under the National Disaster Risk Reduction Management Council provides data on the number of hazardous and disaster events as well as the number of people evacuated/displaced and affected; number of damaged and destroyed dwellings due to disaster events. In addition, the following data sources are generating information on:

- Department of Health (DOH) – provides data on number of deaths and number of injured due to disaster; damaged and destroyed critical infrastructures e.g. health facilities; medical cost of people injured or ill due to disaster events;

- Department of Public Works and Highways- Bureau of Maintenance (DPWH-BOM) – reports data on damaged infrastructures like roads, bridges, etc due to disaster events;
- Philippine Institute of Volcanology and Seismology (PHIVOCS) – reports on earthquake occurrences and volcanic activities;
- National Water Regulatory Board, Department of Environment and Natural Resources (DENR-NWRB) – provides data on water permits granted by type of use e.g. municipal, industrial, irrigation, power, fisheries, livestock, recreation, etc;
- Biodiversity Management Bureau (DENR-BMB) – provide reports on protected areas, classification of caves, wetlands including community based resources management, biodiversity financing;
- Land Management Bureau (DENR-LMB) - generates data on land through surveys to include data on land disposition, land uses, patrimonial properties, forest lands, shore lands, etc;
- National Mapping and Resource Information Authority (DENR-NAMRIA) – generates maps e.g. hazard maps, land cover statistics, etc.

Compilation Process (for government expenditures on disaster risk reduction)

1. Consolidate the total government expenditure on disaster risk reduction activities by type of program and by type of project that are sourced from the National Expenditure (NEP) of the Department of Budget and Management for years prior to the disbursement of the actual expenditures and the actual expenditures provided from the Annual Financial Report of the Commission on Audit (COA); all DRR activities are identified by agency and by year.
2. Examine the available budget from the National Disaster Risk Reduction Management Fund (NDRRMF) and exclude them from the total budget by different source agencies to make consistent and to avoid overlaps in reporting of the budget.
3. Define from the reported expenditures of the NDRRMF all the DRR activities and categorize them into four (4) thematic areas namely: (a) disaster prevention and mitigation; (b) disaster preparedness; (c) disaster response; and (d) disaster rehabilitation and recovery.

   The total government expenditures for DRR activities is a sum of the : (a) expenditures reported from the National Expenditure Program and/or the Commission on Audit; (b) expenditures reported by the National Disaster Risk Reduction Management Fund; and (c) expenditures reported by the Overseas Development Assistance Fund (ODA).
4. Compute for the share of government expenditures on disaster risk reduction to the total reported expenditures of the government revenue

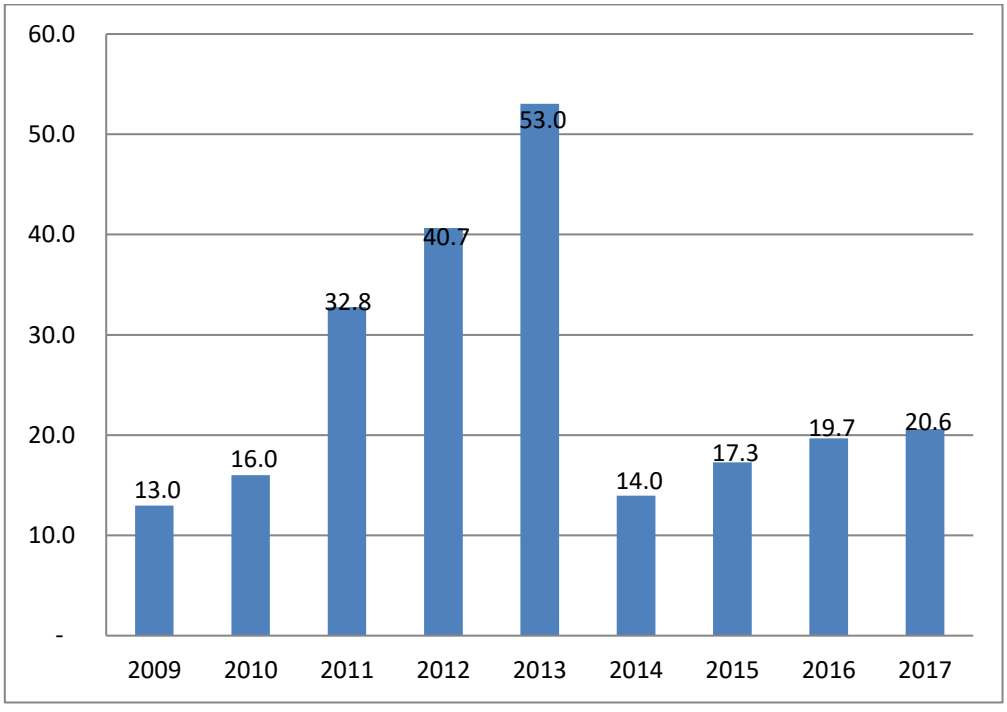
of the Commission on Audit and to the total government final consumption expenditures of the national accounts.

There are eleven (11) government agencies included in the compilation of government expenditures on DRR for years 2009 to 2017 namely: (1) Department of Education (DepEd); (2) State Universities and Colleges (SUCs); (3) Department of Agriculture; (4) Department of Health; (5) Department of Interior and Local Government (DILG); (6) Department of Justice (DOJ); (7) Department of National Defense (DND); (8) Department of Public Works and Highways (DPWH); (9) Department of Science and Technology (DOST); (10) Department of Social Welfare and Development (DSWD); and (11) Department of Transportation (DOTr).

4. Some Preliminary Results : Expenditures on Disaster Risk Reduction

1. The Highest Government Expenditures from 2009 to 2017 was Recorded in 2013 at 53.0 Billion Pesos!

This was mainly due to the country's most hazardous event of Yolanda disaster in November 2013. After 2013, this was followed by disaster expenditures in 2012 and 2011 at 40.7 million and 32.8 million, respectively.



2. In 2016, Total Actual Government Spending for Disaster Risk Reduction was Highest at 82.3 Percent to Finance Disaster Rehabilitation and Recovery!

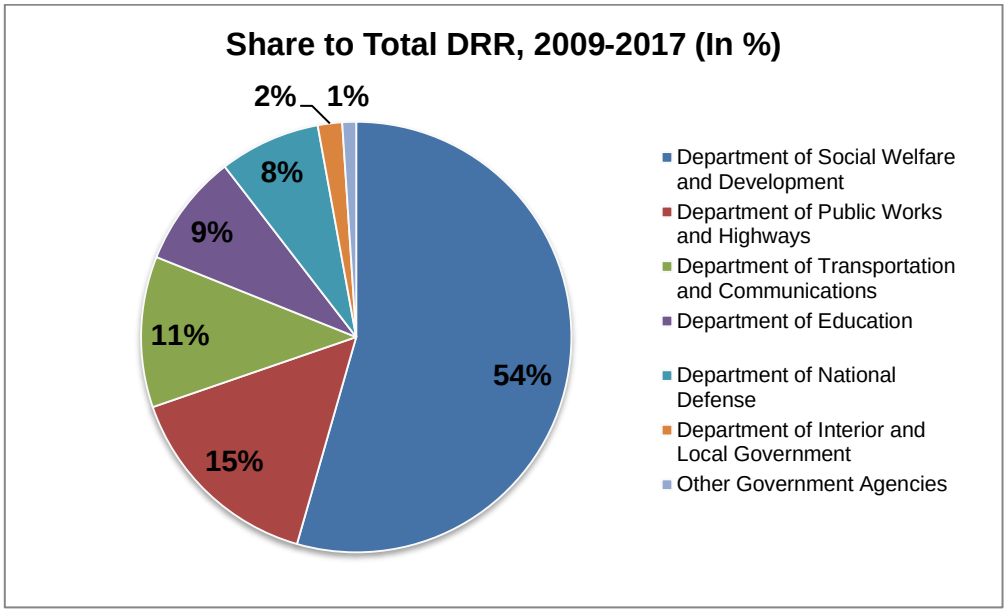
This represent 16.2 billion pesos out of the total government expenditures of 19.7 billion pesos.

Disaster Thematic Area	2016 Expenditures (in 000 pesos)	Share to Total Expenditure (in Percent)
Disaster Prevention and Mitigation	1,852,442	9.41
Disaster Preparedness	692,689	3.52
Disaster Response	932,485	4.74
Disaster Rehabilitation and Recovery	16,197,736	82.32
Total DRR Expenditure	19,675,352	100.00

3. Government Spending in Disaster Risk Reduction was Recorded by the Agencies on Engaged in Social Welfare, Public Works and Highways as well as Transportation!

In 2016, most of the expenditure went to the Department of Social Welfare and Development (DSWD) with Php 10.0 billion followed by Department of Public Works and Highways and Department of Transportation with Php 3.0 billion and Php 2.3 billion, respectively. On the average, 54 percent of the total expenditure on DRR activities went to DSWD from 2009-2017.

Agency	2009	2010	2011	2012	2013	2014	2015	2016	2017
Department of Education	1,498,286	1,376,011	1,465,635	1,617,469	1,869,339	1,297,440	1,974,948	2,096,077	2,642,999
State Universities and Colleges	9,771	11,012	12,272	13,511	15,906	16,431	18,820	20,077	22,211
Department of Agriculture	117,891	121,732	94,470	127,960	159,279	139,073	152,684	176,973	191,305
Department of Health	5,307	6,231	7,985	9,444	10,612	11,157	14,032	17,114	20,981
Department of Interior & Local Government	451,166	366,183	270,661	292,511	298,315	308,695	311,151	373,176	553,301
Department of Justice	8,898	7,977	9,559	9,393	14,639	11,978	12,331	12,296	13,451
Department of National Defense	1,341,888	1,376,704	1,463,560	1,725,559	1,744,419	1,479,959	1,504,184	1,659,387	1,987,087
Department of Public Works and Highways	2,871,646	3,117,055	2,634,897	4,296,968	3,740,000	2,624,198	3,000,435	3,020,448	3,872,096
Department of Science and Technology	21,920	21,202	29,460	30,065	34,593	36,288	40,216	39,755	41,851
Department of Social Welfare and Development	5,393,268	7,539,454	23,731,501	30,245,676	42,472,966	5,343,191	7,662,109	9,991,172	8,459,045
Department of Transportation and Communications	1,251,226	2,063,083	3,058,935	2,297,422	2,667,365	2,695,793	2,585,685	2,268,875	2,807,786
Total	**12,971,266**	**16,006,643**	**32,778,934**	**40,665,980**	**53,027,434**	**13,964,203**	**17,276,596**	**19,675,352**	**20,612,113**



4. In 2016, the Country Received 18.8 Billion Pesos from the Official Development Assistance (ODA) on Disaster Risk Reduction, Almost Close to the Total Government Spending at 18.8 Billion Pesos!

DRR Expenditure with ODA

Expenditure by Agency	Amount (Thousand Php)
Total DRR Expenditure	19,675,352
Official Development Assistance**	18,786,560
Total DRR Expenditure with ODA	38,461,912

** still for validation

5. Conclusions and Recommendations

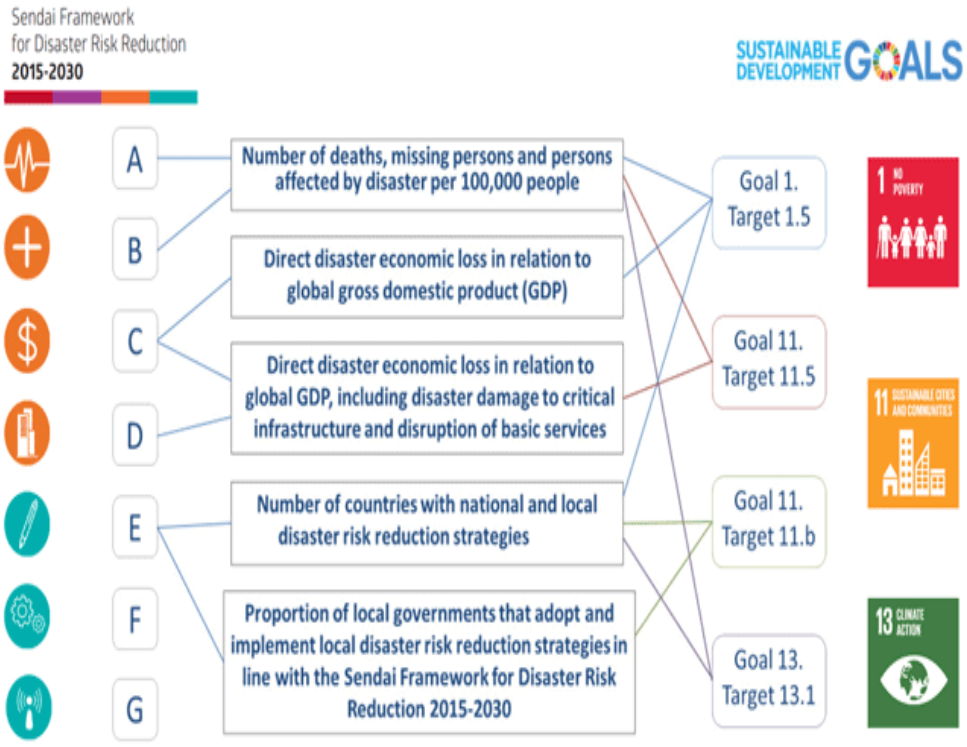
The Disaster Related Statistical Framework (DRSF) and the Disaster Risk Reduction Expenditure (DRRE) Accounts are valuable documents that will help in the production of disaggregated data related to disasters and strengthen evidence-based policy-making at all levels for disaster risk reduction and climate change adaptation. The following are recommendations towards improving the DRSF and DRRE Accounts:

a. Establish a classifications of disaster related activities whose primary purpose are for disaster risk reduction including the classification of products used for DRR activities;
b. Continue to support countries on capacity building particularly among data producers, users and compilers of disaster related statistics and DRRE accounts how to strenghthen disaster related statistics;
c. Advocacy on the uses and application of DRR statistics and DRRE accounts towards more policy recommendations for more effective DRR management at all levels.

References

1. Philippine Statistics Authority. Philippine Statistical Development Program, 2017-2023.
2. Philippine Statistics Authority. Unpublished Technical Report; DRR Expenditure Accounts, 2018
3. UNESCAP 70th Session; (13 June 2014); Resolution 70/2 2014; Disaster related statistics for Asia and the Pacific (E/ESCAP/Res 70/2).
4. UNISDR. Progress and Challenges in Disaster Risk Reduction. Summary & Main Findings. 2014
5. UNGA (2015) UN General Assembly; 74th Session (1 December 2016); Report of the open-ended intergovernmental expert working group on indicators and terminology relating to DRR
6. United Nations University (2016); World Risk Report 2016; worldriskreport.org; Bundis Entwishlung Lift Berlin.
7. UN System of National Accounts, 2008.

Annex I. Sustainable Development Goals (SDGs) and the Sendai Framework



Source: UNISDR

Annex III. The SDG targets aligned with Sendai Framework Global Target

Table 1. Detailed Description of Indicators Source: UNISDR SDG indicators		Sendai Framework indicators
Goal 1. End poverty in all its forms everywhere		
1.5.1	Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population	A1 and B1
1.5.2	Direct economic loss attributed to disasters in relation to global gross domestic product (GDP)	C1
1.5.3	Number of countries that adopt and implement national disaster risk reduction strategies in line with the Sendai Framework for Disaster Risk Reduction 2015-2030	E1
1.5.4	Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies	E2
Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable		
11.5.1	Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population	A1 and B1
11.5.2	Direct economic loss in relation to global GDP, damage to critical infrastructure and number of disruptions to basic services, attributed to disasters	C1, D1, D5
11.b.1	Number of countries that adopt and implement national disaster risk reduction strategies in line with the Sendai Framework for Disaster Risk Reduction 2015-2030	E1
11.b.2	Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies	E2
Goal 13. Take urgent action to combat climate change and its impacts		
13.1.1	Number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population	A1 and B1
13.1.2	Number of countries that adopt and implement national disaster risk reduction strategies in line with the Sendai Framework for Disaster Risk Reduction 2015-2030	E1
13.1.3	Proportion of local governments that adopt and implement local disaster risk reduction strategies in line with national disaster risk reduction strategies	E2

Source: UNISDR

The new Portuguese Central Credit Register: A powerful tool for a Central Bank

Luís Teles Dias*, António Jorge Silva+
Banco de Portugal, Lisboa, Portugal

Abstract

The use of integrated micro-databases has been the cornerstone of the *Banco de Portugal's* long-term strategy for data management, allowing for the fulfilment of its statistical requirements and also contributing with key information to several business areas within the central bank's remit. The Portuguese Central Credit Register (CCR), developed and managed by *Banco de Portugal,* constitutes an important case-in-point of this approach. In fact, CCR data have proved their relevance for a variety of purposes, from, *inter alia,* the compilation of very comprehensive and detailed statistics on credit, to the promotion of a better understanding of the risks underlying banks' balance sheets. With the goal of streamlining in a single reporting framework and in a single database the credit and credit risk data available to *Banco de Portugal,* and increasing significantly the detail of information reported, a new loan-by-loan CCR has recently been developed. The new CCR went considerable beyond the previous CCR, covering more than 200 attributes, way above the 29 offered by the former. This new and enhanced CCR allows to significantly improve the depth and completeness of the credit and credit risk data available for several policy and analytical purposes, namely, banking supervision, financial stability, monetary policy and economic research. An interesting case-study is the use of data from the new CCR to assess the commercial banks compliance with a macroprudential measure that has been issued by *Banco de Portugal* in 2018. This macroprudential measure introduced limits in new credit granted to consumers from 1 July 2018 onwards in terms of loan-to-value ratio (LTV), debt-service-to-income ratio (DSTI) and loan maturity at origination. The aim is to prevent the accumulation of excessive risk in the banks' balance sheets and ensure that households obtain sustainable financing.

Keywords

Credit register data; AnaCredit Regulation; Macroprudential analysis; Microdata JEL classification: C80; E42; E58; G21

1. Introduction

The Great Financial Crisis of 2007-08 (GFC) drew attention to the existing gaps in the availability of detailed information needed for timely risk assessment. In fact, although very relevant for central banks, traditional aggregate statistics have proved insufficient to monitor and analyse the many aspects of the monetary transmission mechanism and the evolution of credit to companies and households.

Early on, the *Banco de Portugal* (hereinafter referred to as "the Bank") understood that complementing aggregate statistics with more granular data is not only an answer to the need for more flexible and detailed information, but also a movement towards more efficient and reliable reporting systems[1]. This is particularly true in situations where a single report of granular data has the potential to substitute several reports of aggregate information, representing different perspectives on a given reality (e.g., credit). In such cases – of which the Portuguese Central Credit Register (CCR) is a good example –, central banks have the possibility to build a consistent and coherent multipurpose granular database, which could then be used by the various functions of those institutions, each of them analysing the granular dataset from its own perspective. This choice would increase not only the flexibility of the data, but also the efficiency of the reporting systems themselves.

The Portuguese CCR was created in the late 1970s with the objective of providing the participating institutions with relevant information to assist them in the assessment of risks when granting credit. Despite the significant changes that the CCR has undergone over the ensuing four decades, its original objective still remains as its main drive. Nevertheless, the CCR is nowadays used for a very diverse set of purposes related to specific functions of the Bank – e.g., statistics, prudential and conduct supervision, economic research, monetary policy, financial stability analysis and risk management.

In this paper, we present the main recent changes in the Bank's CCR and explain how the new, redesigned, loan-by-loan database will help the Bank to fulfil its mission, highlighting the importance of the availability of such granular information to assess the compliance of financial institutions with the recently enacted macroprudential measures. In section 2, we provide a brief overview of the key features of the new Portuguese CCR; section 3 focuses on the role of the CCR in supporting the Bank's macroprudential analysis and policymaking function; section 4 presents some final remarks.

[1] For a more detailed analysis on how the *Banco de Portugal* has been complementing traditional aggregate statistics with granular data, particularly in the field of monetary and financial statistics, please see Dias & Silva (2017).

2. The New Portuguese Central Credit Register

The Portuguese CCR is an information system managed by the Statistics Department of the Bank, which contains monthly granular information on credit granted[2] by the institutions participating in the system – all resident credit-granting institutions[3].

Over time, the CCR has shown a significant and increasing potential and usefulness to support several central bank's functions. Currently, given its granularity and virtually complete coverage, CCR data are used by the Bank, *inter alia*, for the following tasks[4]:

- Compiling comprehensive statistics on credit, with breakdowns by, *inter alia*, institutional sector of the borrower, sector of activity, type of instrument, purpose, size of firms, location/region, original and residual maturity, type of guarantees, and amount of credit exposure;
- Assessing credit concentration and distribution;
- Watching carefully the evolution of overdue loans and overdue loans' ratios;
- Understanding the risks underlying banks' balance sheets;
- Contributing to the Bank's own in-house credit risk assessment system;
- Monitoring the use of credit claims as collateral for Eurosystem[5] credit operations.

The Impact of AnaCredit in the Portuguese CCR

Since 2007, the European System of Central Banks has been exploring the potential of CCRs for statistical purposes, in order to gain a better overview of credit developments in the Member-States of the European Union. In particular, it was sought to understand how the scope and content of national CCRs could be improved and adapted to meet European statistical needs, while fostering a reduction of the reporting burden of the participants and an increase in transparency.

In 2011, the European Central Bank (ECB) together with all euro area and some non-euro area national central banks, launched the AnaCredit project[6].

---

[2] There is a virtually complete coverage – all loans with an initial amount of EUR 50 or more must be reported

[3] The participating institutions are the following: Monetary Financial Institutions (MFIs) – that is, banks, savings banks and mutual agricultural credit banks; non-monetary financial institutions that grant credit; public agencies that grant credit, and non-financial corporations acquiring loans from the resident financial sector. By law, the participation of these types of institutions in the CCR is mandatory.

[4] For a more detailed discussion on the Portuguese CCR and its several uses please see Matos (2015) and Matos & Dias (2017).

[5] The Eurosystem comprises the European Central Bank and the national central banks of the European Union Member-States that have adopted the euro.

[6] The name stands for "analytical credit datasets".

AnaCredit is a dataset containing detailed and monthly updated information on individual bank loans in the euro area, harmonised across all Member-States. It uses new data and existing national credit registers to achieve a harmonised database that supports several central banking functions, such as decision-making in monetary policy and macroprudential supervision.

Two main reasons help to explain the importance of the creation of AnaCredit at the European level: (i) the GFC showed that economic sectors in different countries do not exhibit a homogeneous response to economics shocks and, therefore, the availability of granular data could play an essential role in monitoring such responses; and, (ii) in the wake of the GFC, the ECB and some national central banks in Europe have taken on new macroprudential tasks that require, namely, comparable good quality granular information on credit.

With a view to fulfilling the AnaCredit's requirements, the Portuguese CCR has been completely redesigned in 2018 and adopted a new data model: a loan-by-loan basis instead of the borrower-by-borrower approach that had been in place since its inception. Although the first stage of AnaCredit comprises only loans granted by banks to legal entities (thus excluding, for the moment, households) with an exposure above EUR 25,000, the Portuguese CCR has kept its extensive coverage, both in terms of its participating institutions (all resident credit-granting institutions), borrowers (legal and natural persons), and threshold (EUR 50), in an attempt to cover all the attributes for most of this universe.

In reality, the redesign of the Portuguese CCR was not just a move to meet the AnaCredit's requirements – rather, there was a paradigm shift in data management at the Bank, according to which the CCR now operates as the single entry point for all credit and credit risk data, thus creating a multipurpose hub of credit information that can be used by various business areas of the Bank.

The New CCR

The implementation of the new CCR information system took into careful account other data needs (not related with AnaCredit) and specific functionalities identified as relevant by the main stakeholders. The resulting new data model includes not only the 94 attributes requested by AnaCredit but also other credit data attributes needed by the Bank's internal users, allowing for the rationalization of data submissions by financial intermediaries, through the use of the single entry point approach (as shown in Figure 1 below), and permitting to achieve a high standard of data integration.

Figure 1 – **Rationalization of credit data reports to the *Banco de Portugal***

In order to improve the performance of the Bank's tasks related to monetary policy-making, risk management, statistical compilation, supervision and financial stability, the new CCR covers more than 200 attributes. This means that, when a loan is eligible to be reported to the CCR, the participant institutions have to submit information on the instrument, the debtor(s), the protection/guarantees and the accounting and risk information. Moreover, to meet a need of the financial intermediaries, the CCR will also deal with daily data[7] on relevant credit events, thus allowing for a better evaluation of the credit risk of the borrowers and enabling the Bank to monitor the credit evolution within the financial system with a much smaller time lag.

Institutions do not need to report any reference data on resident legal and natural persons. For the identification of resident borrowers (and guarantors) it is sufficient that its taxpayer number is reported to the Bank. The enrichment of the database with reference data on these entities is done by the Bank through its business register. This procedure ensures that different participant institutions do not report to the Bank different classifications (e.g., by sector or by size) for the same borrower.

[7] The system is fully prepared to deal with daily data but this module will only go live after an amendment in the legal act regulating the CCR that is under approval process by the Portuguese government.

Figure 2 – The new CCR data attributes

3. Financial Stability and Macroprudential Policy — What's in it for the New CCR?

As discussed in the previous section, the Bank has been leveraging the Portuguese CCR to meet its tasks in several domains. One of the most relevant task entrusted to the Bank in its Organic Law is ensuring "*the stability of the national financial system, performing for this purpose, in particular, the functions of lender of last resort and national macro-prudential authority*" and participating "*in the European system for the prevention and mitigation of risks to financial stability and in other bodies pursuing the same goal*". To meet this challenge, the Bank resorts to a number of different inputs and techniques that allow for a systemic view of the financial system and of the build-up of systemic risks.

In this context, data from the CCR are an instrumental and extensively used input, analysing the various dimensions and characteristics attached to the loans, debtors and/or creditors. Indeed, in light of its intrinsic homogeneity and of the possibility to compare its data with other databases, the CCR data allows for a complementary analysis to the "traditional" aggregate data by providing the underlying distribution measures and by enabling the enhancement of the testing and monitoring (e.g., stress testing) of the banks' results in ever-changing and increasingly complex scenarios.

Indeed, Lima & Drumond (2015) discussed the insufficiencies attached to aggregate data when assessing financial stability and showed how microdata databases, such as the CCR, enable an evaluation of the causes of the movements behind the aggregates and thus uncover the potential build¬up of imbalances. Moreover, they also recognize that some macroprudential tools require specifically the use of characteristics that are only available in granular datasets – such as the collateral amount of real estate and debt instalments.

Macroprudential Measure on New Credit Agreements for Consumers

With the aim of promoting financial stability, the Bank announced in February 2018 a macroprudential measure (hereinafter referred to as "the Recommendation") addressed to credit institutions and financial companies.[8] More specifically, it fosters the adoption of prudent credit standards on loans granted by the Portuguese financial system to consumers, with the aim of enhancing the resilience of the financial sector and the sustainability of households' financing, thereby minimizing the risk of defaults. Furthermore, the Bank seeks to prevent excessive risk taking by the financial sector, in a context where less restrictive credit standards have been observed and are expected to become even looser.[9]

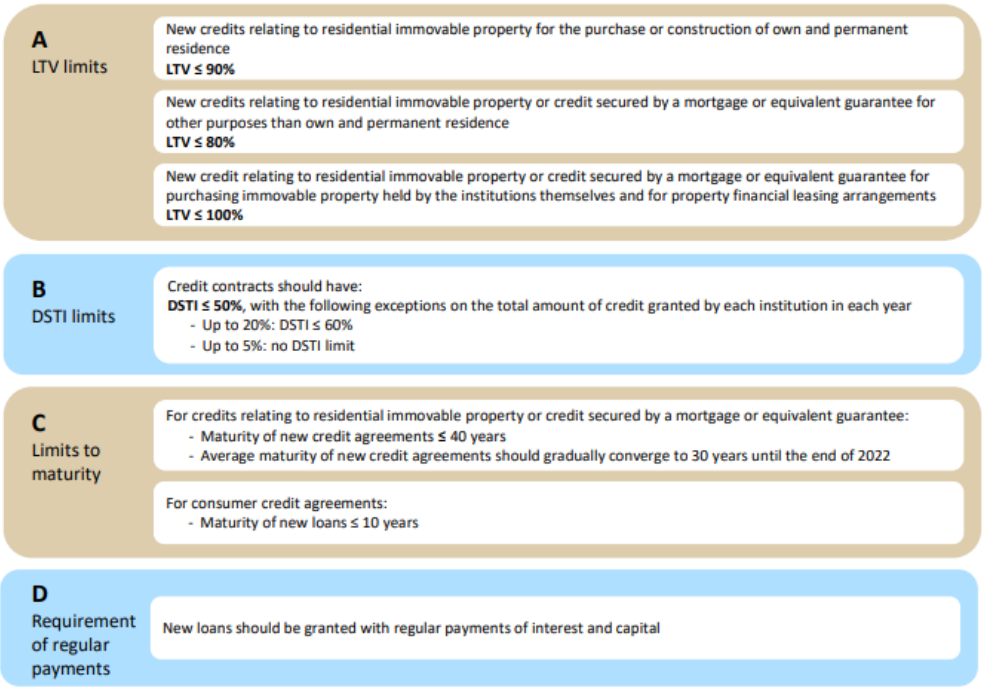
This Recommendation, which is based on the "comply-or-explain" principle, introduced the following terms for new credit granted as from 1 July 2018:

- ↑ Limits to the loan-to-value ratio (LTV) – defined as the ratio between housing loan(s) and the minimum between the purchase price and the appraisal value of the house granted as collateral;
- ↑ Limits to the debt service-to-income ratio (DSTI) – defined as the ratio between monthly instalments of total credit agreements and the borrower's income, net of taxes and contributions to social security;
- ↑ Limits to the maturity of loans;
- ↑ Requirement of regular payments of interest and capital.

[8] The set of recommendations can be found in *Banco de Portugal* (2018a).

[9] For a detailed analysis of the Recommendation and its justification please see *Banco de Portugal* (2018b).

Figure 3 – Summary of the Recommendation



According to the Recommendation, "*Banco de Portugal shall monitor the implementation of this Recommendation at least once a year as well as the evolution of credit agreements for consumers excluded from the scope of this Recommendation*." In previous years, the compliance with a Recommendation such as this one would have to be done via an annual *ad hoc* request addressed to all credit institutions, or, alternatively, directed to only a sample of institutions with the inherent negative impact on the preservation of a level playing field amongst credit institutions in Portugal. Presently, the universal coverage of the CCR, together with the level of detailed information (given the number of attributes) available on a loan-by-loan basis, allows the Bank to assess the compliance with the Recommendation for all credit institutions without needing to set up a dedicated report to obtain additional information from them.

Furthermore, CCR data allow a more in-depth analysis that makes it possible for the Bank to calibrate the Recommendation almost in real time, should the need arises, rather than having to wait several months to receive *ad hoc* information from the credit institutions to conduct such analyses.

4. Final Remarks

The development of the new Portuguese CCR was a major challenge, given the ambitious objectives set for this project – *inter alia*, to act as the single entry point of all credit and credit risk data with a very high level of granularity.

This allowed for the enhancement of the service provided to the participant institutions (and to the public at large, who now have an easier access to their own credit reports) and facilitated the reporting procedures of the institutions – not only because much less transformation rules have to be applied, but also due to the integration in a single reporting system of a number of autonomous reports on credit data (with different concepts, nomenclatures, formats, frequencies, timeliness, etc.) that they had to comply with.

That said, the main challenges that the new CCR had to face were of internal nature, as the project induced important organizational changes in data management: (i) while the Statistics Department of the Bank is the system owner of the new CCR, the ownership of the data is collective (it involves 5 departments of the Bank); (ii) data management, including data quality assurance, is a shared responsibility implying individual (business areas) commitment in contributing to the overall quality of the data; (iii) a governance/relationship model had to be established between the different data owners; (iv) the new CCR champions the paradigm of data-sharing and helps eradicating the traditional data silos landscape; and, (v) the high volumes and granularity of the data call for new skills among the staff.

Despite the daunting challenges faced at the outset, the benefits gathered at the end are inescapable. The policy measure described above – i.e., the macroprudential recommendation related to new loans to consumers – is just one fine example of the usefulness of having a credit register offering the features that the new Portuguese CCR provides. The designing of the said measure was highly facilitated by the extensive availability of underlying data via the new CRC; the decision-makers only needed to focus on the definition of the rules that should be met by the financial institutions when granting credit and definitely not on how the data needed to monitor the recommendation would have to be obtained. This was assured by the new CCR.

References

1. Banco de Portugal (2018a), Recommendation of Banco de Portugal within the legal framework of new credit agreements for consumers.
2. Banco de Portugal (2018b), Macroprudential measure within the legal framework of credit for consumers.
3. Dias, Luís Teles & Silva, António Jorge (2017), Upgrading monetary and financial statistics in the wake of the financial crisis– There's life beyond aggregate data, ISI Regional Statistics Conference 2017, Bali, Indonesia.
4. Lima, Filipa & Drumond, Inês (2015), How to keep statistics' customers happy? Use micro-databases!, IFC Workshop on Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges, Warsaw, Poland.

5. Matos, João Cadete (2015), The Portuguese Central Credit Register: a powerful multi-purpose tool, relevant for many central bank's functions, IFC Workshop on combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges, Warsaw, Poland.
6. Matos, João Cadete & Dias, André Cardoso (2017), The Portuguese Central Credit Register as a key input to the analysis of financial stability... and beyond!, IFC – National Bank of Belgium Workshop on "Data needs and Statistics compilation for macroprudential analysis", Brussels, Belgium.

Have domestic prudential policies been effective: Insights from bank-level property loan data

Veronica B. Bayangos[1], Jeremy De Jesus[2]
[1]Supervisory Policy and Research Department, Bangko Sentral ng Pilipinas
[2]Department of Economic Statistics, Bangko Sentral ng Pilipinas

Abstract
The study examines the effectiveness of domestic prudential policies in restraining growth of real bank loan commitments and in preserving the quality of bank loans in the Philippines using panel bank data regression from the first quarter of 2014 to the fourth quarter of 2017. The study reveals important findings for the BSP. First, tightening of domestic prudential policies, particularly those tightening measures meant to preserve resilience of the banking system are effective in curbing growth of real bank loan commitments to borrowers for acquiring new residential properties. Second, this study highlights the bigger negative impact of tightening prudential measures on real bank loan commitments by universal and commercial banks compared to thrift banks. Third, the share of bank deposits to total liabilities, liquidity position and capital adequacy gap are important drivers of growth in real bank loan commitments to borrowers. Fourth, restricting both instruments meant to promote resilience of banking system and to address cyclical movements limits weakening of bank loan quality, with the latter type of instruments having bigger negative impact. Fifth, tightening of domestic prudential policies varies with monetary policy conditions and over the business and financial cycles in the Philippines. JEL classification: E52, E58, G18, G28.

Keywords
Macroprudential policies, microprudential policy, financial stability, real estate loans

1. Introduction

The study examines the effectiveness of changes in comprehensive domestic macroprudential policies in restraining growth of real loan commitments by universal, commercial and thrift banks to non-financial sector in the Philippines. In recent findings, the use of domestic prudential policies to promote financial stability and prevent the occurrence of financial crisis, which, in turn prevent output losses associated with macroeconomic and financial volatility and financial crises has been highlighted. The use of macroprudential tools to promote financial stability has likewise allowed many central banks to keep monetary policy focused on its primary objective of

maintaining price stability. This has helped enhance monetary policy's credibility in maintaining price stability. In turn, central banks recognize that financial stability policy interacts and influence banking regulations as well as monetary policy actions, implying that central banks need to consider the extent of policy interactions.

Many studies have defined macroprudential policy as a set of measures that prevent or mitigate systemic risk, either over time or across institutions and markets. There are variations on the national/institutional definitions of what constitutes macroprudential policy, but these often hover around these themes - the use of instruments or tools that either increase the resilience of the financial system or constrain systemic risks that are often associated with financial booms. This study covers a more comprehensive set of domestic macroprudential policies classified by instrument, such as those related to credit (or asset-side instruments), liquidity which address the build-up of liquidity and foreign-exchange risks associated with lending booms, capital, banks' reserve requirements on domestic deposits and deposit substitutes, structural or interconnectedness (Orsmond and Price 2016), and currency (Bruno et al. 2015). This study then estimates the effectiveness of these policies in curbing growth of real bank loan commitments to non-financial borrowers who are acquiring new residential properties using an unbalanced panel data regression from 2014 to 2017.

In the Philippines, a detailed study on the effectiveness of the use of prudential policies on the growth of bank credit is yet to be completed. Most of the studies are part of a bigger study or across jurisdictions. In particular, the latest study by Bayangos (2017) found that after controlling for episodes of sterilization of capital inflows across nine Asian emerging market economies for the period 2004-2015, capital inflow restrictions and domestic macroprudential policy are effective in curbing overall real bank and real housing credit and real house prices. Moreover, monetary policy tightening complements tight domestic macroprudential policy in restraining movements in real bank credit and real house prices.

This study is broadly related to a growing area of empirical research on financial stability. The empirical literature on the effectiveness of domestic macroprudential policies in dampening credit cycles across economies remains relevant since the Global Financial Crisis (GFC). In recent past years, empirical evidence of the efficiency of macroprudential policies in restraining excessive credit growth has expanded to include bank-level data and credit registry data. However, credit registry data in many countries, including the Philippines, are limited and confidential. This study uses bank-level data from residential property loan reports involving 101 universal/commercial banks (U/KBs) and thrift banks (TBs).

This study raises five main questions: First, are domestic prudential policies effective in restraining growth of bank loan commitments using bank-level residential property loan data in the Philippines? Second, do responses to a domestic prudential shock differ by type of bank? Third, do responses to domestic prudential policies vary over monetary policy conditions? Fourth, do responses to domestic prudential policies vary over financial cycles of the Philippines? and fifth, do responses to domestic prudential policies restrict bank riskiness?.

This study has three possible contributions to empirical studies. First, this study updates Bayangos (2017) study that documents a database of domestic prudential measures and changes in monetary policy stance for the Philippines to include changes in prudential limits from the first quarter of 2014 to the fourth quarter of 2017. Second, it develops a new database using data from the quarterly bank reports on the Residential Real Estate Price Index (RREPI) from the first quarter of 2014 to the fourth quarter of 2017. Third, the study uses these databases to examine the effectiveness of both tightening and easing of domestic prudential policies to growth of real bank loan commitments and the overall quality of bank loan portfolio. The study will then examine the importance of monetary policy reaction to address changes in real bank loan commitments and in changes in the quality of loan portfolio and the interaction among different instruments of domestic prudential policies. The rest of this study is organized as follows. Section II discusses baseline database and empirical methodology, while Section III highlights the main findings of the paper. Section IV concludes.

2. Baseline Database and Empirical Methodology

Measure of bank loan commitment. This database compiles the volume or number of loans granted for purchases of new residential properties, the average acquisition cost of the property, the appraised value of the residential unit, the appraised value of the lot, the location of these properties, the type of residential property classified into single- detached, duplex, apartments and condominiums from 101 banks. The focus of this database is the compilation of average acquisition cost of residential property as indicator of the commitment of banks to grant loans based on the acquisition cost of the property. The data are generated from the quarterly report submitted by U/KBs and TBs on all Residential Real Estate Loans (RRELs) granted for the generation of RREPI for the Philippines. This study used the appraised value of the residential unit and the appraised value of lot in real terms.

Database for domestic prudential policies. This database includes all the domestic prudential measures adopted by the BSP, classified by instrument, such as, *capital-related* measures that aim to strengthen banks' ability to absorb risks, adjusts banks' capital requirements. These measures include

Basel III capital requirements and adjustments in risk-weights as well as provision requirement; *liquidity-related instruments* which address the build-up of liquidity and foreign-exchange risks associated with lending booms. These instruments include the liquidity coverage ratio and intraday liquidity requirements; *structural or interconnectedness instruments* that aim to address vulnerabilities from interconnectedness and limit contagion. These include interbank exposure limits and additional loss-absorbing capacity for systemically important banks; *asset-related measures (or credit-related instruments)* that place restrictions or caps on amount that can be lent by banks such as loan-to-value (LTV) ratio as well as administrative measures in relation to credit or credit growth; *reserve requirements* imposed against bank deposits and deposit substitutes; and *currency-related instruments* that place limits on net open currency positions and foreign currency lending of banks. The first category captures the measures that are intended to preserve resilience of the banking system. These include capital and liquidity-based measures as well as structural or interconnectedness measures. The second category includes those measures that are expected to address excessive cyclical swings. These include asset-side instruments, banks' reserve requirements and currency-related instruments. These two categories are then aggregated to capture both the measures that are meant to promote resilience of the banking system and to contain excessive cyclical movements.

Moreover, the dataset is classified into tightening and loosening measures. Such a classification is used to verify the extent of asymmetric effects of each type of tightening and loosening measures. This study follows the approach by Kuttner and Shim (2013) and McDonald (2015) in estimating the magnitude of the effectiveness of each instrument. A one year window (or a four-quarter effect) is used to account for the most appropriate lag effects in the implementation of a tightening or loosening of domestic macroprudential policy. A separate index is constructed for each type of prudential instrument. The idea is that a dummy variable is assigned to a value of positive one (1) if all the measures are tightening; 0, otherwise. For loosening measures, a dummy variable is assigned to a value of positive one (1) if all the measures are loosening; or, 0 otherwise. The database includes a measure of the intensity of implementation of prudential policy by considering the number of times a policy is implemented. The the average of these measures is also used.

The study compiles data on the use of domestic prudential instruments. The database shows that majority of the macroprudential measures implemented from 2002 to the fourth quarter of 2017 were currency instruments (41.8% of total), followed by capital-based instruments (29.5%), liquidity-based instruments (13.1%), asset-side instruments (6.1%), and interconnectedness instruments (1.2%). During the same period, a total of 108 tightening measures and 102 loosening measures were recorded. Thirty-four

(34) measures were classified as being neutral, largely pertain to changes in reportorial requirements. On balance, the BSP implemented more tightening than loosening measures. In particular, majority of the tightening measures were capital- and liquidity -related measures for Basel III compliance while those of loosening measures were for currency-related instruments and were implemented in connection to the liberalization of the BSP's foreign exchange framework starting in 2007. Similarly, there were more resilience-based instruments (at 56.3%of the total instruments) adopted compared with cyclical-based instruments (at 43.7%) from the first quarter of 2014 to the fourth quarter of 2017. Of the total measures adopted, 44.3% were tightening measures, 41.8% were loosening and 13.9% were neutral measures.

Measures of monetary policy actions. This database compiles and updates monetary policy actions by the BSP based on Bayangos (2017) database to include Term Deposit Facility (TDF) rates under the Interest Rate Corridor (IRC) system introduced in June 2016. This database is an index of both tightening and loosening policy actions using a four-quarter window based on the estimates. Similar to previous specifications, for each of the central bank official policy rate, a dummy variable is assigned to a value of positive one (1) if the hike in policy rate is accompanied by a rise in TDF rates; hence, the monetary policy stance is tight; 0, otherwise, or when the reduction in policy rate is accompanied by a drop in TDF rates; hence, the monetary policy stance is loose. The database also includes a measure of the intensity of monetary policy actions by considering the number of times a policy is implemented. Taking the average of these measures is also used.

Vector of controls. This dataset includes macro-financial indicators and bank-specific characteristics used in the study. These include changes in real Gross Domestic Product (GDP), inflation, real overseas Filipino remittances, monetary policy rate, TDF rate, bank lending rate, neutral rate of interest rate, output gap, bank credit to GDP ratio gap, nominal peso-dollar rate, real effective exchange rates. The bank-specific characteristics in the dataset include the size of a bank (or total resources in real terms), liquidity ratio defined as liquid assets relative to total assets, capital ratios using capital adequacy ratio and Common Equity Tier 1 ratio to total assets, funding composition using outstanding deposits relative to total liabilities, profitability of banks using real net interest income, and quality of bank loans using nonperforming loans, non-performing assets and non-performing coverage ratio.

Estimation method. In this study, the parameters in the models are estimated using unbalanced panel Generalized Method of Moment (GMM) that is a more appropriate empirical methodology to address the endogeneity between real bank loan commitments and non-performing loans with bank-specific characteristics and macroeconomic indicators. To handle cross-section

fixed effects, data are transformed into first difference. Moreover, residuals are clustered by banks.

Empirical analysis. The empirical analysis includes two parts. The first part estimates the impact of each prudential tool or measure on bank lending to household borrowers as well as the impact on monetary policy conditions and economic cycles. The second part looks at the impact on the loan quality using nonperforming loans by banks.

Impact on bank lending to household borrowers. Using a panel methodology, the impact at the loan level can be seen in equation 1,

$$\Delta \log Loans_{b,t} = a_b + \sum_{j=1}^{k} \gamma_j \Delta \log Loans_{b,t-j} + \sum_{j=1}^{k} \beta_j \Delta MaP_{t-j} + \vartheta X_{b,t-1} + \theta macrovars_{b,t} + \varepsilon_{b,t} \quad \text{(eq. 1)}$$

where $\Delta \log Loans_{b,t}$ is the quarterly change t in the logarithm of loan committed by bank b to a household borrower based on the acquisition cost of the property in real prices over a given period after the introduction or change in a macroprudential tool, a_b are bank fixed effects, $\vartheta X_{b,t-1}$ are bank characteristics, $\theta macrovars_{b,t}$ are macro-financial indicators. The main coefficient of interest is $\sum_{j=1}^{k} \beta_j$ which represents the impact of changes in a domestic macroprudential policy on bank loan commitment to household borrowers.[1]

In arriving at the main dependent variable, this study considered the quarterly growth of appraised value of the residential unit, the appraised value of the lot and the average acquisition cost of the property. These indicators are converted in real terms using the Implicit Price Deflator for Real Estate Activities from the National Income Accounts. Among these three variables, the real average acquisition cost of the property proved to be statistically stable and reliable as the main dependent variable. Moreover, this study considers two separate dummies for tightening actions and loosening actions. Such an approach could help verify asymmetric effects of each prudential tool (Kuttner and Shim, 2016; Bruno et al. 2017). The exercise also considers the intensity or the total number of times each prudential tool has been used, classified by tightening and loosening measures and by resilience- and cyclical- based measures. Moreover, the estimation uses the net intensity on the use of each prudential tool, that is, net tightening and net loosening.

The exercise also estimates how much it takes for a given prudential tool to propagate its effects on lending or the optimal k in equation 1. *Equation 1* considers only the effect after one quarter. However, the propagation effects could be longer, especially with respect to the implementation of a bank loan

[1] In the empirical estimation, the lag effects included contemporaneous impact. However, these contemporaneous estimates did not yield significant results.

commitment. The study considers a specification that considers the sum of one lag to four quarter lag effects to capture the total effects in t as seen in equation 1. In the exercise, the propagation of the effects of changes in domestic macroprudential policy to changes in loan commitment is significant across different regression models up to four quarters.

Do responses to a macroprudential shock differ by type of banks? This section looks at the difference between domestic U/KBs and TBs responses to a macroprudential shock. In the database, there are 101 respondent banks, 38 of which are U/KBs and the remaining 63 are TBs. However, only 56 banks consistently reported residential property loans granted on a quarterly basis from the first quarter of 2014 to the fourth quarter of 2017. Hence, data of these 56 banks are used in all regression models.

To test for the difference, interaction terms that are the product of macroprudential policy indicator and bank-specific characteristic X are included as seen in equation 2,

$$\Delta \log Loans_{b,t} = a_b + \sum_{j=1}^{k} \gamma_j \Delta \log Loans_{b,t-j} + \sum_{j=1}^{k} \beta_j \Delta MaP_{t-j} + \vartheta X_{b,t-1} + \sum_{j=1}^{k} \delta_j \Delta MaP_{t-j} * X_{b,t-1} + \theta macrovars_{b,t} + \varepsilon_{b,t} \quad (eq.\ 2)$$

The test is on the overall significance of $\sum_{j=1}^{k} \delta_j$. This approach builds on the bank lending channel literature. In order to discriminate between loan supply and loan demand movements, the literature has focused on cross-sectional differences between banks. Following Gambacorta (2005), this equation relieson the hypothesis that certain bank-specific characteristics, such as, size, liquidity, the deposit-to-totalfunding ratio and capitalization, influence only the loan supply movements, while a bank's loan demand is independent of these characteristics. This approach basically assumes that after a prudential policy tightening, the ability to shield loan portfolios is different between highly-capitalized and less-capitalized banks.

Do responses to macroprudential policies vary over monetary policy conditions? In this section, additional interaction terms are introduced which combine macroprudential policy indicators and monetary policy actions (measured by the neutral interest rate or NRR based on Taylor rule[2]). This is seen in equation 3 as,

$$\Delta \log Loans_{b,t} = a_b + \sum_{j=1}^{k} \gamma_j \Delta \log Loans_{b,t-j} + \sum_{j=1}^{k} \beta_j \Delta MaP_{t-j} + \sum_{j=0}^{k} \vartheta_j r_{t-j} + \sum_{j=1}^{k} \rho_j \Delta MaP_{t-j} * r_{t-1} + \sigma X_{b,t-1} + \theta macrovars_{b,t} + \varepsilon_{b,t} \quad (eq.\ 3)$$

[2] The neutral interest rate (NRR) is derived as NRR=(10-year average of real 1-year secondary rates)-((real 1-year secondary rates - real 5-year secondary rates) - (real 1-year secondary average - real 5-year secondary average)).

Following Bruno et al (2017), *equation 3* estimates the effectiveness of macroprudential tools when changes in monetary policy push in the same or opposite direction.[3] The test is on the overall significance of $\sum_{j=1}^{k} \rho_j$.

Do responses to macroprudential policies vary over the financial cycles? Additional interaction terms which combine macroprudential policy indicators and real GDP growth (measured by the output gap or the difference between the actual real GDP growth and the average output gap from four approaches[4]). This is seen in equation 4 as,

$$\Delta \log Loans_{b,t} = a_b + \sum_{j=1}^{k} \gamma_j \, \Delta \log Loans_{b,t-j} + \sum_{j=1}^{k} \beta_j \, \Delta MaP_{t-j} + \sum_{j=0}^{k} \vartheta_j \, \Delta logGDP_{t-j} + \sum_{j=1}^{k} \mu_j \, \Delta MaP_{t-j} * \Delta logGDP_{t-j} + \sigma X_{b,t-1} + \theta macrovars_{b,t} + \varepsilon_{b,t} \quad \textit{(eq. 4)}$$

The goal of this exercise is to determine possible presence of endogeneity between output gap and macroprudential tools or their effects may be higher when output gap has widened or vice versa. The test is on the overall significance of $\sum_{j=1}^{k} \mu_j$. In this study, a measure of financial cycle using credit-to-GDP gap or the difference between the actual credit-to-GDP ratio and its trend is used in the regression model.[5] In the exercise, this study also considered separate consumer loans-to-GDP ratios for U/KBs and TBs.

Impact on bank risk. In literature, the use of macroprudential tools is also intended to limit excessive bank risk-taking activities in lending and consequently, the probability of the occurrence of a financial crisis. This study looks at how macroprudential tools have an impact on specific measures of bank riskiness such as gross non-performing loans over total assets. This is seen in equation 5 as,

$$NPL_{b,t} = a_b + \sum_{j=1}^{k} \gamma_j \, \Delta \log NPL_{b,t-j} + \sum_{j=1}^{k} \beta_j \, \Delta MaP_{t-j} + \vartheta X_{b,t-1} + \theta macrovars_{b,t} + \varepsilon_{b,t} \quad \textit{(eq. 5)}$$

Where a_b are bank fixed effects, $\vartheta X_{b,t-1}$ are bank characteristics, $\theta macrovars_{b,t}$ are macro-financial indicators. The main coefficient of interest is $\sum_{j=1}^{k} \beta_j$ which represents the impact of changes in a domestic macroprudential policy on bank risk-taking activities as seen in non-

[3] In the estimation of $\sum_{j=0}^{k} \vartheta_j \, r_{t-j}$, the contemporaneous impact is considered.

[4] These approaches include (1) production function approach, (2) structural vector autoregression (SVAR), (3) macroeconomic unobserved components model (MUCM), and (4) Hodrick-Prescott (HP) filter.

[5] Credit-to-GDP gaps are derived, in line with the Basel III guidelines for the countercyclical capital buffer, as the deviations of the credit-to-GDP ratios from their (real-time) long-term trend. Consumer loans-to-GDP was also used in the estimation.

performing loans (NPL). Following Chavan and Gambacorta (2016), $NPL_{b,t}$ is a logit function of the ratio of gross NPL to total loans for bank *b* at time *t*. In particular, the logit function is given in equation 6 as,

$$NPL_{b,t} = \ln\left[\frac{NPL}{1-NPL\ ratio}\right] \quad \textit{(eq. 6)}$$

Given the perceived persistence of NPL, this study uses a dynamic specification that includes lagged value of the NPL as an explanatory variable. The study considers a specification that takes into account the sum of one lag to four quarter lag effects to capture the total effects in *t*. Similar to specifications in *equations 1 to 4*, the bank-specific characteristics $\vartheta X_{b,t-1}$ in equation 5 includes the size of a bank (or total resources in real terms), liquidity ratio, capital ratios using capital adequacy ratio and Common Equity Tier 1 ratio to total assets, funding composition using outstanding deposits relative to total liabilities, and profitability of banks using real net interest income.

Robustness checks. Diagnostic tests are used to check for normality of residuals across equations at 1%, 5% and 10% levels of significance. The results are broadly robust against normality tests and different specifications of dependent and independent variables. The residual tests show that all estimated coefficients are significant and that the instruments used are not correlated with the residuals (using Hansen test). The standard errors of regression are robust and that the errors in the first difference regression exhibit no second order serial correlation (using serial correlation test).

3. Results

Following diagnostic and robustness checks, the results reveal important findings. First, tightening of domestic prudential, particularly those tightening measures that are meant to preserve resilience of the banking system, are effective in curbing growth of real bank loan commitments to borrowers for acquiring new residential properties. The results show that tightening macroprudential policies have direct and negative impact that can last up to four quarters on real bank loan commitments to borrowers based on real acquisition cost of new properties from March 2014 to December 2017. Importantly, results reveal that tightening domestic macroprudential policies vary with both business and financial cycles. Overall, these findings confirm other studies' observation that prudential policies are more likely to find effectiveness.

Second, this study highlights the bigger negative impact of tightening prudential measures on real bank loan commitments to maintain banks' resilience in by U/KBs compared to TBs, an indication of the presence of bank lending channel. Third, real bank loan commitments to household borrowers are driven by bank deposits (relative to total liabilities), liquidity position and

capital adequacy (gap relative to the regulatory threshold). Moreover, monetary policy tightening complements tightening of prudential policies in restraining growth of real bank loan commitments. Meanwhile, real exchange rate appreciation reacts to tightening of prudential measures.

Fourth, in general, restricting prudential measures limits risk-taking activities by banks. The results show the negative impact of tightening domestic macroprudential measures on non-performing loans relative to total loans. Meanwhile, the results reveal that the impact of both business (output gap) and financial cycles (credit-to-GDP ratio) on the movements of NPL ratio are positive and significant. However, when prudential measures are adopted, the impact on the NPL ratio becomes negative.

In general, the results are consistent with analysis that despite the relative rise in the level of bank loan portfolio, banks have been more risk-sensitive[6], although the level of bank loan portfolio is not the focus of these estimations. Moreover, the response of the growth of NPL to shocks to growth in TLP appears to be modest, highlighting that growth in NPL remains stable amid growth in TLP.

4. Conclusion

This study examined the effectiveness of changes in a comprehensive measure of domestic prudential policies in restraining the growth of real loan commitments of U/KBs and TBs to borrowers for new purchases of residential property and riskiness of these banks' loan portfolio using panel data regression from the first quarter of 2014 to the fourth quarter of 2017. There are improvements that the study intends to pursue moving forward. From the technical point of view, the study intends to explore the use of aging of non-performing loans of U/KBs and TBs in assessing the extent of the risk-taking activities by banks and to examine the impact of domestic macroprudential policies on net interest margins of banks. Moreover, the study aims to use difference-in-difference analysis to check the robustness of results and to assess the effects of domestic macroprudential policies on the supply of loan in greater detail.

The use of credit registry data will be a future research area to assess the impact of domestic macroprudential policies on household and firm credit risk. Matching firm balance sheet information with credit registry data could help us to fill this gap. The approval into law of the creation of a Credit Information System on 31 October 2008 known as Republic Act No. 9510, "An Act Establishing the Credit Information System and for other purposes" and the establishment of the Credit Information Corporation (CIC) to address the

[6] See Cachuela, Rafael Augusto (2018). Technical Box Article: "Does Expansion of Bank Lending Leads to Weakening of Loan Quality in the Philippines", Status Report on the Philippine Financial System, First Semester 2018. Bangko Sentral ng Pilipinas.

need for comprehensive, centralized and reliable credit information system is indeed a significant development. The main purpose of the Corporation is to strengthen the submission of basic credit data, both positive and negative credit information in the entire data subject provided by submitting entities.

Nevertheless, the study's findings have important policy implications. First, the finding that tightening domestic macroprudential policies are effective in reducing growth of real bank loan commitments underscores the critical role for structural policies to enhance the capacity of the economy to cope with volatility, along with improved regulation and supervision of the financial sector. Second, given the influence of real effective exchange rate appreciation in driving growth in real loan commitments, there is a need for more in-depth understanding of exchange rate dynamics, its impact on the economy and the effectiveness of policy instruments, both in the short and longer term, as well as the risk-taking channel of currency appreciation.

Third, an important point to consider is the role of domestic macroprudential measures on cross-border issues. The cross-border effects of prudential measures can be both positive and negative. The positive effect concerns the public good aspect of financial stability, wherein actions enhancing financial stability in one country also benefit others. Policies that prevent the build-up of systemic risk in one jurisdiction may reduce the probability of crises that subsequently spread elsewhere. And fourth, the finding that tightening of domestic macroprudential policies restricts risk-taking activities by banks underscores the role of bank supervision and the resulting microprudential policy in managing risks to banking stability in the Philippines. Importantly, the BSP, cognizant that a "one size fits all" framework is not appropriate for all banks, adheres to the principle of proportionality in the adoption and application of certain prudential regulations.

References

1. Bayangos, V. (2017), "Capital Flow Measures and Domestic Macro Prudential Policy in Asian Emerging Economies: Have These Been Effective?" Bangko Sentral ng Pilipinas Working Paper Series No. 201701, June.
2. Bruno, V, I Shim and H Shin (2017): "Comparative assessment of macroprudential policies", Journal of Financial Stability, Vol 28, pp 183 – 202, February.
3. Bruno, V., I. Shim and H.S. Shin (2015), "Comparative Assessment of Macro Prudential Policies," BIS Working Paper No. 502, Bank for International Settlements, June.
4. Chavan, P. and L. Gambacorta (2016), "Bank lending and loan quality: the case of India", BIS Working Paper No. 595, Bank for International

Settlements, December 2016. Gambacorta, L. (2005): "Inside the bank lending channel", European Economic Review, Vol. 49, pp 1737 – 1759.

5. Kuttner, K N and I Shim (2016): "Can non-interest rate policies stabilize housing markets? Evidence from a panel of 57 Economies". Journal of Financial Stability, Vol 26, pp 31-44, October.
6. Kuttner, K. and I. Shim (2013), "Can Non-Interest Policies Stabilise Housing Markets? Evidence from a Panel of 57 Economies," BIS Working Paper No. 433, Bank for International Settlements.
7. McDonald, C. (2015), "When is Macroprudential Policy Effective?", BIS Working Paper No. 496, Bank for International Settlements, March. Orsmond, D. and F. Price (2016). "Macroprudential Policy Frameworks and Tools" Reserve Bank of Australia Bulletin, December.

Effectiveness of policies in addressing household indebtedness: Evidence from credit registry data in Malaysia

Muizz Aziz, Siow Zhen Shing
Central Bank of Malaysia

The authors would like to thank Qaiser Iskandar Anwarudin, Cheah Su Ling, Karen Lee, Siti Hanifah Borhan Nordin, Rafidah Mohamad Zahari and Nik Ahmad Rusydan bin Nik Hafizi for their invaluable assistance and feedback. The authors can be contacted at muizz@bnm.gov.my and siowzs@bnm.gov.my.

Disclaimer: This paper represents the views of the authors and may not necessarily be those of Bank Negara Malaysia (BNM) or BNM policy. The views expressed herein should therefore be attributed to the authors and not to BNM.

Abstract

In this study, we assess the impact of macroprudential policies on debt service ratio (DSR) of households using borrower-level data from the credit registry. The findings are twofold. First, policies are found to be generally effective in improving borrowers' DSR levels for personal financing, particularly among low-income borrowers who are typically highly leveraged. Second, we found that the effectiveness of the policies in improving borrowers' DSR levels for residential property loan may be constrained by the simultaneous increase in house prices. These insights lend support for a more holistic approach in addressing household indebtedness.

Keywords

Macroprudential Policy; Financial Regulation; Household Debt; Household Debt Service Ratio

1. Introduction

In the decade since the Global Financial Crisis, the immense economic costs arising from the vulnerabilities in the credit and housing markets have brought macroprudential policies into sharper focus. Traditionally, the most common tools in policymakers' toolbox have been monetary and microprudential policies. While monetary policy proved to be too blunt, microprudential policy, on the other hand, tend to have limited optics on system-wide financial vulnerabilities. In turn, a growing number of countries have embraced macroprudential policies to fill the gap, by adopting a more targeted and systemic approach to financial regulation and supervision.

Macroprudential policies have been used in Malaysia to address systemic risk to the financial system since the early 1990s[1]. In the recent decade, the Government and the Central Bank of Malaysia (the Bank) implemented a series of measures to address excessive accumulation of household debt and to promote a more sustainable housing market. Since the implementation of these measures, credit-fuelled speculative purchases of residential properties (measured by the annual growth of the number of borrowers with three or more residential property loans) declined to 1.7% as at end-2018 (2010: 15.8% (peak)). Personal financing, which also drove the earlier rapid debt accumulation by households, moderated significantly to 2.3% as at end-2018 (2008: 25.2% (peak)). Against this backdrop, the household debt-to-GDP ratio declined to 82.1% (2015: 86.9% (peak)). More importantly, this decline occurred without adversely affecting private consumption and economic growth.

Two notable studies assessing the effectiveness of macroprudential policies in Malaysia against the objective that they were designed to achieve, found supportive evidences. Rauf (2017) found that introduction of additional policies is associated with a decline in growth of residential property loans. Of significance, policies affecting the supply and demand of credit are found to be more effective vis-à-vis fiscal policies that affect the cost of homeownership. Similarly, A. Rani and Lau (forthcoming) found that a limit on loan-to-value (LTV) ratio did dampen demand for residential property loans although the effects appear to diminish over time. Both studies measured the effectiveness of macroprudential policies at the bank- and banking system-level, respectively.

Our study extends existing work, by assessing the impact of macroprudential policies on the debt service ratio[2] (DSR) of households, at the borrower-level. The paper is organised as follows: in Section 2, we will provide an overview of the Malaysian household debt over the years and the macroprudential policy framework in Malaysia; in Section 3, we describe the data and methodology used; in Section 4, we present our results; while in Section 5, we draw some conclusions and outline possible policy implications of our work.

[1] A series of macroprudential policies were implemented to curb large capital inflows that led to a strong growth in assets prices, which includes, among others, a cap on LTV ratio and a limit on the expansion of bank financing to less productive economic sectors.

[2] The ratio of total monthly bank and non-bank debt obligations to monthly disposable income (net of statutory deductions).

2. Background and Stylised Facts

Household debt in Malaysia stood at RM1,188 billion, or 82.1% of GDP as at end-2018. The current elevated [3] state of household indebtedness in Malaysia is preceded by a period of rapid debt accumulation. The annual growth of household debt hovered around 10% to 14.2% between 2008 and 2013, significantly above income growth. This trend was primarily driven by loans for the purchase of residential properties and personal use, which accounted for 52.3% and 14.5% of household debt, respectively, as at end-2018 (2008: 48.4% and 10.6%, respectively) (**Chart 1**).

The acceleration in the growth of residential property loans coincided with the spurt in house prices. The annual growth of Malaysian House Price Index (MHPI) increased from 5.6% in 4Q 2009 to 14.3% in 4Q 2012 (peak) (**Chart 2**). The MHPI trend during this period could be attributed to, among others, more relaxed credit underwriting practices of some lenders, increased speculative activities under an environment of low interest rates and inflation, tax regime that was less punitive to speculators (e.g. low Real Property Gains Tax), and the continued strong demand for residential properties especially in urban areas such as Kuala Lumpur, Selangor and Johor Bahru.

Meanwhile, the growth of personal financing largely reflected the increasing role of non-bank financial institutions (NBFIs) in providing financing to households, coupled with aggressive marketing and advertising strategies by both banks and NBFIs to attract customers. This includes development financial institutions, which lent heavily to the low-income group. On the demand side, sustained income growth bolstered households' confidence and shaped optimistic expectations on their future incomes, which subsequently, incentivised households to borrow in order to smooth their consumption over the life cycle.

Against this backdrop, the aggregate leverage (measured as a ratio of outstanding debt to annual income) of individual borrowers in Malaysia stood at elevated levels. A closer look at the leverage ratio by income group revealed a stark difference, with the low-income borrowers (those who are earning <RM3,000 per month) being more vulnerable to financial distress given higher leverage of 8.8 times as at end-2018 (**Chart 3**). This group also held the largest share of personal financing to their total borrowings, compared to other income groups. Their conditions were further compounded by continued reduction in housing affordability, as higher house prices require them to take on bigger loan amount. Compared to other income groups, low-income borrowers experienced the highest increase in house prices (**Chart 4**). Given

[3] Malaysia's household debt-to-GDP ratio remains high, compared to regional and rating peers (Emerging market economies aggregate as at end-2018: 40%; Singapore: 67%; Thailand: 69%; China: 53%; Chile: 45%).

limited financial buffers and susceptibility to potential macroeconomic or financial shocks, lowering their debt obligations to more sustainable levels would be imperative.

Chart 1: Contribution to Growth of Household Debt by Purpose

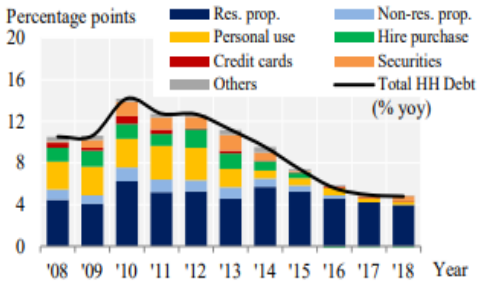


Chart 2: Annual Growth of Malaysian House Price Index

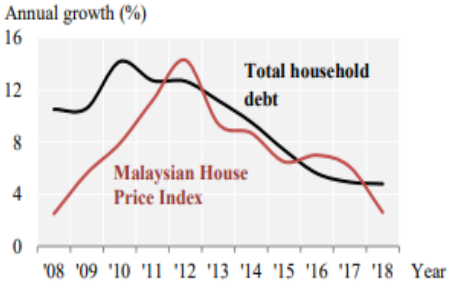


Chart 3: Aggregate Leverage by Income Group

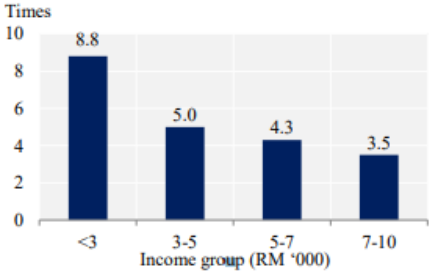
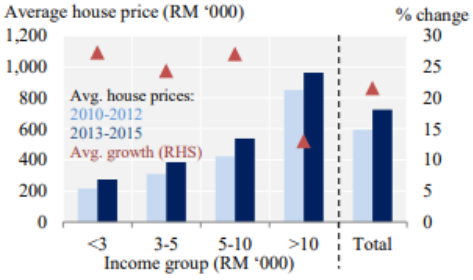


Chart 4: Average Growth of House Prices Purchased by Borrowers



Source: Bank Negara Malaysia and National Property Information Centre

Concerns over the sustainability of household debt and housing market led the Government and the Bank to implement a series of measures aimed at addressing risks associated with excessive accumulation of household debt and to reinforce responsible lending practices by both banks and NBFIs (**Table 1**). The series of policies implemented between 2010 and 2013 (shaded rows in Table 1) will be the focus of this study.

Table 1: Policies Implemented since 2010

Date	Policy
4Q 2010	Introduction of maximum 70% loan-to-value (LTV) on 3rd residential property loan and above
1Q 2011	Stricter credit card requirements such as introduction of minimum eligibility criteria based on income and age
1Q 2011	Higher risk weights for capital adequacy requirements for residential property loans with LTV ratio >90%
4Q 2011	Introduction of maximum 60% LTV on residential property loan for non-individuals
1Q 2012	Implementation of Guidelines on Responsible Financing - 2Q12: Implicit DSR limit of 60% for the low-income borrowers
3Q 2013	Introduction of maximum loan tenure: - Personal financing: 10 years - Residential property loans: 35 years - Car loans: 9 years
3Q 2013	Implementation of Policy Document on Personal Financing that prohibits offering of pre-approved personal financing products
4Q 2013	Prohibition of developers' interest bearing scheme and related financing by the financial institutions
4Q 2013	Implementation of Policy Document on Risk-Informed Pricing
4Q 2015	Introduction of minimum collective impairment provisions and regulatory reserves of 1.2%

Note: These measures were also complimented by monetary and fiscal measures (e.g. Real Property Gains Tax).

3. Data and Empirical Methodology

This study utilises data from the public credit registry (Central Credit Registry Information System, CCRIS), administered by the Bank. CCRIS collates information of all borrowers who obtained credit from financial institutions regulated by the Bank and selected large NBFIs. Our data is a repeated cross-section covering a sample of newly-approved personal financing and residential property loan borrowers from 1Q 2009 to 4Q 2015.

The macroprudential policies that were implemented between 2010 and 2013 can affect potential borrowers in two ways.

First, the policies may have the effect of keeping some individuals out of the credit market. For instance, some individuals may decide not to apply for a loan at all after the implementation of these measures. For those who choose to apply for a loan, they may be rejected[4] by the financial institutions if they are applying for a loan that is beyond their affordability level. To measure this effect, we examine whether there are any significant changes in the

[4] In many instances, those who were rejected do not have enough income or were unable to pledge adequate equity.

distribution of income of approved borrowers, using the Kolmogorov-Smirnov test (Tzur-Ilan, 2018).

Second, prospective borrowers who were able to obtain financing after the implementation of the policies, may have opted to take on a smaller loan that is more in line with their affordability. Consequently, we expect borrowers' DSR for personal financing and residential property loan to decline after the implementation of the policies. To measure this impact, we estimate the following regression using ordinary least squares for personal financing and residential property loan borrowers, respectively (Bekkum et al., 2019):

$$y_{i,l,t} = \alpha_l + \beta Policy_t + \theta X_{i,t} + \epsilon_{i,l,t} \quad (1)$$

$$y_{i,l,t} = \alpha_l + \beta_1 Policy_t + \beta_2 LowInc_i + \beta_3 Policy_t^* LowInc_i + \theta X_{i,t} + \epsilon_{i,l,t} \quad (2)$$

where i indexes individual borrowers, l indexes state of residence, and t indexes time (quarters). The dependent variable, $y_{i,l,t}$, is the borrower's DSR. $Policy_t$ is a policy index, which would increase by 1 if a new tightening measure is introduced and decline by 1 if a measure is relaxed (Cerutti et al., 2017). However, it is important to note that no relaxation of measures took place during the period of estimation. The state fixed effects (α_l) control for fixed differences across regions, such as housing affordability. We also control for changes in the overnight policy rate (OPR) as this may have a direct impact on the borrowers' DSR. $X_{i,t}$ is a vector of control variables measured at the borrower-level, and $\epsilon_{i,l,t}$ the error term. The main parameter of interest, β, measures the average borrowers' DSR in the period after the implementation of the policies relative to (unconstrained) borrowers acquiring loans in the period before. We also estimate equation (2) to investigate the impact of the policies on the low-income group.

For home buyers that were impacted by the policies, they may have adjusted their housing choices by buying cheaper houses that may be smaller, farther from the city centre or in a lower quality neighbourhood compared to the average borrower. To measure this, we estimate equation (1) and (2) with house price as the dependent variable.

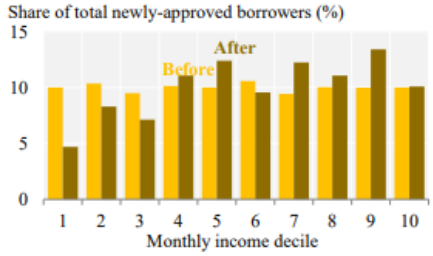
4. Results

A. Changes in the Distribution of Borrower Characteristics

From the Kolmogorov-Sminov test, we found that the distribution of income of newly-approved borrowers is statistically different after the implementation of the policies, for both personal financing and residential property loan (**Chart 5 and 6**). Notably, the average income of the population is higher after the introduction of the measures. This is within expectations - with the improvement in loan affordability assessment which takes into

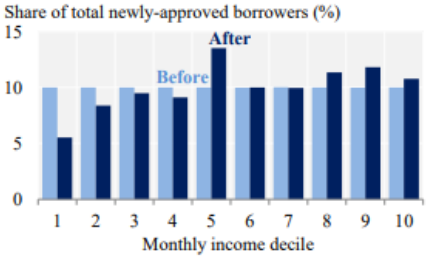
consideration the cost of living and the strong growth in house prices, some low-income borrowers, particularly those who are already highly leveraged, will no longer be able to qualify for financing

Chart 5: Income Distribution for Newly-Approved Personal Financing Borrowers



Range of income deciles (RM) – 1: <1.6k; 2: 1.6k-2.3k; 3: 2.3k-2.9k; 4: 2.9k-3.5k; 5: 3.5k-4.6k; 6: 4.6k-5.8k; 7: 5.8k-7.3k; 8: 7.3k-11.2k; 9: 11.2k-24.6k; 10: >24.6k

Chart 6: Income Distribution for Newly-Approved Residential Property Loan Borrowers



Range of income deciles (RM) – 1: <2.9k; 2: 2.9k-4k; 3: 4k-5.5k; 4: 5.5k-6.9k; 5: 6.9k-9.5k; 6: 9.5k-11.6k; 7: 11.6k-15.2k; 8: 15.2k-22.3k; 9: 22.3k-43.2k; 10: >43.2k

Source: Bank Negara Malaysia

B. Impact of policies on borrowers' DSR

For those who are able to obtain personal financing, we found that the coefficient for *policy* is consistently negative and significant (**Table 2, Eq 1**). On average, the introduction of one additional policy is associated with 1.32 percentage points (ppt) reduction in DSR for personal financing. We also extend the study to see the impact of the policies on the low-income group. For this group, we found that the DSR for personal financing declined more than the other income groups after the implementation of the policies (**Table 2, Eq 2**).

Interestingly, we found that the coefficient for *policy* for those who are able to obtain residential property loans is positive, controlling for borrower specific factors such as income and age, and changes in the overnight policy rate (**Table 3, Eq 1a, 1b and 2a**). This may suggest that while the policies are expected to reduce borrowers' DSR for residential property loan, there are opposing factors, which were not considered in this equation. We postulate that rising house prices during the period have led to borrowers having to take on larger home financing, placing upward pressure on their DSR. Controlling for house prices, we found that this is indeed the case (**Table 3, Eq 1c and 2b**).

Finally, using house price as the dependent variable, we found that on average, borrowers purchased more expensive homes even after the introduction of the policies (**Table 4, Eq 1 and 2**). This is within our expectations given accelerated MHPI growth in the period of estimation. While

the point estimate of 0.060 (**Table 4, Eq 2b**) indicates that the average borrower bought houses that are priced 6.0 ppt higher after an additional measure, the low-income group, on the other hand, purchased cheaper homes compared to average borrowers, in line with their affordability (**Table 4, Eq 2**).

Table 2: DSR for Personal Financing

	Eq 1a	Eq 1b	Eq 1c
Policy	-0.966***	-1.333***	-1.324***
State	Y	Y	Y
Income	Y	Y	Y
OPR	Y	Y	Y
Age	N	Y	Y
Tenure	N	N	Y
R^2	0.137	0.163	0.1634

	Eq 2a	Eq 2b
Policy	-0.413*	-0.898**
Low Income	13.082***	7.079***
Interaction (Low Inc. & Policy)	-1.227***	-1.051***
State	Y	Y
OPR	Y	Y
Other controls	N	Y
R^2	0.117	0.178

Legend: * p<.05; ** p<.01; *** p<.001

Table 3: DSR for Residential Property Loan

	Eq 1a	Eq 1b	Eq 1c
Policy	0.818***	0.730***	-0.220***
State	Y	Y	Y
Income	Y	Y	Y
OPR	Y	Y	Y
Age	N	Y	Y
Tenure	N	N	Y
House price	N	N	Y
R^2	0.320	0.321	0.514

	Eq 2a	Eq 2b
Policy	0.923***	-0.124***
Low Income	23.323***	9.871***
Interaction (Low Inc. & Policy)	0.704***	1.062***
State	Y	Y
House price	N	Y
OPR	Y	Y
Other controls	N	Y
R^2	0.126	0.533

Legend: * p<.05; ** p<.01; *** p<.001

Table 4: House Price

	Eq 1a	Eq 1b	Eq 1c
Policy	0.105***	0.088***	0.066***
State	Y	Y	Y
Income	Y	Y	Y
OPR	Y	Y	Y
Age	N	Y	Y
Tenure	N	N	Y
R^2	0.261	0.281	0.495

	Eq 2a	Eq 2b
Policy	0.111***	0.060***
Low Income	-0.750***	-0.440***
Interaction (Low Inc. & Policy)	-0.029***	-0.044***
State	Y	Y
OPR	Y	Y
Other controls	N	Y
R^2	0.190	0.512

Legend: * p<.05; ** p<.01; *** p<.001

5. Discussion and Conclusion

In this paper, we assess the impact of macroprudential policies on the DSR of households, using borrower-level credit registry data. We found that the policies are generally effective in improving borrowers' DSR for personal financing, particularly those in the low-income group who are typically highly leveraged. However, for residential property loans, we found that macroprudential policies alone are insufficient, as borrowers' DSR for residential property continued to increase after the measures. This can be

attributed to the rising house prices that took place during the same period. This finding lends support for a more holistic approach in addressing indebtedness of the household sector.

We acknowledge that the methodology applied in this paper has some limitations that could be explored for further research. These include expanding the scope to assess (i) the impact of the policies on the different forms (e.g. hire purchase and credit cards) and sources (e.g. banks and nonbank financial institutions) of credit extended to households, and (ii) the impact of other policies introduced including Basel III-related prudential requirements on indebtedness of households.

References

1. A. Rani and Lau (forthcoming). Implementing Loan-to-Value and Debt-to-Income Ratios: Learning from Country Experiences – Malaysia
2. Bekkum S. V., Gabarro M., Irani R. M. & Peydro (2019). Macroprudential Policy and Household Leverage: Evidence from Administrative Household-Level Data
3. Cerutti, E., S. Claessens and L. (2017). The Use and Effectiveness of Macroprudential Policies: New Evidence. Journal of Financial Stability, Vol. 39, pp. 153-185
4. Rauf (2017). Measuring the Effectiveness of Macroprudential Policies – The Malaysian Experience
5. Tzur-Ilan N. (2018). LTV Limits and Borrower Risk. Bank of Israel Discussion Paper

Nowcasting advance estimate of U.S. quarterly personal consumption of services

Baoline Chen, Kyle Hood
Bureau of Economic Analysis Washington, DC 20230

Abstract

This study evaluates two nowcasting techniques—the general bridge equation framework and bridging with factors model—for compiling advance estimates of personal consumption expenditure (PCE) of services in the U.S. national accounts. The proposed approaches use current and lagged monthly and quarterly information to provide more accurate information on the longer-term dynamics and short-term movements in the quarterly target variables. In addition to the monthly indicators assigned to each component, the bridging with factor model employs common factors extracted from a set of all available monthly price and quantity indicators for the service sector to provide general information on the economic and business conditions of the sector. We apply the proposed approaches to the detailed components of PCE services with the objective of reducing revisions in the advance estimates that occur when quarterly information becomes available. We measure improvement in accuracy in terms of reduction in the root mean squared revision statistic.

Keywords

National economic accounts, Now-casting, Bridging with factors, Real time data

1. Introduction

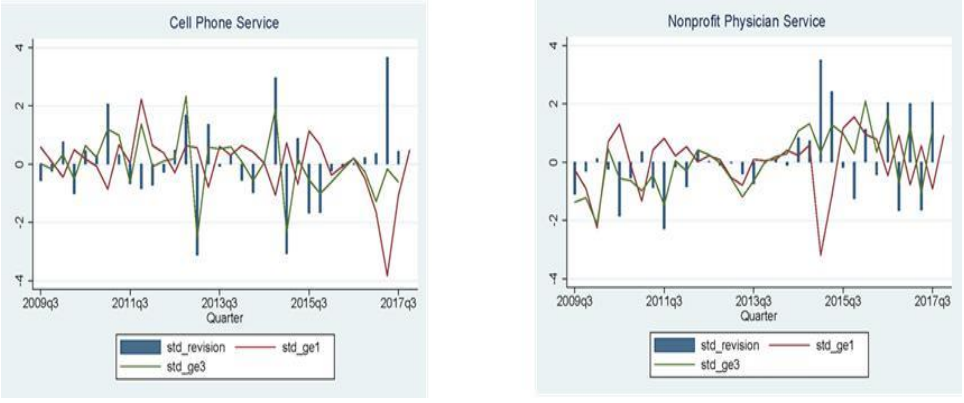
For each quarter, the U.S. quarterly GDP is released three times with a one-,two- and three-month delay, respectively. The first or "advance" release is roughly one month after a quarter has ended. The second release is roughly two months after a quarter has ended and the third and final release is roughly three months after a quarter has ended. The first and second releases are primarily based on selected monthly indicators because quarterly source data for the GDP components are not yet available. By the time the third estimates are being compiled, quarterly source data have become available. Because of delays in the quarterly source data needed for compilation, early releases are often later revised. Revisions for certain components may sometimes exceed 5%, especially during periods of rapidly changing macroeconomic environments.

Among the major components of GDP, personal consumption expenditure (PCE), which consists of PCE goods and services, constitutes two thirds of total

GDP, with PCE services accounting for roughly two thirds of PCE. Because the data used to compile PCE goods and services come from different sources, compilation of PCE goods and services present different challenges to the national accounts. As delays in the quarterly source data affect the first two estimates of PCE services similarly, in this study, we focus on the first or the advance estimate of detailed PCE services.

There are four major reasons for revisions in the advance estimate of PCE services: 1) source data used to compile the three estimates are from difference sources and become available at different frequencies; 2) there are an insufficient number of relevant monthly indicators available for all detailed components; 3) there is lack of close correlation between the monthly and quarterly indicators; and 4) information included in the current extrapolation method is insufficient to reflect the longer-term dynamics of the quarterly PCE service component series and the monthly indicators.

Figure 1: Revision in advance estimate of selected PCE services components



Given delays in the quarterly source data available for compiling early estimates, reducing revisions in the early estimates hinges on producing them more accurately using all available information for the quarter that has just ended. This amounts to a nowcasting problem, which is defined as the prediction of the present, the very near future and the very recent past (Bańbura, Giannone, Modugno and Reichlin, 2013). The basic principle of nowcasting is to exploit information published early and possibly at higher frequencies than the target variable in order to obtain a more accurate early estimate before official estimate based on quarterly source data becomes available.

The objective of this study is to introduce two nowcasting approaches to compile advance estimates of detailed PCE services—the bridge equation (GB) framework and the bridging with factors (BF) model. These approaches are capable of exploiting information on the longer-term dynamics of PCE services as well as the short-term movements in the monthly indicators driven by the changes in the economic conditions. We apply these approaches to compiling

advance estimates of detailed PCE services with the objective of reducing revisions when quarterly source data become available. We measure improvements in accuracy in terms of reduction in the root mean squared revision (RMSR) for each detailed component.

The plan for the paper is as follows: Section 2 describes the current extrapolation method and the proposed bridge equation and bridging with factor frameworks for nowcasting the advance estimate. Section 3 describes the application and the strategy for estimation and nowcasting. Section 4 reports estimation results. Section 5 discusses further research and concludes the paper.

2. Current and Alternative Methods for Compiling Advance Estimate of Detailed PCE Services

Currently, the U.S. national accounts compile advance estimate of PCE services using a simple extrapolation method which takes two steps: 1) extrapolating monthly estimates from the previous month for each of the three months in the quarter using monthly indicators; and 2) computing the advance quarterly estimate as the quarterly averages of the three monthly estimates. The current extrapolation method uses information on the monthly indicators for the current quarter. However, it does not utilize information on the longer-term dynamics of the PCE services, nor does it utilize information on the longer-term dynamics of the monthly indicators. To reduce revisions in the advance estimate of PCE services, we need to allow lagged information on the quarterly PCE services and the current and lagged information on the relevant monthly indicators to be included in the estimation. Since compilation of advance estimate of PCE services is equivalent to a nowcasting problem, we consider two widely-used nowcasting techniques in this study, the bridge equation framework and the bridging with factors model.

The bridge equation approach was first developed by Klein et al. (1989) and has since been further developed and implemented in many studies (Kitchen and Manaco, 2003; and Higgins, 2014). It has been described as a tracking model which tracks quarterly growth in real GDP by tracking the arrival of new information in real time.

In this study, we express variables in growth rates. Let y_t^A denote the quarterly growth rate of the advance estimate of a detailed PCE services component in quarter t; and let $\bar{x}_t = (\bar{x}_{1,t,} \dots, \bar{x}_{s,t})$ denote the quarterly growth rates of the quarterly averages of s monthly indicator variables. For each detailed PCE service component, the general bridge equation framework can be expressed as

(1) $$y_t^A = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{s} \sum_{i=1}^{k} \delta_{j,i} \bar{x}_{j,t-i} + \varepsilon_t,$$

where α is a constant, p is the number of autoregressive parameters, k is the number of lags of the indicator variables, and $\varepsilon_t \sim iid\ N(\mu_\varepsilon, \sigma_\varepsilon^2)$.

Although bridge equations allow for multiple high-frequency indicator variables, our small sample size limits the number of high-frequency variables that could practically be included in the bridge equation. To be able to utilize information from a much larger set of monthly indicators in a parsimonious regression framework, we also consider the bridging with factors model, which replaces the monthly indicators in the bridge equation framework with a small number of common factors to capture the main co-movement of a much larger set of indicators (Giannone, Reichlin and Small, 2008).

Let $F_t = (f_{1,t}, f_{2,t}, \dots, f_{r,t})'$ be the vector of r common factors from monthly factor models aggregated to the quarterly frequency, where $r \ll \min(n, T)$, n is the number of monthly indicator variables and T is the number of quarterly observations in the sample. For each PCE services component, bridging with factor model can be expressed as

(2) $$y_t^A = c + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{r} \sum_{i=0}^{k} \gamma_{j,i} f_{j,t-i} + \varepsilon_t,$$

where c is a constant, p is the number of autoregressive parameters, k is the number of lags of factor j; and $\varepsilon_t \sim iid\ N(\mu_\varepsilon, \sigma_\varepsilon^2)$. The number of factors is determined according to the Bai-Ng criterion proposed by Bai and Ng (2002).

The added advantage of the bridging with factors model is that it allows us to extract common factors from all available monthly indicators for all PCE services and use them in the estimation. Because we have insufficient number of designated monthly indicators for all detailed PCE service components, bridging with factors also allows us to incorporate information on the general business conditions of the service sector via extracted common factors in the estimation.

3. Application: Bridge Equation and Bridging with Factors Models for Estimation of PCE Services

We apply the bridge equation framework and the bridging with factors model outlined in the previous section to compile advance estimate of detailed PCE services using real time data from the U.S. national accounts. 121 detailed components from 9 sub-groups of PCE services are included in the application. To be able to compare revisions with the current extrapolation method, we use the real-time data that were used to compile the advance estimate of detailed PCE services from 2009Q2 to 2017Q4, a maximum of 34 quarters. Quarterly data used in the application include quarterly growth rates

of the first and third estimates of the of detailed PCE services from the 2009Q2 to 2017Q4 vintages; monthly data include percentage changes in population, average wage earnings for the relevant PCE services, and the corresponding consumer and producer price indices (CPI and PPI) from the 2009M7 to 2017M12 vintages. Each vintage of the quarterly data include four lagged quarterly growth rates and each vintage of the monthly indicators includes lagged values to compute 4 lagged quarterly growth rates of the monthly indicators.

For estimation using bridge equations, we allow a maximum of 4 lagged growth rates of the third quarterly estimate of the PCE services and the current and a maximum of 4 lagged quarterly growth rates of the monthly indicators. For estimation of the bridging with factors model, a monthly factor model is first estimated using a total of 92 monthly indicators designated for the components in the 9 sub-groups of PCE services. Two common factors are selected according to the Bai-Ng criterion. Like estimation with bridge equations, we allow a maximum of 4 lagged quarterly growth rates of the target quarterly variable and the current and 4 lagged quarterly growth rates of the selected common factors aggregated from the monthly factors.

The estimation and nowcasting exercise is done in three steps: 1) in-sample estimation using 75% of the sample; 2) one-step-ahead pseudo out-of-sample predictions using the remaining 25% of the sample; and 3) comparison of the proposed methods with the current method. To fully utilize the information from our small samples, we choose the recursive approach to compute out-of-sample predictions. The in-sample estimation is conducted using the STATA program VSELECT and the number of explanatory variables for each PCE service component is determined according to the Akaike information criteria corrected for small samples (AICC).

To examine the impact of the outliers on estimation results, we also estimate models with the outliers ($|y_{t-i}| \geq 3\sigma$) removed. Thus, we evaluate four models: bridge equation (M1), bridge equation without outliers (M2), bridging with factors (M3), and bridging with factors without outliers (M4). We measure improvements in accuracy by comparing model root mean squared revisions (RMSR) with those from the current method.

4. Results from In-Sample Estimation and Out-of-Sample Predictions

4.1 Results from in-sample estimation

Results from the in-sample estimation validate the choices of using bridge equations and bridging with factors models to compile advance estimate of detailed PCE services. Table 1 shows results for selected PCE service components from the estimated bridge equations. Column 1 identifies the regression models. Columns 2 and 3 identify the PCE services components and the service groups they belong to. Columns 4 to 8 display the estimated

coefficients of the constant, the lagged quarterly growth rates of the selected PCE services, and columns 9 to 13 contain the estimated coefficients of the current and lagged quarterly growth rates of the monthly indicators. Numbers in parentheses are the t-statistics for the estimated coefficients. The last two columns display the adjusted R^2's and the p-values for the F-tests.

Table 1: In-sample estimation of selected PCE services from bridge equations

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Model	PCE GROUP	HLC CODE	β_0	β_1	β_2	β_3	β_4	α_0	α_1	α_2	α_3	α_4	Adj R^2	Prob > F
M1	HLC	OPO	.0029 (1.85)	.8994 (6.45)		.2205 (2.01)				.5115 (2.11)			0.76	0.00
M2	HLC	OPO	.0016 (.83)	.7253 (6.58)	.3663 (2.32)	-.2075 (-1.23)							0.73	0.00
M1	TRS	PFT	.0002 (.87)	1.5175 (24.95)	-.6109 (-7.78)	.2171 (5.78)		1.0010 (93.78)	-1.4177 (-23.38)	.5992 (7.57)	-.2196 (-6.05)		0.99	0.00
M2	TRS	PFT	.0002 (.15)	1.0315 (16.73)				.9568 (11.35)	-.9917 (-9.98)				0.95	0.00
M1	RCA	PIC	.0001 (.41)	1.4022 (23.44)	-.6120 (-8.07)	.2306 (6.87)		.9950 (.78.35)		2.2217 (14.04)	-.8236 (-7.86)	.2295 (6.07)	0.99	0.00
M2	RCA	PIC	-.0001 (-.13)	1.0726 (14.77)	-.0877 (-2.58)			.9721 (27.43)		1.0101 (12.39)			0.98	0.00
M1	PER	SCL	.0000 (13)	1.2630 (20.04)	-.5484 (-8.56)	.3280 (9.54)		.9896 (73.04)	-1.2615 (-19.07)	.5514 (8.72)	-.3263 (-8.77)		0.99	0.00
M2	PER	SCL	.0073 (4.41)			.6653 (4.62)					-.7182 (-3.47)		0.46	0.00
M1	SOC	SHO	-0.0048 (.90)					1.0153 (2.66)					0.21	0.01
M2	SOC	SHO	.0001 (-.03)	-.9010 (-5.09)		.1567 (1.88)	-.1494 (-2.51)	.6013 (9.92)	.8919 (7.54)	.4316 (4.68)	-.1119 (-1.56)		0.95	0.00

Note: HLC-Health care (OPO-Nonprofit other medical service); TRS-Transportation (PFT-Parking fees); RCA-Recreational services (PIC-Cinema); PER-Personal services (SCL-Repair of footwear); SOC-Social services (SHO-Nonprofit home for elderly).

The main observations from in-sample estimation are that 1) lagged quarterly growth of PCE services and lagged monthly indicator variables (or lagged common factors in the bridging with factors model) play a significant role in the estimation of advance estimate of most PCE service components; 2) dynamic characteristics of each service component and of its indicators determine the choice of the model for compiling its advance estimate; 3) in estimation using the bridging with factors model, common factors extracted from all available monthly indicators of the 9 PCE service groups play a significant role in compiling the advance estimate; and 4) as shown in the estimated M1 and M2 models in Table 1, outliers do affect selection of the variables in the estimated model.

Lagged growth rates of the PCE services are shown to impact the estimation. Although the maximum number of lagged growth rates is set to 4, for most service components fewer than 4 lags were selected by AICC. Similarly, fewer than 4 lags of the indicators (or common factors) were selected. For some PCE services, no lagged growth rates of the quarterly target variables were selected at all; whereas for some other services only lagged growth rates of the quarterly target variables but no current or lagged growth

rates of the indicators were selected. Based on the dynamic characteristics of PCE services component and the dynamics of its indicators (or common factors), different combinations of lagged dependent variables and current and lagged independent variables were selected.

4.2 Results from out-of-sample prediction

Because our samples have short time spans, to efficiently use the information from the data, we chose the recursive estimation approach to compute pseudo-out-of-sample predictions. This means that each one-step-ahead prediction is computed from estimated model using data up to the quarter prior to the quarter being predicted. We measure improvement in accuracy by reductions in the RMSR relative to the RMSR of the current extrapolation method.

Of the 121 PCE service components, 36 component series use identical indicator data for the first and the third estimates. For these components, revisions are zero or close to zero with any minor revisions coming from revisions in the indicator data. The remaining 85 components use distinct source data to compile the first and the third estimates. Our evaluation of improvement in accuracy is based on the changes in the RMSR of these 85 components.

The main observations from the one-step-ahead pseudo-out-of-sample predictions are that 1) out-of-sample predictions from both bridge equation and bridging with factors models resulted in reductions in RMSR for 63 PCE service components that used distinct indicators for the first and the third estimates; 2) out-of-sample predictions from estimated bridge equations outperformed those from bridging with factors model in 48 out of 63 components (76%); 3) out-of-sample predictions from both bridge equation and bridging with factors models led to reductions in the RMSR at the sub-group aggregates of the PCE services, and 4) degrees of reduction in the RMSR varied from the one-step-ahead predictions computed from different estimated models and differed across service groups.

Table 2: Percentage changes in the RMSR from bridge equation and bridging with factor models for professional services (PRS)

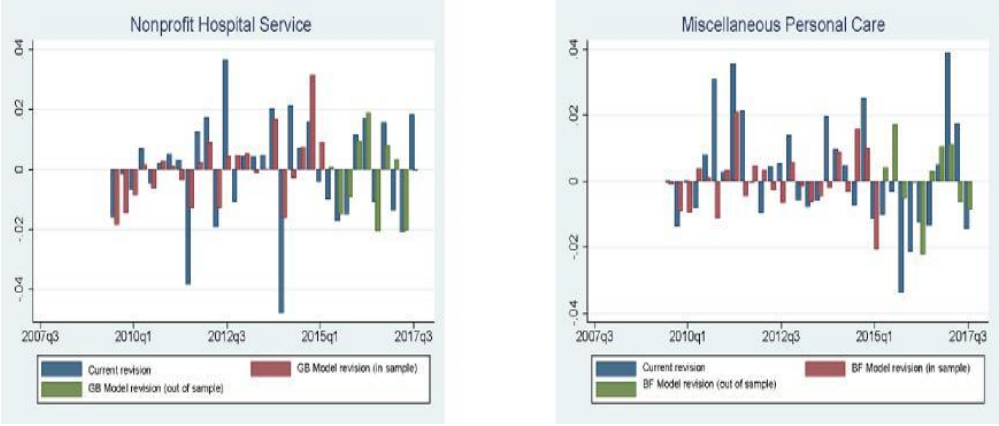
PCE_PRS	ΔRMSR_BE Model	ΔRMSR_BE Outlier Removed	ΔRMSR_BF Model	ΔRMSR_BF Outliers Removed
GAL	-48.15	-44.36	-27.64	-27.64
GAH	-48.04	-44.10	-29.49	-29.49
AXS	-43.62	-10.83	-25.28	-25.28
AXO	-59.76	-47.08	-44.97	-38.90
AOO	-74.15	-66.03	-62.21	-68.52

Note: GAL-Legal service; GAH-Private and public legal service; AXS-Nonprofit professional association service; AXO-Private professional association service; AOO-All other organizational service.

Table 2 shows the percentage changes in the RMSR from the one-step-ahead pseudo-out-of-sample predictions relative to the RMSR using the current extrapolation method for the components in professional services (PRS). Negative values indicate reductions in the RMSR from the one-step-ahead predictions. Reductions in the RMSR are seen in all 5 detailed professional services from all models, ranging from 11% to 74%. For the 5 components from professional services, reduction in the RMSR is the largest from the one-step-ahead predictions computed from the estimated bridge equations. The stronger performance of the bridge equation framework is also evident in other PCE service groups

We can graphically compare revisions from the in-sample and out-of-sample estimation from the estimated bridge equation and bridging with factors models with those from the current extrapolation method. Figure 2 illustrates revisions in nonprofit hospital services and miscellaneous personal care services. Bars shown in blue are revisions from the current extrapolation method, bars in red are revisions computed from the in-sample estimation, and bars in green show revisions from the one-step-ahead predictions computed from the estimated bridge equations (left) or bridging with factors models (right). Not surprisingly, revisions from the in-sample estimation are generally smaller than those from the out-of-sample predictions. Another observation is that revisions from both in-sample and out-of-sample estimation are noticeably smaller in the periods where the current extrapolation method resulted in large spikes in the revisions. However, we also notice that a revision reduction is not seen in every quarter.

Figure 2: Revision comparison from estimated bridge equation model for selected PCE services



5. Conclusion

In this study we have demonstrated that bridge equation framework and bridging with factors model are potentially useful methods for compiling advance estimates of detailed PCE services, because these methods allow all available information on the dynamics of the quarterly target variables and the monthly indicators to be incorporated in the estimation. The one-step-ahead out-of-sample predictions from these two models resulted in reductions in the RMSR. However, we do not see reductions in revisions for every component, nor in every period. To explore further improvements in accuracy, we plan to experiment with various forecast combination and model averaging techniques.

References

1. Bańbura M., D. Giannone, M. Modugno and L. Reichlin, (2013) "*Now-casting and the real-time data flow*, European Central Bank, Working Paper Series: no. 1564, pp. 1-54, European Central Bank.
2. Giannone, D., L. Reichlin, and D. Small (2008): "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55(4), 665–676.
3. Klein, L.R. and E. Sojo, (1989) "Combinations of High and Low Frequency Data in Macroeconometric Models," pp. 1-13 in Economics in Theory and Practice: An Eclectic Approach, L.R. Klein and Marquez (eds.), Kluwer Academic Publisher.
4. Kitchen, J. and R. Monaco, (2003) "Real-Time Forecasting in Practice: The U.S. Treasury Staff's Real-time GDP Forecast System," MPRA Paper: no. 21068, U.S. Department of Treasury.
5. Higgins, P., (2014) "GDP Now: A Model for GDP 'Now-casting,'" Working paper: 2014-7, Federal Reserve Bank of Atlanta.

Forecasting quarterly GDP at real-time monthly intervals using Bayesian Linear Least-Squares Methods*

Jonathan Weinhagen[1]; Peter Zadrozny[2]**

[1] Bureau of Labor Statistics Division of Industrial Prices 2 Massachusetts Avenue, NE, Room 3865 Washington, DC 20212

[2] Bureau of Labor Statistics Division of Price Index Number Research 2 Massachusetts Avenue NE, Room 3105 Washington, DC 20212

Abstract

The paper proposes and illustrates a Bayesian method that requires only linear least-squares methods for estimating a monthly VAR model of GDP, employment, and industrial production using quarterly observations on GDP and monthly observations on employment and industrial production in order to forecast GDP at monthly intervals, using the latest available monthly real-time information. The Bayesian method is Theil and Goldberger's (1962) mixed estimation method that is used to impose equality restrictions on model coefficients at different degrees of Bayesian tightness (cf., Shiller, 1974; Litterman, 1986). The restrictions reflect the implication of stationarity that VAR coefficients of the same variables and the same implied monthly lags should be equal. The GDP forecasts of the best-forecasting model were about 9% lower in root mean squared error (RMSE) than those of a baseline univariate AR model. Tight restrictions resulted in slightly worse forecasting models with higher RMSEs. The methodological contribution of the paper is that its method of "stacking" a model for mixed-frequency data at the lowest frequency immediately generalizes to any number and types of frequencies.

1. Introduction

At any moment a forecaster has available only real-time data that have been released up to that time. Economic data are available at different frequencies, some at monthly or shorter intervals, others at longer intervals. For example, GDP is observed quarterly and employment (EP) and industrial production (IP) data are observed monthly. Different econometric methods have been used to estimate vector autoregressive moving-average (VARMA) models with mixed frequency data (MFD). For example, Zadrozny (1990a,b) first discussed and illustrated estimating a VARMA model of quarterly GNP and monthly employment using maximum likelihood estimation (MLE) and, then, using the estimated model and Kalman filtering to forecast the GNP at monthly intervals. However, MLE, especially applied to MFD, is difficult to

* This work represents the authors' views and does not necessarily represent any official positions of BLS.

** Corresponding author.

apply. Even today, it requires special computer programing that is often not included in available statistical and econometric programs. Even with a program in hand, MLE requires some experience in setting starting values of iterative nonlinear computations so that they converge successfully. Moreover, as a model gets larger with more variables and more parameters to estimate, the top of the likelihood tends to get flatter in all parameter dimensions, so that convergence, if it can be achieved at all, starts to take an impractical amount of time.

Therefore, other estimation methods have been developed to avoid these problems. Although Bayesian methods can be equally or more computationally intensive, they are, at least mathematically, simpler than MLE because using them one doesn't require numerically scaling a peak, but only requires simulating a peak and computing some of its sample moments. Properly programmed, a Bayesian method should compute conclusively, although that may take a long time.

Therefore, there has been a need for quicker, reliable, linear methods for estimating VARMA models using MFD. Chen and Zadrozny (1998) proposed and illustrated the extended Yule-Walker (XYW) method for estimating a VAR model with MFD, which is a linear generalized least squares (GLS) method. Instead of estimating VAR or VARMA models using MFD with feedbacks both from high-frequency to low-frequency variables and vice versa, Ghysels et al. (2007) introduced the more easily implemented mixed-data sampling (MIDAS) which regresses one or more low-frequency variables onto one or more high-frequency variables using an exponential distributed lag with few parameters to estimate. The basic "stacking" idea in MIDAS of estimating a model of mixed-frequency data in low-frequency form originated with Friedman (1962).

Given the desire to estimate a VARMA model using MFD in a simple yet effective way, in this paper we describe and illustrate a stacking method for estimating a monthly VAR model using monthly-quarterly data. The stacking introduces a relatively large number of additional, possibly insignificant, parameters to be estimated. Here the number of these additional parameters to be estimated is reduced by equating feedbacks of the same variables at the same lags in different months of quarters. The restrictions are implied by stationarity and are implemented using Theil and Goldberg's (1961) linear Bayesian mixed-estimation strategy. Ghysels (2016) addresses the parameter proliferation problem arising from stacking slightly differently, partly with a MIDAS-like exponential distributed lag.

2. Econometric Method

We propose estimating a stacked monthly VAR model using monthly-quarterly data, first expressed as

(1) $Ay_t + By_{t-1} + \mu_t,$

where $y_t = (y_{1t}, \ldots, y_{nt})^T$, denotes an nx1 vector of observed variables, t denotes quarters, A and B denote nxn matrices of constant parameters (or functions of constant "deeper" parameters), $u_t = (u_{1t}, \ldots, u_{nt})^T$ denotes an nx1 vector of unobserved disturbances, and superscript T denotes vector or matrix transposition. We assume (a) that A is a sub-lower triangular matrix with all elements on or above the principal diagonal equal to zero and (b) that the disturbance covariance matrix is the identity matrix, $\sum_u = 1_n$. The parameter elements of A, B, and $\sum_u$ are identified and estimated efficiently (unbiasedly with minimum variance) by sequentially applying OLS to eqution (1), equation by equation, from top to bottom.

The structure of coefficient matrix A reflects the Granger-causal ordering of the variables in y_t induced by the order in which they are observed in a quarter. Thus, y_{1t} in y_t is observed first in quarter t, y_{2t} is observed second in quarter t, and so on. Forecasting with estimated equation (1) is similarly
sequenced: y_{1t} is forecast using the first equation, y_{2t} is then forecast using the second equation and the previously computed forecast of y_{1t} , and so on.

Equation (2) illustrates equation (1) with only quarterly GDP and monthly industrial production, so that stacked $y_t = (IP_{1t}, IP_{2t}, IP_{3t}, GPD_t)^T$ and equation (1) is

(2)
$$\begin{bmatrix} IP_{1t} \\ IP_{2t} \\ IP_{3t} \\ GDP_t \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \alpha_{21} & 0 & 0 & 0 \\ \alpha_{31} & \alpha_{32} & 0 & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 0 \end{bmatrix} \begin{bmatrix} IP_{1t} \\ IP_{2t} \\ IP_{3t} \\ GDP_t \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \\ \beta_{31} & \beta_{32} & \beta_{33} & \beta_{34} \\ \beta_{41} & \beta_{42} & \beta_{43} & \beta_{44} \end{bmatrix} \begin{bmatrix} IP_{1,t-1} \\ IP_{2,t-1} \\ IP_{3,t-1} \\ GDP_{t-1} \end{bmatrix} + \begin{bmatrix} U_{1t} \\ U_{2t} \\ U_{3t} \\ U_{4t} \end{bmatrix}.$$

Because IP_{1t} is observed first in quarter t, it depends only on values of variables observed in the previous quarter. Because IP_{2t} is observed second in quarter t, it depends on IP_{1t} first observed earlier in quarter t and on values observed in the previous quarter. The remaining two scalar equations in vector equation (2) are structured analogously.

In practice, economic variables are observed with delays after the periods they represent have passed. Here, for simplicity, the delays are ignored, so that the "real-time" analysis is really a "pseudo real-time" analysis. To state a monthly VAR model of quarterly GDP and monthly IP in stacked quarterly form, monthly IP must be considered as three different quarterly variables, $IP_{1t}, IP_{2t}, and\ IP_{3t}$ = IP of the first, second, and, third months of quarter t, respectively. An example of the present equality restrictions mitigating the parameter proliferation problem that arises from stacking is as follows: in (2), α_{21}, β_{12}, β_{23} reflect the same 2-month feedback of industrial production onto itself, so that stationarity suggests that $\alpha_{21} = \beta_{12} = \beta_{23}$.

To impose the coefficient restrictions across any scalar equations in vector equation (1), we first write equation (1) for each t = 1, ..., T as

(3) $y_t = Cx_t + U,$

where $C = [A, B]$ =nx2n and $x_t = (y_t^T, y_{t-1}^T)^T$ = 2nx1 and, then, in transposed form for all t as

(4) $Y = XC^T + U,$

where $Y = [y_t, \dots, y_T]^T$ =nxT, $X = [x, \dots, x_T]^T$= 2nxT, and $U = [u_t, \dots, u_T]^T$=nxT

Consider column vectorization rule $vec(ABC) = [C^T \otimes A]vec(B)$, where, for the moment, A, B, and C denote any matrices conformable to the matrix multiplication ABC, vec(◦) denotes the columnwise vectorization of a matrix (column one on top of column two, etc.) and ⊗ denotes the Kronecker product. Applying the rule to equation (4), gives

(5) $\bar{y} = (1_n \otimes \mathrm{X})\gamma + \bar{u},$

where $\bar{y} = vec(Y)$ =nTx1, $\gamma = vec(C^T)$ = 2n^{2}x1, and $\bar{u} = vec(U)$ = nTx1. The n-variable generalization of the pattern of A and B in 4-variable equation (2) implies that

(6) $\gamma = (0, \dots 0, \beta_{11}, \dots, \beta_{1n}, \alpha_{21}, 0, \dots 0, \beta_{21,} \dots, \beta_{2n,} \alpha_{31}, \alpha_{32}, 0, \dots, 0, \beta_{31}, \dots \beta_{3n},$
$\alpha_{n1}, \dots, \alpha_{n,n-1}, \beta_{n1}, \dots, \beta_{nn})^T.$

Because zeros in γ make corresponding columns of $(I_n \otimes X)$unnecessary, we eliminate these unnecessary elements of γ and columns of $(I_n \otimes X)$ and write equation (7) as

(7) $\bar{y} = \bar{X}\bar{\gamma} + \bar{u},$

where $\bar{X}$ and $\bar{\gamma}$, respectively, denote $I_n \otimes X$ and γ with the unnecessary columns and elements removed, so that $\bar{\gamma}$ = kx1, where k denotes the number of non-identically zero elements of γ.

Theil-Goldberger mixed-estimation imposition of equality restrictions on elements of $\bar{\gamma}$ goes as follows by adding pseudo data to the bottom of the data matrix $[\bar{y} \quad \bar{X}]$. In exact form, parameter equality restrictions may be expressed as $0 = R\bar{\gamma}$, where R is a qxk matrix of 0, 1, and -1 elements and q denotes the number of restrictions on the k elements of $\bar{\gamma}$, so that q < k. According to the mixed estimation strategy, we impose restrictions with a chosen degree of looseness or tightness regulated by positive scalar λ, such that a higher value of λ indicates greater Bayesian tightness. Thus, we extend $0 = R\bar{\gamma}$ to

(8) $0 = \lambda R\bar{\gamma} + v$

where v is an error or pseudo error analogous to $\bar{u}$. For example, if q = 1 and $\bar{\gamma}$ has three elements, then, $\bar{y}_1, \bar{y}_2 \; and \; \bar{y}_3$ are restricted by $\bar{y}_1 = \bar{y}_3 \; and \; R = [1, 0, -1]$.

The Bayesian tightness restrictions work as follows in OLS. We append equation (8) to the bottom of equation (7) and obtain

$$(9) \quad \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix} = \begin{bmatrix} \bar{X} \\ \lambda R \end{bmatrix} \bar{\gamma} + \begin{bmatrix} \bar{u} \\ v \end{bmatrix},$$

which in general is a (nT+q)x1 "mixed" regression equation with data matrix $\begin{bmatrix} \bar{y} & \bar{X} \\ 0 & \lambda R \end{bmatrix}$, coefficient vector $\bar{\gamma}$, and real-and-pseudo-error vector $(\bar{u}^T, v^T)^T$. OLS works as follows with differing values of λ . OLS sets $\bar{\gamma}$ to $\ddot{\bar{\gamma}}$ so as to minimize the least squares measure of the overall residual vector, $(\ddot{\bar{u}}^T, \hat{v}^T)^T$, in particular, to maintain a minimum least squares balance of $\bar{y} \cong \bar{X}\bar{\gamma}$ and $0 \cong \lambda R\bar{\gamma}$. As λ increases, $\ddot{\bar{\gamma}}$ must be ever more directed toward maintaining $0 \cong \lambda R\bar{\gamma}$ so that the coefficient restrictions are enforced ever more.

The methodological contribution of the paper is that its method of "stacking" a model for mixed-frequency data at the lowest frequency immediately generalizes to any number and types of frequencies of observation of data. In fact, there's no need to distinguish among observation frequencies. All one needs to use the method is to write model (1) (or an extension of it that includes more lags of variables) in stacked form in terms of the lowest frequency of observation, that reflects the Granger-causal ordering of the variables induced by the order in which they are observed in each lower-frequency observation period.

3. Application to Forecasting Quarterly GDP at Monthly Intervals

This section discusses application of the above method to forecasting quarterly GDP at monthly intervals. Model (1) is defined in terms of employment (EP), industrial production (IP), and GDP. EP and IP are observed monthly and GDP is observed quarterly. The model was estimated for these variables for 5 values of Bayesian tightness, λ = 0 (no tightness), 1, 10, 100, and 1000. Observations on the variables from January 1996 to March 2017 were used and were obtained from the real-time database at the Federal Reserve Bank of Philadelphia. The variables are ordered in the model according to dates at which they are available publicly in each quarter. The data are used in standardized (minus sample means, divided by sample standard deviations) period-to-period percentage-change form. EP and IP are further denoted by their release months as EP1, EP2, EP3, IP1, IP2, and IP3. Thus, in the application, the stacked-form model contains 7 variables time-indexed in quarters, t.

Model (1) was sequentially estimated and used to forecast, as follows, The

model was first estimated using data through quarter 4 of 1995. The estimated model was used to forecast GDP for quarters 1-4 of 1996. The model was then reestimated using in addition data for quarter 1 of 1996. The reestimated model was used to forecast GDP for quarter 2 of 1996 to quarter 1 of 1997. This process was repeated moving forward in time until the data were exhausted in March 2017.

Accuracy of forecasting GDP at quarterly and monthly intervals is compared for 4 strategies: using an unrestricted benchmark quarterly univariate AR(1) model estimated conventionally and abbreviated as UAR; a conventionally formulated and estimated quarterly VAR(1) model, abbreviated as VAR; and model (1) under two "scenarios", abbreviated, respectively, as M1S1 and M1S2. The two scenarios are considered in order to guage the contribution of real-time monthly informaton to the accuracy of the GDP forecasts. In scenario 1 (M1S1), intra-quarterly monthly information handled by the first right-side term in equation (1) is ignored in forecasting GDP; in scenario 2 (M1S2), the right-side term and its monthly information are included in the forecasting. Forecast accuracy is measured by root-mean squared errors (RMSE). Table 1 reports RMSEs of the 4 strategies.

There are 12 cases in Table 1. The table shows that UAR forecasts are more accurate than forecasts of the other strategies in only 4 of 12 cases (algebraically larger table entries). M1S1 and M1S2 forecasts are more accurate than VAR forecasts (algebraically smaller table entries). M1S2 forecasts are significantly more accurate than M1S1 forecasts in all cases.

Table 1: % differences in RMSEs compared to UAR, with no coefficient restrictions

Model	1 quarter ahead	2 quarts ahead	3 quarts ahead	4 quarts ahead
UAR	0.0000	0.0000	0.0000	0.0000
VAR	-5.6295	0.2346	-0.4644	0.4298
M1S1	-5.6800	0.1883	-0.4937	0.4126
M1S2	-9.6184	-11.8477	-2.7153	0.0341

No coefficient restrictions are imposed on either M1S1 or M1S2 in Table 1, so that the Bayesian tightness parameter is $\lambda = 0$. Table 2 extends Table 1 by considering different degrees of Bayesian tightness on M1S1 and M1S2. The value $\lambda = 100$ enforces equality up to 3 decimal digits. Per M1S1 and M1S2 strategy, increasing tightness results in higher RMSE and, hence, in lower GDP forecast accuracy, so that, in this application, ignoring the restrictions and setting $\lambda = 0$ results in the most accurate forecasts. The restrictions should be tested with other data. Other data may find them beneficial or not. If not, then,

at least considering their natural provenance from stationarity, their rejection should motivate thinking about why a considered model may or may not be misspecified.

Table 2: % differences in RMSE compared to UAR, with Bayesian coefficient restrictions

	λ	1 quart ahead	2 quarts ahead	3 quarts ahead	4 quarts ahead
M1S1	0	-5.680	0.1883	-0.4937	0.4126
	1	-5.634	0.2730	-0.4685	0.4207
	10	-3.692	2.359	0.2009	0.7340
	100	-2.733	3.200	0.4821	0.8845
	1000	-2.719	3.212	0.4863	0.8867
M1S2	0	-9.618	-11.84	-2.715	0.0341
	1	-9.618	-11.85	-2.691	0.0392
	10	-9.618	-11.45	-1.634	0.5007
	100	-9.618	-11.18	-1.087	0.7715
	1000	-9.618	-11.17	-1.080	0.7755

4. Conclusion

The paper has described and illustrated a method for forecasting a low frequency variable using higher frequency variables. The method allows forecasts to incorporate all relevant information (data) that has been released prior to the time the forecast is made. Theil-Goldberger mixed estimation was used to consider equality restrictions on coefficients indicated by statonarity in several degrees of Bayesian tightness. The paper illustrates the method by forecasting quarterly GDP at monthly intervals using U.S. monthly employment and industrial production data from January 1995 to March 2017. The results clearly show that using the latest available monthly employment and industrial production data significantly improves accuracy of GDP forecasts. However, in all cases considered, imposing the equality restrictions at any Bayesian degree of tightness resulted in slightly to significantly less accurate GDP forecasts and never improved them.

References

1. Chen, B. and P.A. Zadrozny (1998), "An Estimated Yule-Walker Method for Estimating Vector Autoregressive Models with Mixed-Frequency Data," *Advances in Econometrics* 13: 47-73.
   Friedman, M. (1962), "The Interpolation of Time Series by Related Series," *Journal of the American Statistical Association* 57: 729-757.
2. Ghysels, E. (2016), "Macroeconomics and the Reality of Mixed Frequency Data," *Journal of Econometrics* 193: 294-314.

3. Ghysels, E., A. Sinko, and R. Valkanov (2007), "MIDAS Regressions: Further Results and New Directions," *Econometric Reviews* 25: 53-90.
4. Litterman, R.B. (1986), "Forecasting with Bayesian Vector Autoregressions: Five Years of Experience," *Journal of Business and Economic Statistics* 4: 25-38.
5. Shiller, R. (1974), "A Distributed Lag Estimator Derived from Smoothness Priors," *Econometrica* 41: 775-788.
6. Theil, H., and A.S. Goldberger (1961), "On Pure and Mixed Statistical Estimation in Economics," *International Economic Review* 2: 65-78,
7. Zadrozny, P.A. (1990a), "Estimating a Multivariate ARMA Model with Mixed-Frequency Data: An Application to Forecasting U.S. GNP at Monthly Intervals," working paper, Research Dept., Federal Reserve Bank of Atlanta and Center for Economic Studies, U.S. Census Bureau, Washington, DC.

Nowcasting finnish real economic activity: A machine learning approach

Paolo Fornaro[1], Henri Luomaranta[2]
[1] Research Institute of the Finnish Economy
[2] Statistics Finland and University Of Toulouse 1

Abstract
We develop a nowcasting framework based on micro-level data in order to provide faster estimates of the Finnish monthly real economic activity indicator, the Trend Indicator of Output (TIO), and of quarterly GDP. In particular, we rely on firm-level turnovers, which are available shortly after the end of the reference month, to form our set of predictors. The nowcasts are obtained from a range of statistical models and machine learning methodologies which are able to handle high-dimensional information sets. The results of our pseudo-real-time analysis indicate that a simple nowcasts' combination based on these models provides faster estimates of the TIO and GDP, without increasing substantially the revision error. Finally, we examine the nowcasting accuracy obtained by relying on traffic data extracted from the Finnish Transport Agency website and find that using machine learning techniques in combination with this big-data source provides competitive predictions of real economic activity.

Keywords
Flash Estimates, Machine Learning, Micro-level Data, Nowcasting

1. Introduction

Statistical agencies, central banks, research institutes and private businesses have access to (and produce) thousands of economic and financial indicators. However, this wealth of information has not been directly translated into a faster and more accurate production of important economic statistics, such as the GDP. Statistical institutes publish economic indicators with considerable lag and the initial estimates are revised considerably over time. The advantages of having a timely picture of the state of the economy are multiple and concern a range of economic actors such as the central bank, the government and private investors and businesses. Providing this type of information in a timely manner would be invaluable, because it would contribute in reducing the uncertainty of the current state of the economy, thus leading to better informed decisions. The economic advantages of having a timely picture of the economy have not been disregarded by the statistical and academic community.

Nowcasting and the production of economic activity indicators in real time have been the focus of a growing literature. An early work related to the tracking of economic conditions in real time by creating new high-frequency indicator is Aruoba, Diebold, and Scotti (2009). The nowcasting literature is interested in estimating an existing economic indicator (usually quarterly GDP growth) in real-time. Few examples drawn from the nowcasting literature are Giannone, Reichlin, and Small (2008) and Evans (2005) among many others.

In this study, we combine firm-level datasets and machine learning techniques, as well as traditional statistical models which can deal with large datasets, to provide faster estimates of Finnish real economic activity, both at the quarterly and monthly frequencies. The monthly series we target is the Trend Indicator of Output (TIO)[1], published by Statistics Finland, while the quarterly series is GDP. For both series we compute nowcasts of the year-on-year growth rate. In addition, we examine the predictive power of a novel dataset based on traffic volumes' measurements. The use of novel data sources, such as firm-level turnovers data and traffic measurements, in combination with the use of a wide array of machine learning techniques provides the main contribution of our study to the nowcasting literature.

Our approach of combining predictions obtained by using a large set of machine learning algorithms, based on firm-level data, is able to provide accurate estimates of monthly economic activity growth, with revision errors that are in line with the ones of Statistics Finland, while shortening the publication lags by 30 days. The resulting early estimates of the monthly indicator are used to compute three nowcasts of GDP year-on-year growth. The first two nowcasts provide good accurancy, even though there are some notable revision errors. However, the estimates produced after the end of the quarter are very accurate, while providing a 45 days reduction in the publication lag. We conduct a similar analysis using truck traffic volumes' measurements, and find satisfactory results, albeit inferior to the ones obtained from firm-level data.

The remainder of this paper is divided as follows: in Section 2 we discuss some of the large set of models adopted in the analysis, in Section 3 we describe our target indicators and data sources. In Section 4 we report the results and Section 5 concludes.

2. Methodological Aspects

Given that the main contribution of this study is the use of novel data sources, we keep the description of the models adopted brief. This section does not cover comprehensively the techniques we use for two reasons: firstly, the sheer number of statistical models and machine learning techniques

[1] A description of this indicator is available at http://www.stat.fi/til/ktkk/index_en.html

adopted in the exercise does not allow a thorough discussion. Secondly, these techniques have been used in previous econometric or statistical studies, hence a detailed description would be superfluous. We instead try to give the basic intuition underlying some of the main classes of models used and redirect the interested readers to the original works in which the models we employ were originally developed or to some previous applications in which these models are adopted.

Among the most important models considered in the nowcasting literature we have the dynamic factor model, in the form of Stock and Watson (2002). The basic idea is that a handful of constructed variables, the factors, can summarize the information contained in a large dataset. Stock and Watson (2002) have shown that the factors can be estimated using principal components.[2] Factor models are especially important in our application because, in addition to the basic specifications including raw firm-level data and traffic data as predictors, we estimate specifications where we utilize latent factors (estimated via principal components) as predictors. This is done to see whether reducing the noise in our input data improves the performance of the models.

Another important class of models we use is shrinkage regression, in particular the ridge regression, the Lasso (Tibshirani, 1996) and the elastic-net (Zou and Hastie, 2005). These main intuition of these models is to regularize the coefficients of the predictors, in order to reduce the predictions' variance. Hastie, Tibshirani, and Friedman (2009) provides an in-depth review of these models, while De Mol, Giannone, and Reichlin (2008) offers an economic forecasting application of shrinkage regressions, with an interesting comparison with principal components.

Our nowcasts are then based on a large number of machine learning techniques, which are covered extensively in Hastie et al. (2009): boosting, regression trees and random forests, regression splines, support and relevance vector machines, neural networks and k-nearest neighbors.

All the models utilized in our nowcasting exercise are estimated using the caret package for R. Once considering specifications with different input variables (raw data vs. sets of principal components extracted from the data), we arrive at a total of 130 models to estimate. As benchmark model, we utilize an automated ARIMA procedure. Moreover, we include in our models set an automated ARIMA where we include principal components as external predictors.

[2] An alternative factors estimator can be found in Doz, Giannone, and Reichlin (2011).

3. Data description

The target variables in our exercise are the Trend Indicator of Output (TIO) and quarterly GDP, both measured in real-term year-on-year growth rates. The TIO is a monthly series that describes the development of the volume of produced output in the economy. It is constructed by using early estimates of turnover indexes (not publicly available), which are appropriately weighted to form the monthly aggregate index. The TIO is published monthly at t + 45, and its value for the third month of a quarter is used to compute the flash estimate of GDP, which is also published as an early version at t + 45, and updated at t + 60. The t + 60 version is considered as the first official and reliable estimate of GDP. Thus, given the information we have provided, the TIO in fact represents a GDP nowcast in its own right. We stress the importance of using the realistic vintages, as the data is typically "improved" by many internal processes, and by the accumulation of new data.

The main predictors in our nowcasting application are firm-level sales extracted from the sales inquiry, a monthly survey conducted by Statistics Finland for the purposes of obtaining turnovers from the most important firms in the economy. This dataset covers around 2,000 enterprises and encompasses different industries (services, trade, construction, manufacturing), representing ca. 70% of total turnovers. The data is available soon after the end of the month of interest and a considerable share of the final data is accumulated around 15 to 20 days after the end of the reference month. Formally, Statistics Finland imposes a deadline to the firms, which are supposed to send their data by the end of the 15th day of the month. We compute the nowcast on the 16th day. However, this deadline is not always met, thus our set of firms' sales does not cover the entire sample.

As alternative source of predictors, we examine traffic loop data for real-time estimation purposes, and consider the predictive performance of traffic volumes records obtained by the Finnish Transport Agency website[3]. This dataset contains the number of vehicles passing through a number of measurement points, observed through an automatic traffic monitoring system. For our nowcasting analysis, we collect data for trucks' traffic volumes, in particular their year-on-year growth rate at the different measurement points across the country.

4. Empirical results

The TIO is a monthly indicator of real economic activity. Our nowcasting exercise is centered on providing fast estimates for the year-on-year growth rate of TIO, starting from March 2012 (the first month for which we have the vintage of the data) and ending in December 2018. We start by reporting the

[3] The data is available at https://aineistot.liikennevirasto.fi/lam/reports/LAM/

results of the models which provide the lowest root mean squared error (RMSE), the lowest mean error (ME), mean absolute error (MAE), and finally for the model with the lowest maximum absolute error (MaxE). In addition, we report the results for the simple forecast combination consisting of the unweighted average of the nowcasts with mean error (in absolute terms) below the 20th percentile. As benchmark, we report the results obtained by using an automated ARIMA procedure.

	Lowest ME	Lowest RMSE	Lowest MAE	Lowest MaxE	Combination	ARIMA
ME	-0.00	-0.25	-0.25	-0.25	-0.01	0.11
MAE	1.06	0.75	0.75	0.75	0.78	1.36
RMSE	1.35	0.95	0.95	0.95	0.96	1.79
MaxE	4.60	2.17	2.17	2.17	2.52	5.85

Table 1: ME, MAE, RMSE and MaxE for different nowcasting models. Lowest ME, RMSE, MAE and MaxE indicate the models with the lowest mean error, root mean squared error, mean absolute error and max error, respectively. The Combination column contains performance measures for the simple nowcast combination based on the unweighted average of our models. The set of predictors is based on firm-level turnovers.

As we can see from Table 1, the nowcasting performance of our selected models is better than the one of an automated ARIMA procedure. In the first column, we report the results for the model with lowest mean error (an automated ARIMA with principal components extracted from the firm data), which shows a fairly poor performance in terms of MAE, RMSE and max error. Interestingly, the same model (a boosted generalized additive model with factors as input variables) has the best performance in terms of MAE, RMSE and MaxE, however its mean error is fairly high (indicating biased nowcasts). The simple combination of nowcasts shows very similar performance compared to the best possible model in terms of MAE and RMSE, with a slightly higher maximum error. The benefit brought by the nowcasts combination approach is the very low mean error, which means that the combination of nowcasts does not systematically undershoot or overshoot the TIO. Consequently, for the rest of this paper, e.g. when we look at the results for quarterly GDP growth, we focus on the nowcasts obtained by combining different model predictions.

In Table 2 we report the nowcasting performance obtained when using traffic volumes as predictors. We only report the results for the nowcasts combination approach and for the benchmark ARIMA.

	Combination vs. First	ARIMA vs. First
ME	-0.07	0.11
MAE	0.86	1.36
RMSE	1.09	1.79
MaxE	3.16	5.85

Table 2: ME, MAE, RMSE and MaxE for the nowcast combination approach, evaluated using the first version of TIO growth. The set of predictors is based on trucks' traffic volumes.

Table 2 gives us some really interesting insights. With respect to the first version of TIO, the nowcasts combination based on traffic data provides slightly worse predictions, at least compared to the sales' data. However, the MAE and MaxE are fairly low, and much lower than the ones of the automated ARIMA model, indicating a satisfactory nowcasting performance.

We now turn to the results regarding the estimation of quarterly GDP year-on-year growth, in real terms. In particular, we nowcast the t + 60 release of GDP, which is the first official release made by Statistics Finland. Next, we report the nowcasting performance measures for these three sets of predictions. We also compare our results against the performance of the Statistics Finland's flash estimate of GDP. Notice that even in this application, we are using only the vintage of data which would have been available at the time the nowcasts or flash estimates were to be computed.

	Nowcast second month	Nowcast third month	Nowcasts 16 days after	StatFi Flash
ME	0.28	0.08	0.05	0.01
MAE	1.1	1.06	0.84	0.78
RMSE	1.39	1.31	0.99	0.92
MaxE	3.23	2.97	2.13	1.77

Table 3: ME, MAE, RMSE and MaxE for GDO nowcasts, using nowcasts combinations. The set of predictors is based on firms' sales.

Looking at Table 3, we see that our nowcasting framework is able to predict GDP accurately. As we can expect, the performance of the models improves the later we compute the nowcasts and, from the second estimate onward. In particular, the latest estimates presents a comparable performance compared to the Statistics Finland flash estimates, providing a 30 days reduction in publication lag.

Finally, we examine the performance of the nowcasts based on traffic data in Table 4.

	Nowcast second month	Nowcast third month	Nowcasts 16 days after	StatFi Flash
ME	0.21	-0.12	0.04	0.01
MAE	1.04	1.04	0.79	0.78
RMSE	1.31	1.29	0.98	0.92
MaxE	3.21	3.28	2.15	1.77

Table 4: ME, MAE, RMSE and MaxE for GDO nowcasts, using nowcasts combinations. The set of predictors is based on trucks' traffic volumes.

The results of Table 4 confirm that the nowcasts produced using traffic date have a satisfactory predictive performance, very similar to the one based on firm-level sales. Overall, it is interesting to see that traffic data are allowing us to create fairly precise estimates of GDP growth well before the official publication by Statistics Finland. Given the potentially real-time availability of traffic volumes' measurements, these results indicate the need to further explore the nowcasting ability of models based on these data.

5. Conclusions

We have examined the potential of large micro-level datasets, in combination with statis-tical models and machine learning techniques that are able to handle high-dimensional information sets, for the production of faster estimates of real economic activity in-dicators, both at the monthly and at the quarterly frequency. In particular, we have examined the nowcasting performance of firm-level data, and of trucks' traffic volumes measurements.

We find that a simple combination of the nowcasts obtained from a large set of machine learning techniques and large dimensional statistical models is able to produce accurate estimates of monthly real economic activity, or at least estimates that do not lead to a much larger revision error compared to the current official publications. While the revision errors do not increase substantially, our approach allows for a reduction in the publication lag of roughly 30 days, when considering the monthly indicator. Turning to the results related to quarterly GDP, we find that our nowcasts would produce fairly accurate estimates of GDP growth during the third months of the reference quarter, even though there are few large errors. On the other hand, the nowcasts computed at t + 16 are accurate and do not show large revisions, or at least revisions that are compatible with the ones of Statistics Finland. Even though these estimates would be produced after the end of the quarter, they would still allow for more than a month reduction of the publication lag. Finally, it is important to underline the satisfactory performance of traffic measurements data. The potential of this source of information should be explored further, given its real-time availability.

References

1. S. Boragan Aruoba, Francis X. Diebold, and Chiara Scotti. Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics,* 27(4):417–427, 2009.
2. Christine De Mol, Domenico Giannone, and Lucrezia Reichlin. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics,* 146(2):318–328, October 2008.
3. Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics,* 164(1):188–205, September 2011.
4. Martin D. D. Evans. Where Are We Now? Real-Time Estimates of the Macroeconomy. *International Journal of Central Banking,* 1(2), September 2005.
5. Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics,* 55 (4):665–676, May 2008.
6. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition, 2009.
7. James H. Stock and Mark W. Watson. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics,* 20(2):147–62, April 2002.
8. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 58(1):267–288, 1996.
9. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology,* 67(2): 301–320, 2005. ISSN 1369-7412.

Nowcasting and forecasting with Dynamic Factors Models: Some experiences and lessons*

M. Camachoa[1], R. Doménechb[2]
[1] Universidad de Murcia, Spain
[2] BBVA Research and Universidad de Valencia, Spain

Abstract

Dynamic Factor Models (DFM) have become a useful econometric methodology for GDP nowcasting and forecasting. We review the experience in the use of these models at BBVA Research for a large sample of advanced and emerging countries. Particularly when financial variables are included, DFM forecast GDP growth at least as well as other alternative methodologies, in a very parsimonious ways, allowing to present the contribution of each variable to the innovation of forecasts in a very informative way. We also show that DFM can be used to nowcast GDP growth when the estimates provided by national statistical institutes are not useful indicators of the underlying activity. Finally, DFM can be adapted to include valuable indicators of economic activity at different time frequencies obtained using real-time big data information such as retail sales or credit cards spending, improving significantly the nowcasting performance of our models.

Keywords

Dynamic factor models, GDP, nowcasting, forecasting, financial variables, big data. JEL Classification: E32, C22, E27.

1. Introduction

During the Great Recession in 2008 and 2009, policy makers and other economic agents seemed eager to detect signals of its intensity and length, and those of the subsequent the recovery. Some years later, the same type of concerns attracted attention to the crisis of several emerging countries. And more recently, in one of the most longest lasting expansions since the mid-19th century in the US economy, there is a huge interest to anticipate the signals of a potential downturn in future quarters. As for other peers in the world baking industry, the anticipation of economic conditions is one of the most important challenges for BBVA to position competitively and strategically in its footprint.

* M. Camacho thanks the financial support of grants MINECO ECO2016-76178-P and Seneca Excellence Groups 19884/GERM/15. R. Doménech thanks MINECO CICYT ECO2017-84632 and Generalitat Valen-ciana PROMETEO2016-097 for financial support. Contact: r.domenech@bbva.com.

Given the publication lags of GDP, different nowcasting and forecasting methods have been proposed to asses current and future economic conditions, using real-time data of economic variables that are published on a more frequent and timely basis, such as industrial production, financial variables, or consumer and business confidence. To deal with the mixed frequencies, unbalanced panel data and the potential non-linearity of these variables, different methods have been proposed, such as bridge equations, MIxed DAta Sampling (MIDAS) regression models, and Dynamic Factor Models (DFM).[1]

In this paper, we review the experience at BBVA Research in the use of DFM to nowcast and forecast GDP growth in our footprint, which can be summarized in six lessons. First, DFM forecast GDP growth and recession probabilities at least as well as other models in the large sample of countries (world, USA, EMU, China, Spain, Portugal, Turkey, Argentina and Mexico). Second, DFM forecast in a very parsimonious ways, allowing to present easily the contribution of different indicators to forecasts innovations. Third, financial variables, such as the slope of the yield curve or financial tension indexes, contain valuable information about future growth and can be easily introduced in DFM. Fourth, DFM should be tailored to different countries and variables. Fifth, DFM can be used to estimate the underlying activity in countries where official GDP statistics are not reliable. And sixth, DFM allow the introduction of useful indicators of economic activity obtained using real-time big data (e.g., retail sales, credit cards spending, etc.), improving nowcasting significantly.

2. Methodology

2.1 Small-scale dynamic factor models

DFM were advocated by Geweke (1977) as a time-series extension of factor models previously developed for cross-sectional data. The premise of DFM is that the covariation among economic time series variables at leads and lags can be traced to a few underlying unobserved series, usually known as factors. Although dynamic factor models have been the source of a vast literature, in this paper we focus on the "single-index" dynamic factor model developed by Stock and Watson (1989, 1991). Low-dimensional parametric dynamic factor models are expressed in state space form. This implies that the Kalman Filter can be used to construct the Gaussian likelihood function and thereby to estimate the unknown parameters of the model by maximum likelihood.

2.2 The single-index dynamic factor model

Let x_t denote the $n \times 1$ vector of a set of macroeconomic indicators observed at period t. These indicators are assumed to be covariance stationary

[1] See Camacho, Perez-Quiros and Saiz, 2013, and Foroni and Marcellino, 2013.

and typically refer to either growth rates of hard indicators or the level of soft indicators. Without any loss of generality, we will consider a set of only two economic indicators, which implies $x_t = (x_{1t}, x_{2t})'$. The single-index dynamic factor approach attempts to forecasting a targeted time series, y_t, by using its own dynamics and the dynamics of a set of indicators, x_t that capture aggregate economic activity.

Let us assume that the variables used in the model admit a dynamic factor representation. In this case, the j-th indicator of the model can be written as the sum of two stochastic unobserved components: a common component, f_t, which represents the overall business cycle conditions, and an idiosyncratic component, u_t^j, which refers to the particular dynamics of the time series. Both the unobserved index and the idiosyncratic component are modeled as having linear stochastic structures. In particular, the underlying business cycle conditions are assumed to evolve with $AR(p_f)$ dynamics

$$f_t = \rho_1^f f_{t-1} + \cdots + \rho_{p_f}^f f_{t-p_f} + \epsilon_t^f, \tag{1}$$

where $\epsilon_t^f \sim i.i.d.N(0, \sigma_e^2)$.

Apart from constructing an index of the business cycle conditions, we are interested in computing accurate short-term forecasts of y_t. To compute these forecasts, we start by assuming that the evolution this time series depends linearly on f_t and on their idiosyncratic dynamics u_{yt}, which evolves as an $AR(p_y)$,

$$y_t = \beta_y f_t + u_t^y \tag{2}$$

$$u_t^y = \rho_1^y u_{t-1}^y + \cdots + \rho_{p_y}^y u_{t-p_y}^y + \epsilon_t^y. \tag{3}$$

where $\epsilon_t^y \sim i.i.d.N(0, \sigma_y^2)$.

In the same way, the economic indicators also admit a common factor representation, in the sense that they depend contemporaneously on f_t and on their idiosyncratic dynamics, u_t^i. Again, the idiosyncratic components of the n monthly indicators can be expressed in terms of autoregressive processes of p_i orders

$$x_{it} = \beta_i f_t + u_t^i \tag{4}$$

$$u_t^i = \rho_1^i u_{t-1}^i + \cdots + \rho_{p_i}^i u_{t-p_i}^i + \epsilon_t^i, \tag{5}$$

where $\epsilon_t^i \sim i.i.d.N(0, \sigma_i^2)$ and $i = 1,2$.

The main identifying assumption is that the co-movements of the multiple time series arise from the single source of the common factor f_t. This is made precise by assuming that $(\epsilon_t^f, \epsilon_t^y, \epsilon_t^1, \epsilon_t^2)$ are mutually and serially uncorrelated at all leads and lags. In addition, the scale of f_t is identified by setting the normalization restriction that $\sigma_e^2 = 1$.

The single-index dynamic factor model stated above can easily be represented in state-space form and be estimated by maximum likelihood using the Kalman filter. Additionally, as shown by Camacho and Doménech (2019), the model can be easily extended in the following dimensions:

1. Dealing with data problems and variables in levels and growth rates
2. Mixed frequencies. Following the approach of treating quarterly series as monthly series with missing observations, Mariano and Murasawa (2003) extended the single-index dynamic factor model by including quarterly GDP data and monthly indicators.
3. Unbalanced data sets and missing data due to different publication dates, causing the so-called *ragged-edge* data problem at the end of the sample.
4. Business-cycle nonlinearities estimated using Markov-switching dynamic factor models. This allows us to obtain real-time recession probabilities.
5. Leading indicators, such as the slope of the yield curve or financial tension indexes.
6. Measurement errors in GDP growth.

We call these extensions MICA models because they are factor Models of economic and financial Indicators used to monitor the Current development of the economic Activity.

3. Results

Although BBVA's DFM are continuously updated in real time, we restrict this section to the results obtained in the analysis of the models at the time they were published in academic journals (Camacho and Doménech, 2012, Camacho and García-Serrador, 2014, Camacho and Martínez-Martín, 2014, Camacho, dal Bianco, and Martínez-Martín, 2015a).

The implementation of the methodology described in the previous section requires a selection of the appropriate indicators from a list of potential business cycle indicators. We simplify this process by choosing them according to certain properties: timeliness, high statistical correlation and relevance (high loading factor). A meaningful starting point for selecting indicators entering a factor model is the approach of Stock and Watson (1991). Using this baseline model, new indicators are further added when the have statistically significant loading factors. All monthly series are made stationary by differencing or log-differencing, if needed, so they appear in quarterly growth rates (QGR), in monthly growth rates (MGR), in annual growth rates (AGR) or in levels (L). All variables are standardized by subtracting the mean and dividing by the standard deviation.

After the selection process, the set of economic indicators with statistically significant loading factors includes the following variables:

1. Spain: real GDP, real wage income (RWI), electricity consumption (EC), social security affiliation (SSA), registered unemployment (U), real credit card spending deflated with the consumer price index (CCS), consumer confidence (CC), industry confidence (IC), the slope of the yield curve (10-year bond rate minus 3-month Euribor, SLOPE), the average mortgage rate minus the 12-month Euribor (MR12E), the average mortgage rate minus the 12-month Treasury bill rate (MR12TBR) and real credit to the private sector (RCPS).
2. Eurozone: real GDP, unemployment (UR), industrial production (IP), exports (Exp), Economic Sentiment Indicators for the euro area related to industry (ESII), services (ESIS) and consumers (ESIC), total credit to households (LHH) and term spread (beta)
3. US: real GDP, monthly industrial production (IPI), payroll employment (EMPL), real personal income less transfers (RPI), and trade sales (SALES), industrial new orders (MNO), housing starts (HOUSE), the Conference Board consumer confidence index (CC) and the ISM manufacturing PMI, SP500 and the term spread (SLOPE)
4. Argentina: real GDP, industrial production (IPI), quarterly employment (EMPL), real personal income, and real trade sales (SALES) and construction activity (ISAC).
5. World: GDP, industrial production (IPI), PMI, employment (EMPL), exports orders (NExO) and the VIX.

The pseudo real-time forecasting accuracy of the models is examined in Table 1, which shows the mean-squared forecast errors (MSE), as the average of the deviations of the predictions from the final releases of GDP available in the data set. Two alternative models are included in the forecast evaluation: an autoregressive model of order two (AR) and a random walk (RW) model. This table shows that the forecasts from DFM clearly outperform those from univariate models.

Using our DFM, we have also investigated to what extent there has been deviations between the official figures of GDP in Argentina and reliable indicators of economic activity.

Table 1: *Predictive accuracy*

	backchats			Nowcasts			Forecasts		
	All	Rec	Exp	All	Rec	Exp	All	Rec	Exp
Panel 1: MICA-Spain (1Q1990-2Q2009)									
MICA	0.138	0.470	0.069	0.194	0.795	0.070	0.260	1.078	0.090
RW	0.351	1.434	0.131	0.357	1.475	0.133	0.361	1.469	0.135
AR	0.207	0.808	0.085	0.280	1.193	0.095	0.309	1.303	0.108
Panel 2: MICA-Europe (1Q1990-1Q2010)									
MICA	0.138	0.449	0.080	0.298	0.916	0.108	0.403	1.358	0.223
RW	0.370	1.746	0.111	0.361	1.707	0.101	0.377	1.779	0.113
AR	0.297	1.350	0.098	0.369	1.651	0.127	0.377	1.769	0.114
Panel 3: MICA-US (3Q1989-4Q2011)									
MICA	0.218	0.480	0.182	0.368	1.238	0.245	0.443	1.989	0.220
RW	0.404	2.531	0.210	0.500	2.554	0.211	0.504	2.569	0.217
AR	0.358	1.435	0.208	0.431	2.007	0.208	0.491	2.413	0.209
Panel 4: MICA-Argentine (1Q2002-1Q2012)									
MICA	1.049	-	-	1.552	-	-	1.884	-	-
RW	2.999	-	-	2.187	-	-	2.174	-	-
AR	2.851	-	-	2.090	-	-	2.139	-	-
Panel 4: MICA-World (1Q2000-3Q2013)									
MICA	0.12	0.93	0.11	0.19	2.06	0.11	0.24	2.55	0.11
RW	0.31	3.50	0.15	0.32	3.50	0.15	0.33	3.50	0.13
AR	0.21	1.06	0.19	0.24	2.04	0.16	0.25	2.04	0.14

Notes. In each panel the figures show the Mean Squared Errors (MSE) of MICA, Random Walk (RW), autoregressive of order two (AR). All refers to the entire forecasting sample while Rec and Exp are the results of splitting the sample into recessions and expansions according to the NBER (for the US) and ECRI (for Spain and Euro area) business cycle datings.

In particular, Camacho, dal Bianco, and Martínez-Martín (2015b) collect a set of reliable indicators of economic activity, which serve as a proxy of output growth, estimate a single-index DFM that captures the in-sample statistical relationships across indicators and GDP until 2007, and compute out-of-sample GDP forecasts since this year. Our results show an important gap since 2007 between the official GDP and the model-based alternative projections. The official real GDP shows an accumulated positive gap of about 14.4% between 2007 and 2012. To check for the validity of our experiment, we have replicated the analysis for the US. In this case, the DFM produces an estimation of GDP that is almost identical to the official data, both during the recession period (2008-2009) and in the expansion of the following years.

The last extension of the basic single-index dynamic factor model accounts for the evolution of the world GDP across two different regimes: recessions and expansions. The transition between these regimes is governed by a two-

state Markov-switching variable and extended to accounts for leading indicators. Thus, the DFM produces an index of global business cycle condition, yields short-term forecasts of world GDP quarterly growth in real time in a monthly basis, and estimate the real-time probabilities of being in a recessionary regime. Our results are very consistent to the chronology of global recessions proposed by Martinez-Garcia, Grossman and Mack (2015): the probabilities of recession jump quickly around the peaks, remain at high values during recessions and fall to almost zero after the troughs.

4. Discussion and conclusion

Recognizing in advance the evolution of GDP is crucial for economic agents' decisions. In this paper, we have reviewed the experience at BBVA Research in the use of DFM to nowcast and forecast GDP growth in a large sample of advanced and emerging countries. Our results show that DFM forecast GDP growth and recession probabilities at least as well as other alternative models. DFM forecast in a very parsimonious ways, allowing to present easily the contribution different indicators to forecasts innovations of GDP growth. Financial variables (e.g., the slope of the yield curve or financial tension indexes) contain valuable information about future growth and can be easily introduced in DFM. Additionally, DFM should be tailored to different countries and variables, and they can be used to estimate underlying activity in countries where official GDP statistics are not reliable. Finally, DFM allow the introduction of useful indicators of economic activity obtained using real-time big data (e.g., retail sales, credit cards spending, etc.), improving nowcasting and forecasting very significantly.

References

1. Camacho, M., and R. Doménech (2012): "MICA-BBVA: a factor model of economic and financial indicators for short-term GDP forecasting," *SERIEs,* 3, 475–497. https://goo.gl/WGa4df
2. Camacho, M., and R. Doménech (2019): "Nowcasting and Forecasting with DynamicFactors Models: Some Experiences and Lessons," Mimeo. BBVA Research. https://bit.ly/2ZOsotc
3. Camacho, M., G. Perez-Quiros and P. Poncela (2013): "Short-term forecasting for empirical economists: A survey of the recently proposed algorithms," *Foundations and Trends in Econometrics,* 6, 101-161.
4. Camacho, M., and A. García-Serrador (2014): The EURO-STING revisited: the usefulness of financial indicators to obtain Euro area GDP forecasts," *Journal of Forecasting* 33, 186-197.
5. Camacho, M., and J. Martínez-Martín (2014): "Real-time forecasting US GDP from small-scale factor models," *Empirical Economics,* 47, 347-364.

6. Camacho, M., and J. Martínez-Martín (2015): "Monitoring the world business cycle," *Economic Modelling,* 51, 617–625.
7. Camacho, M., M. dal Bianco, and J. Martínez-Martín (2015): "Short-Run Forecasting of Argentine GDP Growth," *Emerging Markets Finance and Trade,* 51, 473-485.
8. Camacho, M., M. dal Bianco, and J. Martínez-Martín (2015): "Toward a more reliable picture of the economic activity: An application to Argentina," *Economics Letters,* 132, 129-132.
9. Foroni, C., and M. Marcellino (2013): "A survey of econometric methods for mixed-frequency data," Norges Bank Working Paper 2013-6.
10. Geweke, J. (1977): "The dynamic factor analysis of economic time series." In: A. Aigner and A. Goldberger (eds.), *Latent Variables in Socio-Economic Models.* Amsterdam: North-Holland.
11. Mariano, R., and Y. Murasawa (2003): "A new coincident index of business cycles based on monthly and quarterly series," *Journal of Applied Econometrics,* 18, 427-443.
12. Martinez-Garcia, E., V. Grossman, and A. Mack (2014): "A contribution to the chronology of turning points in global economic activity (1980–2012)," *Journal of Macroeconomics,* 46: 170-185.
13. Stock, J., and M. Watson (1991): "A probability model of the coincident economic indicators." In K. Lahiri, and G. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records.* Cambridge: Cambridge University Press.

Dual-system estimation by another name: Population size estimation using weighting classes

Owen Abbott
Office for National Statistics, Fareham, UK.

Abstract

This presentation will outline the application of a number of approaches to population size estimation, applied to an administrative based system. The Office for National Statistics (ONS) is exploring models for the future of population statistics. The UK does not have a population register or a set of coherent identifiers across administrative datasets held by government. The current population statistics system is underpinned by the decennial Census, which is expensive and is arguably becoming increasingly unwieldy as a source of data in a rapidly evolving society and with ever increasing demands for more timely, relevant statistics. The ONS is therefore researching how it can transform its population statistics system within that context, and the most important part is the estimation of population size.

The expected sources of data for estimation include administrative data sources and a population coverage survey designed for the purposes of estimating population size, much like a Post-enumeration Survey. Classic capture-recapture estimators can be applied, but they rely on heavy assumptions such as minimal over-coverage, which may not be the case for the administrative datasets. Alternative estimators have been explored, as well as ways of processing the administrative data in advance. Previous research (Abbott *et al*, 2015) applied a weighting-class approach to the measurement of coverage in a Census. This form of estimator can be thought of as a modified dual-system estimation, with a different set of assumptions.

The presentation will discuss the different flavours of dual-system estimation in use across National Statistical Institutes for estimating population size, including the Bayesian approaches developed by Statistics New Zealand, and where they differ in terms of the assumptions that underpin each application. The research into exploring these methods in the UK context will be outlined, with some early results.

Keywords

Administrative data; Official statistics; capture-recpture

1. Introduction

The Office for National Statistics (ONS) is exploring models for the future of population statistics. The UK does not have a population register or a set of coherent identifiers across administrative datasets held by government. The

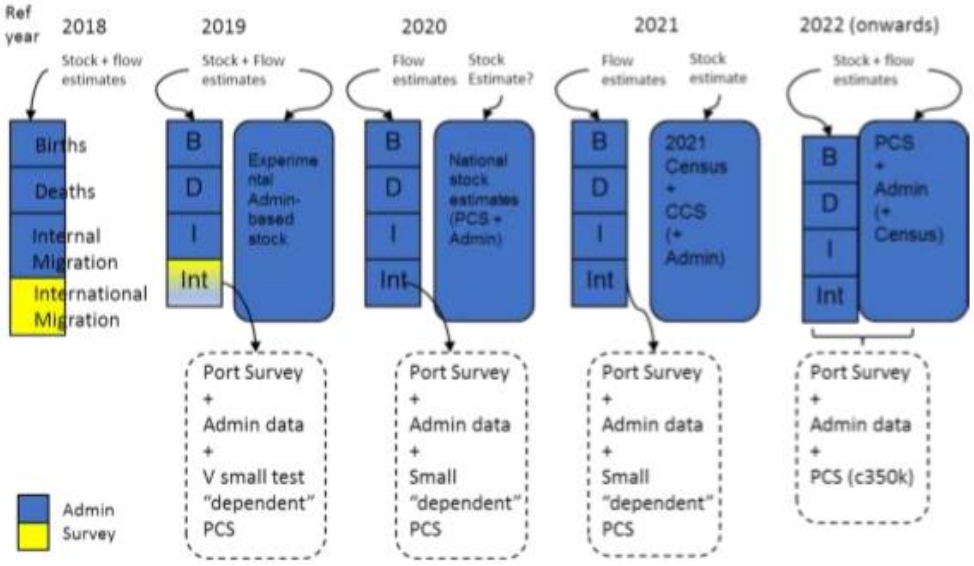
current population statistics system is underpinned by the decennial Census, which is expensive and is arguably becoming increasingly unwieldy as a source of data in a rapidly evolving society and with ever increasing demands for more timely, relevant statistics. The system is also highly reliant on a port-based survey to measure migrant flows to and from UK, with the result that the intercensal population size estimates tend to have an increasing element of bias. The ONS is therefore researching how it can transform its population statistics system within that context (ONS, 2018), and the most important part is the estimation of population size.

This paper will discuss the application of a number of approaches to population size estimation, applied to an administrative based system. These approaches are all based on capture-recapture methods which have been used to measure undercoverage within the UK Census. The assumptions underpinning such methods can be altered to suit different scenarios, for instance when over-coverage becomes more of an issue. However, one does not get something for nothing and whilst some options are less sensitive to over-coverage, they are more sensitive to failures in other assumptions.

Firstly, the context within the UK statistical system is briefly described, followed by a high level view of the current framework for transformation. The presentation will discuss the different flavours of dual-system estimation in use across National Statistical Institutes for estimating population size, including the Bayesian approaches developed by Statistics New Zealand, and where they differ in terms of the assumptions that underpin each application. The research into exploring these methods in the UK context will be outlined, with some early results.

Figure 1 shows the high level of view of transformation plans in the lead up to and immediately after the 2021 Census. The principle is to try to produce the best possible population statistics every year using all available sources, whilst moving away from reliance on large scale data collection activities like the Census. In 2022 and 2023, the quality of the statistical outputs will feed into the decisions around whether a large-scale data collection (not necessarily a census) will be required in 2031.

Figure 1. Proposed Transformation of Population Statistics in England and Wales



This is an ambitious undertaking with many challenges. Research to date has established that population size estimates for local government areas can be derived from administrative sources alone, but the quality varies and is poorer at lower levels of disaggregation. Those quality problems are driven by administrative data lags, resulting in both under-coverage and significant over-coverage. From a methodological perspective, this is the largest challenge. ONS has extensive experience of dealing with under-coverage (in censuses), using a focused coverage survey and capture-recapture techniques (Brown et al, 2018). Dependent interviewing, where a sample of administrative records are drawn and traced in the field, is in theory the best way of estimating over-coverage. However, due to ethical concerns, only dependent sampling (where you can sample but not use data in the field) is being considered. Thus, ONS is currently exploring the options for estimating population size using some form of capture-recapture.

The expected sources of data for estimation include administrative data sources and a population coverage survey designed for the purposes of estimating population size, much like a Post-enumeration Survey.

2. Methodology

2.1 Traditional Dual-system Estimation

Dual-system estimation (DSE) is an established and understood technique when applied to measuring population size from Censuses. In the UK it has been used successfully in the 2001 and 2011 Censuses to derive the key population estimates. It will also be used, albeit in a log-linear modelling form,

for the 2021 Census. Brown *et al* (2018) provide a description of DSE and its underpinning assumptions, which will not be reproduced here.

In the context of using administrative data instead of a Census to derive population size estimates, there will still be a requirement for a second 'count' of the population which is used for estimating coverage (and hence size). In the UK, there are plans to develop a Population Coverage Survey (PCS) which will play that role, much like the equivalent Census Coverage Survey (CCS). The PCS will borrow many sample design elements from the CCS – it will be a short survey focusing on counting households and people, around 300,000 households per annum.

Using the PCS within a DSE framework seems like the obvious answer, but it is susceptible to bias due to over-coverage in either source, and as noted previously many of the administrative lists in the UK suffer from this. In addition this would also require high quality linkage between the administrative data and the survey, which is achievable but costly.

2.2 Weighting class

An alternative to DSE is a weighting-class approach (see Lohr, 1999). The basic idea is to use information known about household responders and non-responders to estimate response rates within classes. In the context of deriving population size from administrative data, the known information are the administrative data entries linked to the address frame being used by ONS for all of its statistical processes. The PCS sampled addresses are linked to the frame so that responding and non-responding addresses are identified.

The key attractions of this approach (compared to DSE) are that high quality individual level linkage is no longer required as the only requirement is linkage to the address frame, and the estimator is less susceptible to over-coverage in the source used to estimate the weights (in this case the administrative data). The reduction in the matching requirement is extremely attractive, as it is a costly and time intensive undertaking to ensure that matching is completed to a high standard.

However, the approach does have some drawbacks. It makes no allowance for persons missed in addresses captured by the survey, which is generally about 2 to 3 per cent in previous census coverage surveys. In addition, it can also be biased in the same way as DSE as it also assumes homogeneity of response probability within classes.

As part of the initial work to explore estimation options for using administrative data in place of a census for population size estimation, ONS (2013) reported on the work undertaken to apply such an approach. This assumed a population coverage survey as described previously was designed and available. Such a survey was simulated from the 2011 Census data, along with administrative data. The weighting class approach performed as expected, being much less sensitive to over-coverage but biased due to survey

within-household non-response. The variance of the estimator was also higher, reflecting its simpler form.

Abbott *et al* (2015) explored the weighting class approach further in the context of a census, showing some of the empirical properties in this context when compared to a standard DSE approach.

ONS (2017) reports on additional work exploring weighting classes when some activity data provides an improved administrative data base. However, it was not possible to separate out the biases due to failures in some of the other assumptions underpinning the method. This made it very difficult to assess whether the approach was 'better' than a standard DSE.

In summary, this work highlighted that the weighting-class approach is analogous to a DSE where the assumption of zero within-household non-response means that the estimator is no longer as susceptible to individual level over-coverage on the first source. It is still susceptible to over-coverage on the frame and over-coverage on the survey, but both of these are deemed less of a risk. Whilst it is an attractive approach due to its over-coverage properties, like DSE, it would require a second estimation process to adjust for biases due to failure of assumptions. In this case it would be for within-household non-response.

2.3 Other methods

There are other flavours of dual-system estimation being explored for the purposes of estimating population size using administrative data. Again, these approaches essentially trade off the underlying assumptions to attempt to reduce bias, choices being made based on the context. Here we describe the work being undertaken by two National Statistical Institutes in this area, although these are not exclusive.

2.4 Trimmed DSE

The Central Statistics Office Ireland is exploring a trimmed DSE approach (Dunne, 2018). The idea behind this approach is heavily based on having good quality activity type information (either directly from a source or via linkage across multiple sources) to deal with over-coverage. The activity information is used to trim out records from the administrative data, calculating a new DSE along the way. If the records being trimmed are genuine over-count, the DSE will decrease (as the records will be in the 'admin only' cell of the DSE and thus be inflating the estimate). Trimming continues until the DSE stabilises and/or the variance grows too much, and at that point the DSE is taken as the best estimate.

This approach makes a lot of sense – it is turning an administrative source from one with over-coverage (and undercoverage) into a source with only under-coverage. The DSE methodology works much better under those conditions, so it is a clever way of getting one assumption to fit much better and thus reduce bias. However, it does perhaps make the estimator more

susceptible to failures in other assumptions (e.g. homogeneity bias) as in some extreme cases the under-coverage rate can be quite large. Work is ongoing to explore this further in Ireland and other countries. ONS have done some early work on such ideas for both Census and Administrative data contexts (ONS, 2014).

2.5 Bayesian Approach

Statistics New Zealand (SNZ) have taken a Bayesian approach to this as described by Bryant and Graham (2015). They are developing a Bayesian hierarchical model, where the observed data comes from either administrative data, or from a coverage survey where the two are linked at individual level.

This model seeks to infer from the observed data people's "true" location in a population-administrative data union, allowing for both under and over-coverage. The key input is the use of prior information that provides reasonable bounds on the total population size N (i.e. it's likely to be similar to previous size estimates), and to the coverage probabilities of the administrative data (e.g. we might expect that some age groups are more likely to interact with admin data than others). ONS are currently experimenting with this approach (jointly with SNZ) to see how strong the prior information needs to be in order for the estimates to be unbiased, and how robust the approach is to failure in other assumptions.

3. Results

To highlight the different performance of the versions of DSE, specifically the classic DSE and the weighting class estimator, ONS (2017) showed the results of applying these methods to a linked administrative data based 'statistical population dataset'. Figure 11 (reproduced from ONS (2017)) shows the results for males by five-year age group when compared to the gold-standard 2011 Census population estimates. It shows that the DSE suffers from failures in the over-coverage assumption, and the Weighting class estimate (WCE) is negatively biased due to within household non-response.

Figure 11: Percentage difference from census estimates for males by five-year age group
England and Wales, 2011

Source: Office for National Statistics

Further work is being undertaken to refresh the simulation framework from the earlier studies to enable a fuller evaluation of the performance of all of these versions of capture-recapture estimators.

4. Discussion and Conclusion

Capture-recapture methods depend heavily on the underlying assumptions being met, and when they are not estimates based upon them are biased. This is clearly not desirable when using such techniques for official statistics about the size of the population. This paper has outlined various ways that National Statistical Institutes are attempting to tackle this issue in the context of using administrative data to drive such estimates. Essentially, it is an exploration of ways in which the assumptions can be balanced in different ways according to the likely qualities of the data.

The methodological challenges of transforming population statistics should not be underestimated. ONS is devoting significant resources to exploring what might be possible, and this paper only touches the surface of those that need and are being addressed. For instance, other methodological challenges not covered here include data linkage, disclosure control, uncertainty measurement and the output definitions. Research is into these is continuing apace.

References

1. Abbott, O., Castaldo, A., Racinskij, V., Ross, H., Smith, P. A. and Brown, J. J. (2015) Developing a weighting-class approach for the 2021 Census. Accessed at: www.researchgate.net/publication/305710321_Developing_a_weighting-class approach for the 2021 Census

2. Brown, J. J., Sexton, C. Abbott, O. and Smith, P.A.(2018) The framework for estimating coverage in the 2011 Census of England and Wales: Combining dual-system estimation with ratio estimation. Statistical Journal of the IAOS vol. Pre-press, no. Pre-press, pp. 1-19, 2018.
3. Dunne, J. (2018) The Irish PECADO project: Population estimates compiled from administrative data only, International context and methodology. Accessed at: www.cso.ie/en/media/csoie/newsevents/documents/administrativedataseminar/7thadministrativedataseminar/Methodology Report.pdf
4. J. R. Bryant and P. Graham (2015) A Bayesian Approach to Population Estimation with Administrative Data, Journal of Official Statistics, vol. 31(3), page 475-page 487.
5. Lohr, S. L. (1999) Sampling: Design and Analysis. Pacific Grove: Duxbury Press.
6. Office for National Statistics (2013) Beyond 2011: Producing Population estimates Using Administrative Data: In Theory. Accessed at: www.webarchive.nationalarchives.gov.uk/20150109083021/http:/www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011--producing-population-estimates-using-administrative-data--in-theory--m8-.pdf
7. Office for National Statistics (2014) Beyond 2011: Statistical Research Update (M13). Accessed at: www.webarchive.nationalarchives.gov.uk/20140730091826/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011--statistical-research-update.pdf
8. Office for National Statistics (2017) Research Outputs: Coverage-adjusted administrative data population estimates for England and Wales, 2011. Accessed at: www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/researchoutputscoverageadjustedadministrativedatapopulationestimatesforenglandandwales2011
9. Office for National Statistics (2018), Annual assessment of ONS's progress on the Administrative Data Census. Accessed at: www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogressontheadministrativedatacensusjuly_2018

New population estimation methods: New Zealand and Ireland

John Dunne[1]*, Patrick Graham[2]
[1] CSO, Ireland
[2] Stats NZ, New Zealand

Abstract

Like many countries, New Zealand and Ireland are researching new systems of population estimates compiled using administrative data. Neither Ireland or New Zealand have a Central Population Register from which the estimates can be compiled. Projects in both countries involve first creating a Statistical Population Dataset (SPD) to represent the population and then adjusting for coverage errors.

In Ireland, the SPD is compiled such that by design it will only have undercoverage. The SPD counts are then adjusted using Dual System Estimation (DSE). An extension of DSE, Trimmed Dual System Estimation (TDSE), is then used to verify that no overcoverage exist, or if overcoverage in the SPD is present then those parts of the SPD identified with overcoverage are removed prior to an application of traditional DSE methods to compile population estimates. While the use of TDSE in official statistics is new, this system is also novel in that it uses an administrative data list as list B in the application of DSE methods.

In New Zealand, the SPD is compiled to represent the population as close as possible and as such is expected to contain both undercoverage and overcoverage. The New Zealand approach is novel in that it uses Bayesian methods and a single sample from the target population to correct for both undercoverage and overcoverage at the same time.

This short paper will describe both the Irish and New Zealand approaches as well as discuss the strengths and weaknesses of each approach.

Keywords

Bayesian, TDSE, overcoverage, coverage, SPD

1. Introduction

Statistical agencies in many countries are investigating methods for replacing traditional census based population estimation systems. Ireland and New Zealand are two such countries. Neither country has a Central Population Register (CPR) but both countries have invested significant resources into the exploitation of administrative data sources for statistical purposes (Dunne,

* Corresponding Author: John Dunne, Central Statistics Office, Ireland; E-mail: John.Dunne@cso.ie.

2015; Statistics New Zealand, 2017). As part of this effort both countries are exploring and evaluating new methods for the compilation of population estimates. While undertaken separately, there are similarities in the respective approaches. Projects in both countries involve first creating a Statistical Population Dataset (SPD) to represent the population and then adjusting for coverage errors.

In the Irish PECADO project (Population Estimates Compiled from Administrative Data Only), the SPD is built using a Signs of Life (SoL) approach and as such by design does not suffer from overcoverage. Ireland does have an official person identification number (PIN) used across Public Services, and linkage based on this number eliminates linkage error. DSE methods are used to adjust for undercoverage using a separate administrative data list. Reassurance with respect to the non existence of overcoverage is further provided through an application of an extension of DSE methods called Trimmed Dual System Estimation (TDSE). The TDSE approach was first proposed by Zhang and Dunne (2018) and we will refer to this method as the Zhang and Dunne method.

Statistics New Zealand build their SPD using a rules based approach to determine if someone belongs to the population or not. The rules are typically based on activity on administrative sources in the two year period up to and including the reference period. In this context the SPD is expected to contain both undercoverage and overcoverage errors. Statistics New Zealand have developed and are evaluating a methodology to adjust SPD counts to obtain population estimates. Their approach is based on a Bayesian model and requires a second sample to be undertaken in the field. The approach is proposed by Graham and Lin (2019) and we will call the Graham and Lin method.

As different conditions and assumptions underpin each approach the methods are not directly comparable. However, we consider their different strengths and weaknesses. This short paper has two more sections. The next section describes the two methods and in the final section we consider their strengths and weaknesses.

2. Methods Illustrated

2.1 *The Irish PECADO Project - Zhang and Dunne Method*

We take as our starting point the DSE model proposed by Zhang and Dunne (2018), and following this approach we develop a DSE estimator for the population size as $\widehat{N} = nx/m$ where n is the size of list B, x is the size of list A and m is the size of the match between the two lists. The assumptions underpinning this DSE model are i) no erroneous records in either list A or list B; ii) no linkage error when matching records between list A and list B and iii)

Every unit i in the population U has an equal chance π of being captured in list B.

Adding an additional assumption, the event that a person is *captured* in list *B* is independent of any other person being captured in that list, Zhang and Dunne (2018) show a variance estimator similar to that derived by Sekar and Deming (1949) and presented in Bishop et al. (1975, page 233).

These assumptions are more relaxed than those presented by Wolter (1986). This DSE model can be applied in many more scenarios where the Wolter assumptions may not hold true. One scenario is where list *A* is derived from administrative data sources.

The Irish PECADO project proposes a system where the SPD, compiled from the activity records in individual public administration systems, is list A (size x) with heterogeneity in the capture rates and a second administrative list as list B (size n) satisfying the homogeneous capture assumption. List *B (DLD*) is composed of those persons applying for or renewing their driver licence in a given year. In Ireland, drivers have to renew their licence at least every 10 years and are required to show that they are resident in the State. We assume neither list has erroneous records and we also assume perfect linkage based on official Identification Numbers. Erroneous records can be considered as a record that is not related to a person that should be included in the population. The population estimate, $\widehat{N}$ is compiled as $\widehat{N} = xn/m$ where m is the size of the match between list A and list B. Post stratification by single year of age, gender and nationality group is also implemented to strengthen the homogeneous capture assumption and provide population estimates by these groups. An additional assumption of no undercoverage for those under 18 years of age in list A is also made as *DLD* has no coverage in this age group. *DLD* is further validated as a suitable list *B* by swapping in a smaller list derived from a survey (underpinned by homogeneous capture assumption) and comparing results. TDSE methods are used to hunt for erroneous records.

The theory underpinning TDSE is based on the concept that if the assumption of homogeneous capture holds, then when list *A* is trimmed of k records to get a new (trimmed) list A_T of size $x - k$, there should be no significant difference between the untrimmed population estimate *N* and the population estimate after trimming N_T. The size of the match between list A_T and list *B* is $m - k_1$ where k_1 is the number of records from the trimmed segment that now need to removed from the match between list A and list B. This provides the trimmed population estimate, $\widehat{N}_T = \frac{n(x-k)}{(m-k_1)}$.

We use TDSE methods to evaluate suspect parts of list A for records that are not part of the population. While in theory the SPD is designed to remove them, in practice, there may be errors in processing of administrative data sources that may result in erroneous records being included in the SPD. To do this we identify parts of the SPD where we suspect there may be erroneous

records. Typically, this is done by recompiling the SPD without one of the underlying data sources to get our trimmed list A_T, then comparing , $\widehat{N}_T$ with N to see if they are the same. If , $\widehat{N}_T$ is less than $\widehat{N}$ then this indicates that capture rate for the trimmed segment (size k) is less than the capture rate for list A_T indicating that there is a higher proportion of erroneous records in list A compared to list A_T. Therefore we consider estimator , $\widehat{N}_T$ to be less biased than $\widehat{N}$. We can apply this idea iteratively to each data source in a strategy to eliminate erroneous records from the SPD and obtain a less biased and possibly an unbiased estimate of the population.

An alternative trimming strategy could involve scoring records using criteria correlated with the likelihood that those records are erroneous and then incrementally trimming based on these scores to identify some point where the trimming is no longer effective in removing bias (population estimates are not changing significantly). The variance of the estimator should also be monitored through the trimming. The variance of the trimmed estimator is estimated in the same manner as that for the DSE estimator but using list sizes and matches after trimming.

Alternative trimming strategies can be deployed. Effective trimming strategies are those where trimming steps remove erroneous records without removing too many valid records. The more valid records that are removed, the smaller the match and hence the greater the variance of the estimator. There is a cost to trimming and poor trimming strategies can lead to unstable and possibly biased estimators with poor precision (large variance).

2.2. *Statistics New Zealand - Graham and Lin Method*

Graham and Lin (2019) provide a comprehensive and detailed account of the methodology. Here we present a greatly simplified account of the approach.

	In SPD	Not in SPD	
In Population	$N_{11}(\phi_{11})$	$N_{10}(\phi_{10})$	N_T
Not in Population	$N_{01}(\phi_{01})$	0(0)	
	N_L		

Table 1. Relationship between SPD and target population using the Graham Lin method to estimate the target population size N_T . Multinomial Probabilities (adding to 1) denoted in parenthesis.

	In SPD	Not in SPD
In Sample	$\pi\phi_{11}$	$\pi\phi_{10}$
Not in Sample	$(1-\pi)\phi_{11}+\phi_{01}$	$(1-\pi)\phi_{10}$

Table 2. Probability distribution for breakdown of SPD and Sample. Graham and Lin Method.

We start by considering an SPD compiled from administrative sources. The SPD is compiled to the best of our abilities but is suspected of suffering from undercoverage as well as overcoverage. We now consider all relevant units (persons) U as including persons in both the population and the SPD (U_V of size N_{11}), persons in the SPD but not in the population (U_O, equating to overcoverage in SPD of size N_{01}) and the number of persons in the population and not in the SPD (U_U or undercoverage of size N_{10}). N_{11}, N_{01} and N_{10} are unobserved but $N_L = N_{11} + N_{01}$, the size of the SPD is observed. The objective is to estimate $N_T = N_{11} + N_{10}$, the size of the target population.

First we consider every unit i in the universe U as a multinomial trial with probabilities $P(i \epsilon U_V) = \phi_{11}, \mathrm{P}(\mathrm{i} \epsilon U_O) = \phi_{01}$ and $\mathrm{P}(\mathrm{i} \epsilon U_u) = \phi_{10}$ with $\phi_{11} + \phi_{01} + \phi_{10} = 1$. Table 1 illustrates this relationship between the target population and the SPD.

To estimate N_T the size of the target population, Graham and Lin (2019) propose sampling the target population with known sample inclusion probabilities and linking the sampled units to the administrative list in an error free way. In practice, an area frame in conjunction with a well maintained dwelling register will allow for sampling dwellings with a known inclusion probability. Known inclusion probabilities for individuals then requires an assumption of no within dwelling non-response. Various field procedures can be used to approximate this assumption as closely as possible. However, to simplify notation and explanation, we consider a simple random sample of individuals with a constant and known inclusion probability π. Table 2 provides the corresponding cell probabilities, for the relationship between the SPD and the sample in terms of π which is assumed known and the multinomial probabilities in Table 1. In practice, the underlying probability model (Table 2) is extended to include covariates such as age, sex, ethnicity and geography. We will use $n_{00}, n_{10}, n_{01} and\ n_{00}$ to denote the cell counts in the cross-tablulation of sample and list inclusion (i.e the table of counts corresponding to Table 2), where $n_{00}, n_{10}\ and\ n_{01}$ are directly observed. We note the count for observed (0,1) cell in the sample - list union, n01 contains a mix of people in the target population but not included the sample and people genuinely not in the target population. Consequently the inference is not a standard DSE problem, which deals only with undercoverage in the observed data.

Graham and Lin (2019) take a Bayesian approach to inference which follows from the joint posterior distribution for $\phi^{under} = \phi_{10}/(\phi_{11} + \phi_{10})$ and ϕ_{01}.The posterior distribution for the remaining cell probabilities can be easily obtained using $\phi_{11} = (1 - \phi_{01})\left(1 - \phi^{under}\right), \phi_{10} = (1 - \phi_{10})\phi^{under}$. Given the posterior distribution for the cell probabilities, the posterior distribution for the total target population size can be obtained. Graham and Lin (2019) evaluate two methods for completing the target population unit record file. The first uses the estimated model probabilities and estimated $N_T$to impute

target population records missed by both the list and the sample into the target population and to impute an overcoverage indicator for the list records in the observed (0, 1) cell. Records imputed as overcoverage are excluded from population estimates. The second approach weights the list records by the ratio $w = \frac{\widehat{\phi}_{11}+\widehat{\phi}_{10}}{\widehat{\phi}_{11}+\widehat{\phi}_{01}}$ which is the ratio of one minus the over-coverage probability to one-minus the undercoverage probability. In practice, the weights are specific to particular covariate combinations. Using either the imputation or weighting approach sub-group population estimates can be readily obtained, either by counting records or summing weights within the sub-groups of interest. To represent uncertainty these calculations are repeated for each draw from sample from the posterior of the cell probabilities.

The key assumptions underpinning this approach are:

- the sample is drawn from the target population with known inclusion probabilities for each person (in reality the sample inclusion probabilities may be estimated)
- selection in the sample is conditionally independent of inclusion on the list, given the covariates included in the model
- linkage between the sample and the SPD is done in an error free way
- there is no misclassification with respect to covariate information

The key innovation in this approach is that it does not require sampling directly from the SPD to estimate overcoverage, an idea first proposed by Zhang (2015).

Graham and Lin (2019) acknowledge there is still work to be undertaken with this approach particularly in areas related to clustering in sample design, record linkage error, misclassification in data and outline a number of directions that this work could take.

3. Discussion and Conclusion

The situation in Ireland and New Zealand are similar. Both countries compile an SPD that may contain overcoverage and undercoverage. In the case of Ireland, the approach is to compile an SPD that only has undercoverage and then focus on statistical methods to adjust for undercoverage which may be sizable. In New Zealand, the approach is different in that the SPD attempts to be as close as possible to the target population and the statistical methods are then required to only make minor adjustments for undercoverage and overcoverage. However, both methods can be deployed in scenarios where both undercoverage and overcoverage exist.

Both methods require a second data source to adjust for coverage error in the SPD. Henceforth, we refer to this second list as the coverage list. In both cases, the coverage list must come from the target population with no

erroneous records (overcoverage), the coverage list must be capable of being linked in an error free way to the SPD and information with respect to the inclusion probabilities of units of the population in the coverage list is required. The Zhang and Dunne method operates under the assumption of *homogeneous capture* and blocking can be deployed where there is a suspicion or knowledge that inclusion probabilities will differ between different population groupings. The coverage list requirements under the Zhang and Dunne method allow for the use of a suitable administrative list provided the conditions are satisfied and negate the need to undertake a survey in the field. The Graham and Lin method requires knowledge of the actual inclusion probabilities and where they are not known then high quality estimates are required. The Graham and Lin method requires a greater knowledge of inclusion probabilities. The coverage list requirements in the Graham and Lin method can probably only be satisfied by conducting an appropriate field survey that will satisfy requirements with respect to knowledge of the inclusion probabilities. This is probably the critical difference in the application of the two methods.

To date, the Zhang and Dunne method has been used to compile population estimates as part of the PECADO project for years 2011 to 2016. The PECADO estimates are considered research in nature. When comparing the estimates with that of the 2016 Census of Population estimates there are some differences, part of which can be explained by different population concepts. However, more work is required to reconcile the differences between the PECADO estimates and the Census of Population counts. So far, the Graham and Lin method has been evaluated only on simulated data and shows good statistical properties.

The Graham and Lin method will eliminate overcoverage errors from the estimated target population. The Zhang and Dunne method depends on deploying an effective trimming strategy to eliminate overcoverage errors if they exist and the only guarantee with respect to completely eliminating overcoverage is the belief that the trimming strategy has effectively evaluated all possible sources of overcoverage.

In conclusion, the authors believe it is possible to compile population estimates from administrative data sources without the requirement of a public administration systems underpinned by a CPR. The work undertaken to date shows this possibility, however, more work is required in developing the respective methods.

References

1. Bishop, Y., Feinberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis.* Springer.

2. Dunne, J. (2015). The Irish Statistical System and the emerging Census opportunity. *Statistical Journal of the IAOS,* 31(3):391–400.
3. Graham, P. and Lin, A. (2019). Bayesian and approximate Bayesian methods for small domain population estimation from an administrative list subject to under and over coverage. *Submitted ISI 2019 paper.*
4. Sekar, C. and Deming, W. E. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical AssociationAmeri-can Statistical Association,* 44(245):101– 115.
5. Statistics New Zealand (2017). Experimental population estimates from linked administrative data 2017 release. Technical report, Statistics New Zealand.
6. Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association,* 81(394):338–346.
7. Zhang, L. and Dunne, J. (2018). Trimmed dual system estimation. In Bohning, D., van der Heijden, P. G., and Bunge, J., editors, *Capture-recapture methods for the Social and Medical Sciences,* chapter Trimmed du, pages 237–258. CRC press.
8. Zhang, L. C. (2015). On modelling register coverage errors. *Journal of Official Statistics,* 31(3):381–396.

Multiple system estimation for the size of the Māori population in New Zealand

Maarten Cruyff[1], Peter G.M. van der Heijden[1,2], Paul A. Smith[2], Christine Bycroft[3], Patrick Graham[3]

[1] Utrecht University, the Netherlands

[2] University of Southampton, UK

[3] Statistics New Zealand

Abstract

We investigate the situation where two or more registers, or lists, of individuals are linked both for the purpose of population size estimation and to investigate the relationship between variables appearing on all or only some of the registers. There is usually no full picture of this relationship because there are individuals that are in only some of the lists, and also individuals that are in none of the lists. These two problems have been solved simultaneously in dual system estimation using the EM algorithm. We extend this approach to four registers (including the population census) to estimate the size of the indigenous Māori population in New Zealand, where the reporting of Māori is not the same in each register and where there is a further missing data problem, with individuals included in one or more registers who did not provide their ethnicity. We consider the implications for estimating the size of the Māori population from administrative data only.

Keywords

dual system estimation, linkage, missing data, register, coverage

1. Introduction

The use of dual system estimation (DSE, also known as capture-recapture or the Lincoln-Peterson estimator) to estimate the size of a population which cannot be completely observed has become widespread in official statistics, particularly as a key part of making estimates from a population census (eg Brown et al. 1999, 2019), though also in situations involving the use of linked administrative data sources. The need to make efficient use of data already available to government in the construction of official statistics outputs has led to better access to administrative data, and linkage of the records from these sources is being widely used to understand and estimate corrections for the under- and over-coverage within them. We will use "registers" as a generic term for all sources containing lists of identifiable units.

When two registers are linked, in general there will be some records from one register which remain unlinked, because there is no corresponding record in the other source. This leads to missing data for any variables which appear

in only one register (item missingness). The linked data are used to estimate the size of the population that is not present in either register, and for these unobserved records all the variables are missing (unit missingness). There is an extensive line of research that studies this problem from a missing data perspective, starting with Zwane and van der Heijden (2007), and summarized in van der Heijden et al. (2018). The latter paper concluded that further practical experience with these methods is needed to demonstrate their usefulness in a variety of situations and encourage their wider application.

Here we consider the methods for estimating the size of the Māori population in New Zealand. Ethnicity is the principal measure of cultural identity in New Zealand, and is used across the official statistics system. The 2005 New Zealand statistical standard for ethnicity states that ethnicity is self-perceived and a person can belong to more than one ethnic group. Identifying the indigenous Māori population is of particular importance.

Ethnicity is regularly included in data collections because of its importance in defining groups of policy interest, for example on health outcomes for indigenous people in New Zealand. However, differences in questions, differences in self-perception depending on the context, and changes over time, can all affect how ethnicity is recorded in these data sources (Statistics New Zealand 2005). Ethnicity is collected independently in a number of administrative sources as well as through the census and household surveys. People do not always report the same ethnicity in each source. Also, people do not always report their ethnicity, so there is an additional missingness problem to deal with.

Official population estimates and projections for major ethnic groups in New Zealand are based principally on the responses people provide in the five yearly census, adjusted for non-response using a post-enumeration survey. As part of its census transformation programme, Statistics NZ is exploring the feasibility of a census based on administrative data (Statistics New Zealand, 2012, 2014). The ability to produce ethnicity data from administrative sources is a key consideration. Using ethnicity information from linked administrative data sources may also improve the current production of official ethnic population estimates.

The aim is to use ethnicity information from linked administrative data to improve official ethnic population estimates in New Zealand. In support of this we analyse a variety of census and administrative sources using the approach of Zwane and van der Heijden (2007), with a specific focus on the estimation of the size of the Māori population at the time of the 2013 population census. The analysis requires the extension of the methods to deal with multiple registers and with a variety of different types of missing data. The methodology falls within the area of data integration of multi-source statistics, see de Waal et al. (2017) and Zhang and Chambers (2018).

Four data sources, the population census and three administrative registers are available, that each have an ethnicity variable. Here we focus on Māori ethnicity in a summarised binary form so that we have two mutually exclusive categories: Māori (with or without other ethnicities) and non-Māori (everyone else). Details of these sources and the procedures which have been used to link them are described in section 2; perfect linkage is an essential assumption for DSE. Then we build up the estimation problem in section 3, starting with two registers, and then four registers, and finally consider using the three administrative sources without the census. Some conclusions are presented in Section 4.

2. Methodolgy

Because a person's reported ethnicity can change over time, and depending on the context, a key question is how to combine ethnicity from multiple sources, when information is sometimes conflicting. Reid, Bycroft, and Gleisner (2016) compared ethnicity data from the 2013 Census with the ethnicity information collected by administrative sources, for a New Zealand resident population derived from administrative sources. They found that nearly everyone in this admin-based New Zealand resident population had ethnicity recorded in at least one administrative data source, but that consistency with census responses varied considerably by source and by ethnic group. The method used to combine these sources has a major impact on the result. Under the assumption that census responses provide the best measure for official statistics purposes, a method that ranks sources based on their consistency with the census has been applied. Using administrative data alone was found to produce a time series that reflects expected patterns of increasing ethnic diversity, with age structure and regional distribution of ethnicity consistently in line with official measures (Stats NZ, 2018). The approach however has some limitations, for example it does not allow for reporting errors or conflicts in higher-ranked sources, which may be better managed through a statistical model.

The population used here is the experimental administrative-based NZ resident population known as the 'IDI-ERP' (Stats NZ, 2017). The data are probabilistically linked in Stats NZ's Integrated Data Infrastructure (IDI). The IDI provides safe access to de-identified linked microdata for research and statistics in the public interest.

We use ethnicity data from the 2013 population census and from three administrative sources:

(i) Department of Internal Affairs (DIA) birth registrations data - which includes the ethnicity of the child as reported at registration (ii) Ministry of Education (MOE) tertiary education enrolment data - which includes ethnicity for students (iii) Ministry of Health (MOH) National Health Index

system, a unified national person list - which includes ethnicity. For a more detailed explanation of these sources, see Reid et al. (2016).

Each of the administrative sources relates to different parts of the population. Birth registrations are for babies born in NZ since 1998, or those up to age 14 in 2013; tertiary education enrolments are available from around the late 1990s, and are mainly for those aged between 18 and 40 years in 2013; both census and health data include all ages, and each has an ethnicity value for around 90 % of the IDI-ERP population. Overall, almost 99 percent of the IDI-ERP population have ethnicity information from at least one of these sources, and many people have information from more than one source.

The aim of the following analyses is to produce aggregate estimates of Māori and non-Māori ethnicity by combining these four independent sources: the 2013 Census and the three administrative sources.

3. Results

3.1 Two registers

We first explain the methodology for two registers and then apply it to four registers. We start by using the two sources with the widest coverage, the Census and the MOH. Being in the Census is denoted by A (A = 1 for 'yes', A = 0 for 'no'), and similarly for MOH, denoted by C. The ethnicity variable in the Census is denoted by a (a = 0 for non-Māori, a = 1 for Māori, a = '-' for individuals who are in A but did not fill in their ethnicity, and a = 'x' for individuals that are not in A). The ethnicity variable in the MOH is denoted by c and coded similarly to a. In comparison to the methods employed by van der Heijden et al. (2018) , the presence of the '-' level in variables a and c is new, and we first extend these methods with two registers.

Figure 1 illustrates the form of the data when they are coded in a matrix of individuals in the rows by variables in the columns. The middle two columns depict A and C, that indicate whether individuals are only in A but not in C ((A; C) = (1; 0)), in both A and C ((A; C) = (1; 1)) or not in A but only in C ((A; C) = (0; 1)). At the bottom is a horizontal band of 'Individuals missed by both lists', and this refers to (A;C) = (0; 0). This last number has to be estimated to arrive at an estimate of the size of the total population of non-Māori and Māori. The first column stands for ethnicity variable a. When individuals are only in A ((A; C) = (1; 0)), there are three types of individuals, namely 0, non-Māori (light grey); 1, Māori (blocks); and '-', those who have a missing value for ethnicity (raster). When individuals are in both A and C ((A; C) = (1; 1)), all three areas are found. When individuals are not in A but only in C, the ethnicity variable a is automatically not measured and denoted by 'x' (white area). The last column stands for ethnicity variable c, and it has similar levels as a. Notice that there are three kinds of missing data: there is item missingness '-' for those individuals that are on a list but did not provide their ethnicity; there is item

missingness 'x' for those individuals that are not on one list, and hence have no value on the corresponding ethnicity variable (if only A = 0, a = 'x', and if only C = 0, c = 'x'). Last, there is unit missingness for those individuals that are missed by both A and C.

A second presentation of the problem is in contingency table format, see Table 1, Panel 1. The original 15 counts in Table 1, Panel 1, will have to be redistributed over 3 subtables of dimension 2×2. I.e., the subtable of size 3×3 has to be reduced to size 2×2, the three values for A = 0; a = 'x' have to lead to a subtable of size 2×2 and similarly for the three values for C = 0; c = 'x'. In a second step the subtable for A = 0; C = 0 has to be estimated, and this refers to the individuals that are missed by both lists. Thus two types of missing data are estimated. Estimates are found using the Expectation- Maximization algorithm. Van der Heijden et al. (2018) show that the maximal loglinear model that can be fitted to the data is [Ac][ac][Ca], where the highest fitted margins are placed between square

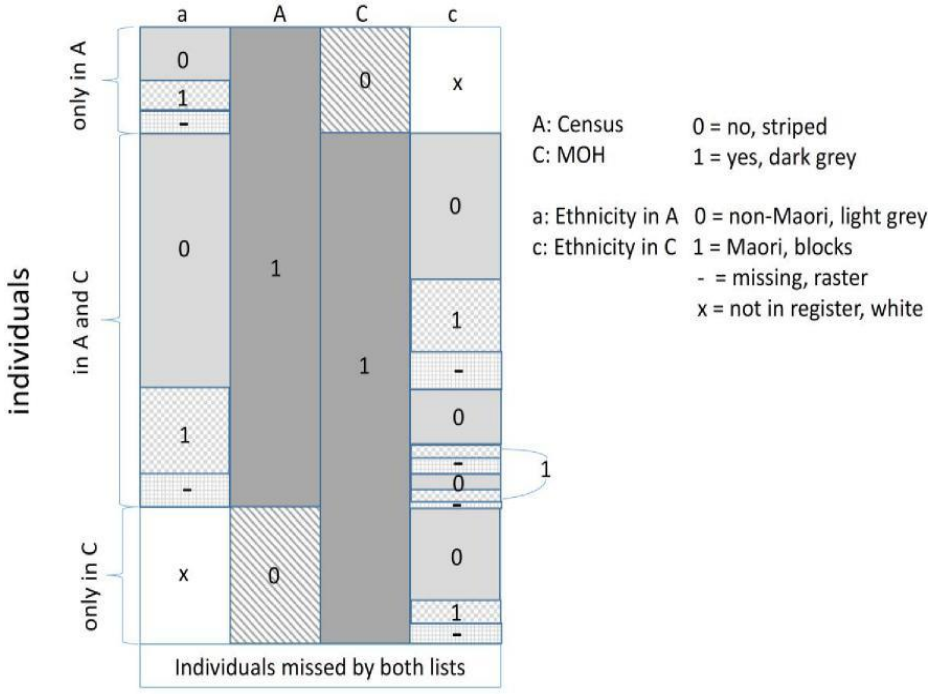


Figure 1: Graphical representation of two linked registers

brackets. The maximal model [Ac][ac][Ca] is saturated in the sense that the fitted values are equal to the observed values. The result is given in Table 1, Panel 2. Due to the fitted model, in each of the three estimated 2×2 subtables the a*c odds ratio is identical and equal to 377.9. The lower right 2×2 table in Panel 2 of Table 1 shows the estimated numbers of people missing from both census and MOH. These numbers there are relatively low, due to the large overlap of the two registers. The estimated total population size for New Zealand is 4,383,613.7. The census and the MOH differ in which part of this

total is Māori. For the Census the estimated number is 721,971. For the MOH it is 640,711.

3.2 Four registers

We now add the other two registers, DIA and MOE, to the analysis. Now the maximal model is [ABCd][ABDc] [ACDb][BCDa] [ABcd][ACbd] [ADbc][BCad] [BDac][CDab] [Abcd][Bacd] [Cabd] [Dabc][abcd]. Notice that, as for the two registers, a capital variable label cannot be in the same interaction term as a lower case variable label, as these interactions cannot be estimated from the data. Notice that the assumptions become less and less demanding as more registers are involved. The number of unique individuals in the four linked registers is 4,401,282, and the estimated number missed by all registers is 25,939, giving an estimated population size of 4,427,221.

The estimated numbers of Maori are displayed in Table 2. To arrive at a final estimate of the number of non-Māori and Māori we describe two approaches, both using the concept of measurement error. Consider the margins of the ethnicity variables a; b; c and d of the four registers. A statistical approach to measurement error is to make use of a latent class model (McCutcheon, 1987). See Table 3. In this latent class model, the first latent class is to be interpreted as the class for non-Māori, and the estimated probability of falling in this class is 0.826. The probability for the Māori class corresponds to an estimated Māori population size of about 770,000. Estimated conditional probabilities of being Māori for each latent class are also shown in Table 3; they are consistently low for the non-Māori latent class and high for the Māori latent class.

Panel 1: Observed counts

			C = 1		C = 0	Totals
		c = 0	c = 1	c = -	c = x	
A = 1	a = 0	3,004,329	31,998	150,855	38,640	3,225,822
	a = 1	108,192	435,468	12,402	4,377	530,439
	a = -	16,512	2,769	894	435	20,160
A = 0	a = x	398,838	146,985	24,642	-	570465
Totals		3,527,871	617,220	188,793	43,452	4,377,336

Panel 2: Fitted values under [Ac][ac][Ca]

		C = 1		C = 0		Totals
		c = 0	c = 1	c = 0	c = 1	
A = 1	a = 0	3,170,298.4	33,791.2	38,619.1	411.6	3,243,120.3
	a = 1	111,244.8	448,084.6	879.3	3,541.9	563,750.6
A = 0	a = 0	402,713.4	10,772.5	4,905.7	131.2	418,522.8
	a = 1	14,131.1	142,848.3	111.7	1,129.2	158,220.3
Totals		3,698,387.7	635,496.6	44,515.8	5,213.9	4,383,613.7

Table 1: Census *(A)* linked to MOH (C). Covariate Ethnicity in *A* is denoted by a and ethnicity in *C* is denoted by c, where a and c have levels '0' (non-Māori), '1' (Māori), '-' (missing) and 'x' (not in register). Observed counts have been randomly rounded to protect confidentiality. Source: Stats NZ.

	Census	DIA	MOH	MOE
non-Māori	3,690,913	3,668,349	3,782,239	3,665,099
Māori	736,308	758,872	644,983	762,122

Table 2: Summary of Census linked to DIA, MOH and MOE, estimated numbers

		census	DIA	MOH	MOE
	π_x	$\pi^a_{r=1\|x}$	$\pi^b_{s=1\|x}$	$\pi^c_{t=1\|x}$	$\pi^d_{u=1\|x}$
Class 1	0.826	0.004	0.012	0.003	0.014
Class 2	0.174	0.939	0.930	0.824	0.924

Table 3: Estimates of latent class model with two latent classes

3.3 Three registers without the Census

We also made estimations for three registers without the Census, see Table 4. We also present estimates derived only from the three administrative data sources, so that we can see what would happen if the census were replaced entirely by an administrative data-based system. The observed number of individuals in at least one of the registers is 4,377,573. We estimate an additional 24,058 individuals missed by all three registers. This leads to a total population size of 4,401,631. This is somewhat less than the four register estimate of 4,427,221.

	DIA	MOH	MOE
non-Māori	3,599,611	3,760,211	3,625,453
Māori	802,020	641,421	776,179

Table 4: Summary of DIA, MOH and MOE, ignoring census, estimated numbers.

4. Discussion and Conclusion

Van der Heijden et al. (2018) presented an approach for estimating the margins of auxiliary variables in the dual system estimation framework. They suggested that more experience with applications of this methodology was needed to be able to judge its usefulness. Here this approach is extended to multiple system estimation with four registers, and a more complicated missing data structure. We conclude that the methods of van der Heijden et al. (2018) provide stable results that allow for detailed interpretation of the processes of inclusion in the registers considered, and of recording Māori status.

References

1. Brown, J.J., C. Sexton, O. Abbott, and P.A. Smith (2019, in press). The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS.*
2. Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society: Series A*, 162(2), 247-267.
3. De Waal, T., A. van Delden and S. Scholtus (2017). Multi-source Statistics: Basic Situations and Methods. The Hague, Statistics Netherlands, discussion paper 2017/12.
4. McCutcheon, A.L. (1987) *Latent class analysis*. Sage, Newby Park.
5. Reid, G., C. Bycroft, and F. Gleisner (2016). Comparison of ethnicity information in administrative data and the census. Statistics New Zealand, Christchurch. https://www.stats.govt.nz/assets/ Research/Comparison-of-ethnicity-information-in-administrative-data-and-the-census/comparison -of-ethnicity-information-in-administrative-data-and-the-census.pdf.
6. Statistics New Zealand (2005). Statistical standard for ethnicity. Christchurch, Statistics New Zealand. http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/ ethnicity.aspx.
7. Statistics New Zealand (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
8. Statistics New Zealand (2014). An overview of progress on the potential use of administrative data for census information in New Zealand: Census Transformation programme. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
9. Statistics New Zealand (2017). Experimental population estimates from linked administrative data: 2017 release. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
10. Statistics New Zealand (2018). Experimental ethnic population estimates from linked administrative data. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
11. Van der Heijden, P.G.M., P.A. Smith, M. Cruyff, and B. Bakker (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. Journal Official Statistics, 34, 239-263.
12. Zhang, L.-C., and R.L. Chambers (Eds) (2018). Analysis of integrated data. Boca Raton: CRC Press.

13. Zwane, E., and P. G. M. van der Heijden (2007). Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. *Statistics in Medicine*, 26, 1069-1089.

A linkage error correction model for population size estimation with multiple sources

Daan Zult[1], Peter – Paul de Wolf[1], Bart Bakker[1,2], Peter van der Heijden[3]
[1]Statistics Netherland[1]
[2]VU University
[3]Utrecht University and University of Southampton

Abstract
A new method is described to do population size estimation, while linkage of sources occurs with errors. Our model is derived from a linkage error correction model introduced by Ding and Fienberg (1994). They show how to use linkage probabilities to correct the capture - recapture estimator for linkage errors, but only in the case of two sources and no covariates. A generalisation is proposed by incorporating the Ding & Fienberg model into the standard log - linear modelling approach used in multiple - recapture estimation. We show how the method performs in a simulation study with data that resemble real data.

Keywords
Multiple – recapture estimation; population size estimation; capture – recapture; record linkage; linkage errors

1. Introduction

This paper is a summary of Zult et al. (2019), which we refer to for a more extensive and elaborate discussion of this topic. The size of a partly observed population is often estimated with the capture – recapture (CR, for two sources) or multiple – recapture (MR, for multiple sources) method. An important assumption for these models is that records in different sources can be identified such that it is known whether these records belong to the same unit or not, i.e. records can be perfectly linked between sources. This assumption of perfect linkage is of particular relevance if identification is not obtained by some perfect identifier (like a tag or id-code) but by indirect identifiers (like name and address). In that case record are usually linked with probabilistic linkage (see Fellegi and Sunter, 1969, Winkler, 1988 or Jaro, 1989) and the perfect linkage assumption is often violated which generally leads to a biased population size estimate (PSE) (Gerritse et al., 2017).

A solution to this problem was provided by Ding and Fienberg (1994) (DF), Di Consiglio and Tuoto (2015) (DC&T_15) and De Wolf et al. (2018) (DW). These authors show how to use linkage probabilities to correct the capture -

[1] The authors like to thank Jan van der Laan from Statistics Netherlands for his review of the final version of this the paper.

recapture estimator for linkage errors. Recently, Di Consiglio and Tuoto (2018) (DC&T_18) extended their method to three sources.

In this paper we provide a general framework that allows us to extend this work further in two ways, with covariates and multiple sources. This is done by generalising the standard log - linear modelling approach used in multiple - recapture estimation such that it incorporates linkage error correction. This leads to the weighted multiple – recapture (WMR) model and is discussed in section 2. In section 3 we show the results of a simulation study that tests the WMR model.

2. Methodology

We first introduce some formal notation. s defines the source, where in standard CR = (1,2) and in MR = (1,2, ...). Next, we define the linked 'register' R_{t-1} as:

$$R_{t-1} = \begin{cases} R_{0=}\, S_1 \\ R_{1=}\, L_1(S_1, S_2) \\ R_{2=}\, L_2(R_1, S_3) \\ \quad \vdots \\ R_{t-1=}\, L_t(R_{t-2}, S_t) \end{cases},$$

where R_t refers to a set of t + 1 sequentially linked sources and L_t refers to the linkage process that links R_{t-1} with S_{t+1}. In case of CR this reduces to $R_1 = R = L(S_1, S_2)$. The *true* cell counts, *estimated* cell counts and *observed* cell counts (i.e. the counts of records that are linked and not linked between R_{t-1} and S_{t+1}) are denoted as $m_{ij} = (m_{11}, m_{10}, m_{01})$, $\hat{m}_{ij} = (\hat{m}_{11}, \hat{m}_{10}, \hat{m}_{01})$ and $n_{ij} = (n_{11}, n_{10}, n_{01})$ respectively. Here $i \in \{1,0\}$ corresponds to records in and not in R_{t-1} and $j \in \{1,0\}$ corresponds to records in and not in S_{t+1}. When there are no linkage errors, the true cell counts are equal to the observed cell counts, i.e. $m_{ij} = n_{ij}$. Furthermore, we define $m^*_{ij} = (m^*_{11}, m^*_{10}, m^*_{01})$ and $n^*_{ij} = (n^*_{11}, n^*_{10}, n^*_{01})$ as the true and observed cell counts in a random sample from R_{t-1} called a rematch or audit study (for a discussion on the difference between rematch and audit sample, which is small, we refer to Zult et al. (2019)). Beside that m^*_{ij} refers to a subsample, the difference between m^*_{ij} and m_{ij} is that in the presence of linkage errors m^*_{ij} is assumed to be known while m_{ij} is not. Finally, we introduce $p = 1, \ldots, P^t$ which are the records in R_t. Under perfect linkage this implies that all records refer to unique units/individuals, but in case of linkage errors two records in R_t might belong to different units/individuals or one record in R_t might represent two or more units.

The derivation of the WMR model follows three steps. First the D&F model is written as log – linear Poisson regression model. Second, the dependent variable in this model is corrected for linkage errors in case of two sources but

with covariates. These two steps are discussed in section 2.1. Third, this model is extended towards multiple – sources, which is discussed in section 2.2.

2.1 Capture - recapture estimation and linkage error correction

In the most basic case of CR the PSE is given by the standard Petersen (Petersen, 1986, Lincoln, 1930) formula:

$$\widehat{M}_{Petersen} = m_{11} + m_{10} + m_{01} + \frac{m_{10}m_{01}}{m_{11}} = \frac{(m_{11}+m_{10})((m_{11}+m_{01})}{m_{11}} = \frac{m_{1+}m_{+1}}{m_{11}} \quad (1),$$

where under the appropriate assumptions $\widehat{M}_{Petersen}$ is an unbiased estimate of the true population size (Wolter, 1986). The Petersen estimator is closely related to a fitted value obtained from a log - linear Poisson regression model with cell counts data (e.g. see Cormack, 1989), i.e.:

$$E[m_{ij}] = e^{(\beta_0 + \beta_1 i + \beta_2 j)} \text{ for } i, j \in \{1,0\} \quad (2),$$

where m_{ij} serves as the dependent variable in the log - linear regression model. The Poisson regression model uses maximum likelihood to obtain estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. An important difference between equation (1) and (2) is that (2) can be easily extended with additional sources or categorical covariates.

When the appropriate assumptions are not met, for instance records are not perfectly linked, $\widehat{M}_{Petersen}$ is biased. Therefore D&F developed a linkage error correction method that uses a rematch study from which they calculate the linkage error probabilities that are used to correct the PSE for linkage errors. DW show that this correction method can be written as:

$$\widehat{M}_{D\&F} = \frac{m_{1+}m_{+1}}{\hat{m}_{11}} \quad (3),$$

where $\hat{m}_{11}$ is the estimated number of links between both sources that takes linkage errors into account. Combining equation (1), (2) and (3) allows us to write:

$$E[\hat{m}_{ij}] = e^{(\beta_0 + \beta_1 i + \beta_2 j)} \text{ for } i, j \in \{1,0\} \quad (4)$$

where $\hat{m}_{11} = n_{11}\frac{m_{11}^*}{n_{11}^*}$, $\hat{m}_{10} = n_{1+} - \hat{m}_{11}$ and $\hat{m}_{01} = n_{+1} - \hat{m}_{11}$ (see Zult et al. (2019) for a more extensive derivation). In words, equation (4) constitutes the same model as equation (2), except the dependent variable m_{ij} is replaced by $\hat{m}_{ij}$, where $\hat{m}_{ij}$ is simply a vector of estimated cell counts that is based on the results of the audit study. Here we should note that the calculation of $\hat{m}_{11}$ is independent of the exact linkage procedure L . In fact, the only thing that matters is that the fraction $\frac{m_{11}^*}{n_{11}^*}$ is a consistent estimate of $\frac{m_{11}}{n_{11}}$, which implies that the audit study should be representative for R.

Equation (4) allows for the inclusion of covariates in the same way as in a regular log - linear Poisson regression, which implies that $\hat{m}_{ij}$ must be separated further into groups (e.g. male/female) and this categorical covariate can be added to the regression equation. We refer to this extension of the D&F model as the weighted CR (WCR) model. Why it is called 'weighted' will become clear in the next section.

2.2 The weighted – multiple recapture model

In section 2.1 we showed how the D&F model can be written as a log – linear Poisson regression model and how (categorical) covariates can be added to this equation by splitting - up $\hat{m}$ into smaller groups. This implies that after this procedure we have for each cell count both an estimated and observed cell count. Here we should note that each cell count consists of records, so for each record we can calculate its weighted contribution to its estimated cell count, i.e.:

$$w_p = \frac{\hat{m}_p}{n_p} \tag{5}$$

where $\hat{m}_p$ and n_p refer to the estimated and observed cell count of record p. E.g., when we ignore covariates and record p is linked between S_1 and S_2 , $\hat{m}_p$ and $n_p = 11$. Now w_p is a record level weight that sums up to the different elements in $\hat{m}$. Adding up over w_p is similar to the case of no linkage errors where each record has a weight of 1 and is added up to obtain (the true and observed) cell counts. However, when we want to extend the model such that it can deal with multiple – sources, we can write w_p as:

$$w_p^t = w_p^{t-1} \frac{\hat{m}_p^t}{n_p^t} \tag{6},$$

with $w_p^{t=0} = 1$, $n_p^t = \sum_{p \epsilon cell\ count} w_p^{t-1}$. Under equation (6) w_p^t is updated after every linkage procedure, which can be repeated for each new source. After the update of w_p^t the estimated cell count elements of $\hat{m}$ can be calculated by summing up w_p^t over the records p that belong to that cell, where $\hat{m}$ does not only distinguish between i and j but may distinguish between any number of sources and categorical covariates. The WMR model can then be written as:

$$[\hat{m}_{zt}] = e^{f(\beta, Z_t)} \tag{7},$$

where $\hat{m}_{zt}$ is the estimated cell count vector that depends on $Z_t = (R_{t-1}, X)$ with a set of categorical covariates, according to some function $f(\beta, Z_t)$ with β a parameter vector.

3. Results

We evaluate the WMR model with a simulation study. In this study the true population size (TPS) is known and will be compared with estimates of the population size. We use a (quasi – real) dataset that is a publicly available fictitious population dataset of 26 625 persons that is representative for the UK population census. It was created in a European project on data integration (McLeod, Heasman and Forbes, 2011) that ran from 2009 to 2011. The dataset has linkage keys such as address and birthdate but also covariates such as gender and age. By generating sources from this quasi - real dataset, outcomes may reflect reality to some extent.

The main goal of this simulation study is to evaluate the performance of the WMR model. The WMR model is applied within different scenarios, where scenarios differ with respect to three elements:

1. Covariate dependence of capture probabilities, which implies that the probability of a record to be in S_1, S_2 and S_3 may vary due to differences in the covariate values of records (e.g. a male may have a higher probability to be in S_1 and a lower probability to be in S_2).
2. Source dependence of capture probabilities, which implies that the probability of a record to be in S_1, S_2 and S_3 may depend on this record being in another source (e.g. a record in S_1, may have a different probability to be in S_2 than a record that is equal in all other aspects except being in S_1).
3. Linkage errors in the linkage procedure; sources are linked either with errors or are linked perfectly without errors.

These three elements are of particular interest, because they are the sources of bias that the WMR model aims to correct for while the alternative models should suffer from at least one of them. They lead to four different scenarios that can be seen in table 1.

Table 1: Simulation study scenarios.

Scenario	Linkage errors	Covariate dependence	Source dependence
1	Yes	No	No
2	Yes	Yes	No
3	Yes	No	Yes
4	Yes	Yes	Yes

Each scenario is replicated 1 050[2] times and in each replication a population of 10 000, together with three sources of approximately 8 000, 5 000 and 2000 records is generated, where the generation of sources differs between

[2] The number is 'only' 1 050 because we use a spark cluster of fifteen cores (available at Statistics Netherlands mainly for Big Data related computations) that each does 70 replications with different random seeds, in which each single replication takes about 10 minutes. In total it takes almost two days to run all four scenarios, which is mainly due to the computation time of the probabilistic linking the three sources.

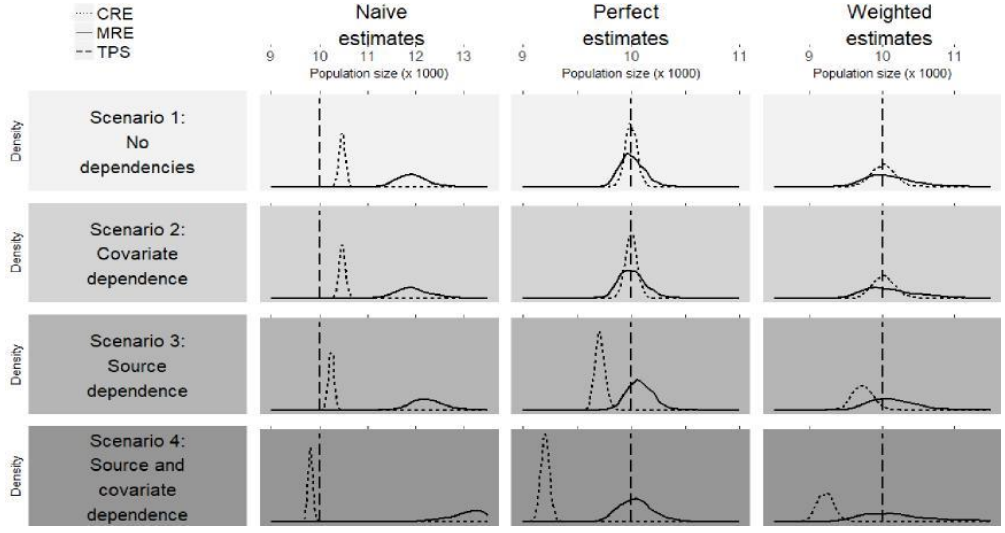
scenarios. For further details on the simulation setup we refer to Zult et al. (2019), in the next section we discuss the results.

3.1 Simulation outcome

In figure 1 below the simulation results of the four scenarios are presented as density plots where each density plot contains the CR estimate (CRE), MR estimate (MRE) and TPS that are calculated in three different ways, i.e. naïve (with linkage errors and without correction), perfect (without linkage errors and without correction) and weighted (with linkage errors and with correction). The results are in figure 1 on the next page.

Ideally the density of an estimate revolves around the TPS of 10 000. However, the first column shows that the densities of the naïve estimates do not, which implies that the linkage errors indeed lead to biased PSEs. Furthermore, in case of perfect linkage, in scenario 1 and 2 both the CREs and MREs revolve around the TPS. However, when source dependence is introduced in scenario 3 and 4 the CR model (necessarily) fails while the MR model still performs well. This failure of the CR model implies that it suffers from source dependence as intended by the simulation setup. Finally, the third column contains the weighted estimates. Here the (weighted) CR model performs well in scenario 1 and 2, which implies the WCR model is able to correct for both linkage errors and covariate dependence simultaneously. However, in scenario 3 and 4 the CR model logically fails, because it is unable to deal with source dependence. Fortunately, in these scenarios the density of the MREs still revolves around the TPS, which implies that the WMR model indeed corrects for linkage errors, covariate dependence and source dependence simultaneously.

Figure 1: Density plots of two PSEs with three dependent variables and four scenarios (table 1).



4. Discussion and Conclusion

In this paper we derived and tested the WMR model for population size estimation corrected for linkage error. The model is derived from the D&F model and is a more general extension than the models developed by DC&T (2015, 2018) and De Wolf et al. (2018) because it can deal with three or more sources and covariates. Furthermore, the WMR model is incorporated in the more general family of log - linear regression models and therefore no longer has to be studied as an isolated issue in CR and MR models. Finally, the WMR model was tested and approved in a simulation study.

In theory the WMR model might be an improvement on the D&F model, they both still require the availability of a rematch (for D&F) or audit (for WMR) study. The advantage of the WMR model is that an audit study might be easier to obtain, because it has lower requirements (it needs to be constructed on the cell count level instead of the much more detailed records matching pair level). However, the incorporation of covariates and additional sources in the WMR model also puts additional constraints on the audit study, in the sense that the audit study should include these same covariates and additional sources. Given that the sample that underlies the audit study must be representative for R^t, this might be more difficult for increasing t.

Also, we should note that we paid little attention to the impact of the exact linkage procedure. In section 2 we developed the WMR model in the context of the common sequential linkage approach, in which first two sources are linked and a third source is linked to this combined source. However, it is also possible that sources are linked pairwise or simultaneously. These approaches are less common because they suffer either from computational (i.e. the number of potential matches between multiple sources increases exponentially) or methodological (e.g. what to do with inconsistent matching patterns like $A \rightarrow B,\ B \rightarrow C,\ C \nrightarrow A$?). Furthermore, in the simulation study of section 3 we applied probabilistic linkage that uses techniques developed by Fellegi and Sunter (1969), Winkler (1988) and Jaro (1989) that aim to optimise the quality of matches on the matching pair level, while matching techniques that are designed to optimise the quality of the matches on the cell count level might already significantly reduce the problem of linkage errors in population size estimation.

Another point that deserves some discussion is the 'individual starting weight of 1'. Lists or registers of individuals sometimes also contain individual sample weights, which indicate the size of the group that this individual represents as part of the total population. There is no reason why these sample weights cannot replace the starting weights of 1 in the WMR model. Furthermore, when additional sources also contain sample weights they can be used to calculate n^t, n^{*t} and m^{*t} in a slightly different way, i.e. simply by adding up sample weights instead of counting. This way we would get 'linkage

error corrected sample weights'. However, we should note that the presence of sample weights usually implies that the source only covers a (very) small part of the population, so when multiple sources contain sample weights the probability of matches becomes low, leading to very low cell counts and an unreliable PSE. How exactly sample weights can be combined with linkage and linkage error correction requires further research.

References

1. Cormack, R. M. (1989). Log-linear models for capture-recapture. Biometrics, 45, 395 – 413.
2. De Wolf, PP., Van Der Laan, J. and Zult, D. (2018). Joining correction methods for linkage error in capture-recapture, 45, Discussion paper, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2018/18/connecting-correction-methods-for-linkage-error-in-crc. To appear in Journal of Official Statistics, September 2019.
3. Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. Journal of Official Statistics, 31, 415 – 429.
4. Di Consiglio, L. and Tuoto, T. (2018). Population Size Estimation and Linkage Errors: the Multiple Lists Case. Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 889–908.
5. Ding, Y. and Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. Survey Methodology, 20, 149 – 158.
6. Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183 – 1210.
7. Gerritse, S.C., Bakker, B.F.M. and Van der Heijden, P.G.M. (2017). The impact of linkage errors and erroneous captures on the population size estimator due to implied coverage. Discussion paper 2017 - 16, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2017/39/impact-of-linkage-errors-and-erroneous-captures
8. Jaro, M. (1989). Advances in Record Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida. Journal of American Statistical Association 84: 414–420.
9. McLeod, P., Heasman, D. and Forbes, I. (2011). Simulated data for the on the job training. Essnet DI. Available at: http://www.cros-portal.eu/content/job-training.
10. Lincoln, F. C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns, U.S. Dept. Agric., 118: 1-4.

11. Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfiord from the German Sea. Report of the Danish Biological Station, 6, 5 – 84.
12. Winkler, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi - Sunter model of record linkage. Section on Survey Research Methods, 667 – 671.
13. Wolter, K.M. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81, 338 – 346.
14. Zult, D.B.,. De Wolf, PP., Bakker, B.F.M. and van der Heijden, P.G.M. (2019). A general framework for multiple - recapture estimation that incorporates linkage error correction. Discussion paper 2019 - 12, Statistics Netherlands, The Hague/Heerlen. Available at: https://www.cbs.nl/en-gb/background/2019/19/correcting-for-linkage-errors-in-the-multiple-capture

Markov-Switching three-pass regression filter

Pierre Guérin[1], Danilo Leiva-Leon[2], Massimiliano Marcellino[3]
[1]OECD
[2]Banco de España
[3]Bocconi University, IGIER and CEPR

Abstract

We introduce a new approach for the estimation of high-dimensional factor models with regime-switching factor loadings by extending the linear three-pass regression filter to settings where parameters can vary according to Markov processes. The new method, denoted as Markov-switching three-pass regression filter (MS-3PRF), is suitable for data sets with large cross-sectional dimensions, since estimation and inference are straightforward. Both Monte Carlo simulations and empirical applications show significant predictive gains when using the proposed framework.

Keywords

Factor model; Markov-switching; Forecasting

1. Introduction

This paper introduces a new approach for the estimation of high-dimensional factor models with regime-switching factor loadings. Our modelling approach builds on Kelly and Pruitt (2015), who developed a new estimator for factor models—the three-pass regression filter (3PRF)—that relies on a series of ordinary least squares (OLS) regressions. As emphasized in Kelly and Pruitt (2015), the key difference between principal component analysis (PCA) and the 3PRF approach is that PCA summarizes the cross-sectional information based on the covariance within the predictors, whereas 3PRF condenses cross-sectional information based on the correlation of the predictors with the target variable of the forecasting exercise, thereby extending partial least squares.

In this paper, we extend the 3PRF approach by introducing regime-switching parameters in the linear 3PRF filter. This new framework is denoted as Markov-switching three-pass regression filter (MS-3PRF). A key advantage of this approach is that it is well suited to handle high-dimensional factor models, as opposed to the existing regime-switching factor models that can handle only models with limited dimensions due to computational complexity (see, e.g., Camacho et al. (2012)). Our approach is attractive in that the estimation strategy only requires estimating a series of univariate Markov-switching regressions. As such, it is computationally straightforward

to implement and offers a great deal of flexibility in modelling time variation since we do not restrict the regime changes in the cross-sectional dimension to be governed by a single or a limited number of Markov chains.

2. Markov-Switching Three-Pass Regression Filter

One key reason for the absence of a significant literature on large-scale Markov-switching factor models relates to the computational challenges associated with the estimation of such models. We present here the Markov-switching three-pass regression filter, which circumvents these difficulties. Our setting is similar to that in Kelly and Pruitt (2015), who introduced the linear 3PRF, but the key novelty is that we include time variation in the model parameters via Markov processes. Specifically, we have the following model:

$$y_{\boldsymbol{t}} = \beta_0(S_{yt}) + \beta(S_{yt})f_{t-1} + \eta_t, t = 1, \dots, T, \quad (1)$$

$$z_{\boldsymbol{jt}} = \lambda_{0,j}(S_{zjt}) + \lambda_j(S_{zjt})f_t + w_{jt}, j = 1, \dots, k_f \quad (2)$$

$$x_{\boldsymbol{it}} = \emptyset_{\boldsymbol{0},\boldsymbol{i}}(S_{xit}) + \emptyset_{\boldsymbol{f},\boldsymbol{i}}(S_{xit})f_t + \emptyset_{g,i}(S_{xit})g_t + \varepsilon_{it}, i = 1, \dots, N, \quad (3)$$

where y is the scalar target variable of interest for forecasting; $f_t = (f_{1t}, \dots, f_{kft})'$ is a $k_f \times 1$ vector of unobservable factors, with associated slope coefficients $\beta(S_y)$; $z_{jt,}\, j = 1, \dots, k_f$, are so-called proxy variables driven by the same factors as y, $\boldsymbol{f_t}$, with variable specific loadings $\boldsymbol{\lambda_j}(S_{zjt})$; $x_{it}\, i = 1, \dots, N$, are variables driven by the $\boldsymbol{f}_t$ factors but also by the k_g (unobservable) factors in the vector $\boldsymbol{g}_t$, with associated variable specific loadings $\emptyset_{\boldsymbol{f},\boldsymbol{i}}(S_{xit})$ and $\emptyset_{\boldsymbol{g},\boldsymbol{i}}(S_{xit})$ respectively; $\beta_0(S_{yt})$, $\lambda_{0,j}(S_{zjt})$, $\emptyset_{\boldsymbol{0},\boldsymbol{i}}(S_{xit})$ are intercepts. As anticipated, the coefficients in (1) to (3) are time-varying and driven by variable specific and independent across variables M-state Markov chains: S_{yt}, S_{zjt} and $S_{xit}\, j = 1, \dots, k_f$ and $i = 1, \dots, N$. Each Markov chain is governed by its own $M \times M$ transition probability matrix,

$$P_q = \begin{pmatrix} p_{q,11} & p_{q,21} & \cdots & p_{q,M1} \\ p_{q,12} & p_{q,22} & \cdots & p_{q,M2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{q,1M} & p_{q,2M} & \cdots & p_{q,MM} \end{pmatrix}, \quad (4)$$

for $q = y, z_1, \dots, z_{kf}, x_1, \dots, x_N$.

Given the model in equations (1) to (3), our algorithm for the MS-3PRF model consists of the following three steps:

- *Step 1:* Time-series regressions of each x_{it} on the proxy variables $z_{jt}, j = 1, \dots, k_f$. Hence, defining $\boldsymbol{z}_t = (z_{1t}, \dots, z_{kft})'$, we run N Markov-switching regressions

$$x_{it} = \emptyset_{\mathbf{0},i}(S_{xit}) + \emptyset_{\boldsymbol{i}}(S_{xit})\boldsymbol{z}_t + \varepsilon_{it}, t = 1, \dots, T, \quad (5)$$

where $\varepsilon_{it} \sim NID(0, \sigma_{xi}^2(S_{xit}))$, and keep the (variable specific) estimates of $\emptyset_{\boldsymbol{i}}(S_{xit})$, denoted by $\widehat{\emptyset}_i(S_{xit})$, where $\widehat{\emptyset}_i(S_{xit})$ is a $1 \times k_f$ vector, for $i = 1, \dots N$. All regime-switching regressions are estimated via (pseudo) maximum likelihood, which is why we have made a normality assumption for ε_{it} that is instead not required when estimating by OLS the linear version of the 3PRF. Note also that the Markov chains in (5) are the same as in (3). We also define for later use in the second step of the algorithm the variables $\widehat{\emptyset}_{A,it}$ and $\widehat{\emptyset}_{B,it}$. The variables $\widehat{\emptyset}_{A,it}$ are a weighted average of the estimated regime-specific factor loadings:

$$\widehat{\emptyset}_{A,it} = \sum_{j=1}^{M} \widehat{\emptyset}_i(S_{xit} = j) P(S_{xit} = j|\Omega_T), \quad (6)$$

where $P(S_{xit} = j|\Omega_T)$ is the smoothed probability of being in regime j given the full sample information Ω_T. the variables $\widehat{\emptyset}_{B,it}$ are instead defined as selected factor loadings:

$$\widehat{\emptyset}_{B,it} = \sum_{j=1}^{M} \widehat{\emptyset}_i(S_{xit} = j) P(S_{xit} = j|\Omega_T), \quad (7)$$

where $I(.)$ is an indicator function that selects the regime with the highest smoothed probability, $P(S_{xit} = j|\Omega_T)$, at time t.

- *Step 2:* Cross-section regressions of the x_{it} on either $\widehat{\emptyset}_{A,it}$ or $\widehat{\emptyset}_{B,it}$. Hence, we run T linear regressions

  $$x_{it} = \alpha_{0,t} + \widehat{\emptyset}_{q,it} f_t + \upsilon_{it}, i = 1, \dots, N, \quad (8)$$

  $\upsilon_{it} \sim IID(0, \sigma_{vt}^2)$, with t = 1,..., T, $q = A$ or $q = B$, and we keep (for each t) the OLS estimates $\hat{\boldsymbol{f}}_{\boldsymbol{t}}$, where $\hat{\boldsymbol{f}}_{\boldsymbol{t}}$ is a $k_f \times 1$ vector.
- *Step 3:* Time-series regression of y_t on $\hat{f}_{t-1}$. hence, we run one Markov-switching regression:

  $$y_t = \beta_0(S_{yt}) + \beta(S_{yt})\hat{f}_{t-1} + \eta_t, \; t = 1, \dots, T, \quad (9)$$

  $\eta_t \sim NID(0, \sigma_\eta^2(S_{yt}))$, and we keep the maximum likelihood estimates $\hat{\beta}_0(S_{yt})$ and $\hat{\beta}(S_{yt})$. We calculate the forecast $\hat{y}_{T+1|T}$ as:

  $$\hat{y}_{T+1|T} = \sum_{j=1}^{M} (P(S_{yT+1} = j|\Omega_T)\hat{\beta}_0(S_{yT+1} = j) + P(S_{yT+1} = j|\Omega_T)\hat{\beta}(S_{yT+1} = j)\hat{f}_T), \quad (10)$$

  where $P(S_{yT+1} = j|\Omega_T)$ is the predicted probability of being in regime j in period $T + 1$ given the information available up to time T, Ω_T.

3. Empirical Application

In this forecasting exercise, we construct factors from a cross-section of nominal bilateral U.S. dollar (USD) exchange rates against a panel of 26

currencies. We extract factors from the MS-3PRF, MSS-3PRF, linear 3PRF, PCA, TPCA and PC-LARS. We then use the resulting factors to forecast selected bilateral exchange rates. (All currency pairs use the USD as numéraire.) The choice of the data set draws from the exercise in Greenaway-McGrevy et al. (2016). The data set is monthly, and the full sample size extends from January 1995 to December 2015. The data are obtained from the International Financial Statistics of the International Monetary Fund, and the monthly data are taken as the monthly average of daily data. The data set consists of the currencies of Australia (AUS),Brazil (BRA), Canada (CAN), Chile (CHI), Columbia (COL), the Czech Republic (CZE), the euro (EUR), Hungary (HUN), Iceland (ICE), India (IND), Israel (ISR), Japan (JPN), Korea (KOR), Mexico (MEX), Norway (NOR), New Zealand (NZE), the Philippines (PHI), Poland (POL), Romania (ROM), Singapore (SIN), South Africa (RSA), Sweden (SWE), Switzerland (SUI), Taiwan (TAI), Turkey (TUR) and the United Kingdom (GBR).

The left-hand side of Table 1 reports point forecasting results for specific currencies: the Canadian dollar (CAD), the euro (EUR), the Japanese yen (JPY) and the British pound (GBP), all relative to the USD. These are G7 currencies, and among the most traded currency pairs according to the Bank for International Settlements Triennial Central Bank Survey. The point forecast results are presented as the MSFE of a specific approach relative to the MSFE obtained from the no-change forecast. The no-change forecast is the standard benchmark in the exchange rate forecasting literature (see, e.g., Rossi (2013)). We also report the results of the Diebold and Mariano (1995) test of equal out-of-sample predictive accuracy using the no-change forecast as a benchmark. First, the models' forecasting performance relative to the no-change forecast is typically the strongest for forecast horizon $h = 1$ (except for the JPY/USD). The improvement in forecast accuracy relative to the random walk is also statistically significant according to the Diebold and Mariano test of equal MSFE when forecasting the Canadian dollar at forecast horizon $h = 1$ across most approaches (this is also true to a lesser extent for the British pound). Second, the PC-LARS approach performs best for forecast horizon $h = 1$ when forecasting the British pound. Moreover, the MS-3PRF (first and third pass) approach performs best when forecasting the Canadian dollar for forecast horizons $h > 1$. Third, for the Canadian dollar and the Japanese yen, modelling time variation in the forecasting equation is relevant in that this leads to substantial forecasting improvement over the no-change forecast at distant forecast horizons $h = \{9\}$ for the Japanese yen and $h = \{2, 3, 6, 9, 12\}$ using the MS-3PRF (first and third passes) approach.

Next, the right-hand side of Table 1 shows the directional accuracy forecasting results, which are broadly in line with the point forecast

results. Under the null hypothesis of no directional accuracy, one would expect a success ratio of 0.5. We also report the results of the Pesaran and Timmermann (2009) test to evaluate the statistical significance of the directional accuracy results. Across all forecasting approaches, the success ratios tend to be stronger for forecast horizon $h = 1$, except for the JPY/USD. In those cases, the improvements in directional accuracy are often statistically significant according to the Pesaran and Timmermann (2009) test. It is also interesting to note that the success ratios are especially strong at distant forecast horizons for selected currencies, as high as 72.6 per cent for the CAD/USD and 77.0 per cent for the JPY/USD in the case of the MS-3PRF with regime changes in the first and third passes.

Overall, while the differences in predictive accuracy tend to be small across forecasting approaches in terms of point forecasts, the gains in terms of directional accuracy are strong with the MS-3PRF approach and typically statistically significant according to the Pesaran and Timmermann (2009) test.[1]

References

1. Camacho, M., Pérez-Quirós, G., and Poncela, P. (2012). Markov-switching dynamic factor models in real time. CEPR Discussion Papers 8866, C.E.P.R. Discussion Papers.
2. Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics,* 13(3):253–63.
3. Greenaway-McGrevy, R., Mark, N., Sul, D., and Wu, J.-L. (2016). Identifying Exchange Rate Common Factors. *Mimeo Notre Dame.*
4. Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics,* 186(2):294–316.
5. Pesaran, M. H. and Timmermann, A. (2009). Testing Dependence Among Serially Correlated Multicategory Variables. *Journal of the American Statistical Association,* 104(485):325–337.
6. Rossi, B. (2013). Exchange Rate Predictability. *Journal of Economic Literature,* 51(4):1063–1119.

[1]We also conducted several Monte Carlo exercises to assess the finite sample performance of the proposed model. We find that the MS-3PRF performs favourably compared with alternative modelling approaches whenever there is structural instability in factor loadings.

Table 1: Out-of-sample exchange rate forecasting

Forecast horizon	1	2	3	6	9	12	1	2	3	6	9	12
	CAD/USD (MSPE)						CAD/USD (Success ratios)					
PCA	**0.871***	0.951	0.980	1.001	1.094	1.175	0.637**	0.602**	0.575*	0.522	0.504	0.407
TPCA	0.872*	0.955	0.986	0.996	1.088	1.182	**0.637****	0.620**	0.593**	0.522	0.487	0.407
PC-LARS	0.872	0.949	0.986	0.998	1.095	1.157	0.611**	0.602**	0.513	0.487	0.469	0.407
3PRF	0.963*	1.011	1.033	1.017	1.045	1.047	0.540	0.584**	0.558**	0.566**	0.504	0.451
MS-3PRF (first pass)	0.890	0.962	0.982	0.971	0.993	1.031	0.611**	0.611**	0.566*	0.540	0.522	0.504
MS-3PRF (first and third pass)	0.888**	**0.892**	**0.944**	**0.915**	**0.919**	**0.941**	0.575**	0.575*	**0.620***	**0.620****	**0.726****	**0.717****
MSS-3PRF (first pass)	0.911*	0.938*	0.987	0.994	1.020	1.058	0.620**	**0.620****	0.549	0.496	0.460	0.469
MSS-3PRF (first and third pass)	0.946	1.057	0.973	0.948	1.078	1.026	0.593**	0.478	0.575	0.584	0.620**	0.655**
	EUR/USD (MSPE)						EUR/USD (Success ratios)					
PCA	0.988	1.033	1.000	1.034	1.085	1.138	0.549	0.487	0.540	0.504	0.416	0.372
TPCA	0.995	1.049	1.010	1.061	1.109	1.155	0.558	0.487	0.566	0.487	0.443	0.354
PC-LARS	**0.961**	1.028	1.010	1.072	1.141	1.207	**0.566***	0.531	0.549	0.540	0.460	0.389
MS-3PRF (first pass)	0.994	1.008	0.996	1.002	1.036	1.095	0.549	0.540	0.575	0.566	0.469	0.469
MS-3PRF (first and third pass)	1.253	1.059	1.269	1.210	1.452	1.470	0.522	**0.549**	0.416	0.425	0.319	0.487
MSS-3PRF (first pass)	1.005	**0.985**	1.014	1.016	1.047	1.079	0.531	0.531	**0.593****	**0.558**	**0.531**	0.460
MSS-3PRF (first and third pass)	1.367	1.182	1.329	1.123	1.476	1.845	0.522	0.496	0.487	0.487	0.372	0.372
	JPY/USD (MSPE)						JPY/USD (Success ratios)					
PCA	1.091	1.093	1.108	1.070	1.101	1.145	0.460	0.496	0.469	0.496	0.504	0.425
TPCA	1.081	1.111	1.098	1.066	1.085	1.121	0.460	0.434	0.451	0.504	0.478	0.425
PC-LARS	1.066	1.066	1.078	1.078	1.074	1.117	0.531	0.504	0.496	0.487	0.469	0.425
3PRF	1.017	1.038	1.050	1.039	1.070	1.117	**0.575****	0.513	0.487	0.478	0.451	0.389
MS-3PRF (first pass)	1.031	1.060	1.054	1.066	1.078	1.122	0.487	0.487	0.496	0.451	0.487	0.363
MS-3PRF (first and third pass)	1.136	1.094	1.091	1.009	0.763**	1.051	0.522	0.478	0.451	**0.549**	**0.770****	0.460
MSS-3PRF (first pass)	1.058	1.061	1.055	1.043	1.074	1.078	0.478	0.522	0.531	0.487	0.460	0.443
MSS-3PRF (first and third pass)	1.345	1.240	1.241	1.159	**0.763***	1.068	0.558	**0.540**	**0.531**	0.549	0.735**	**0.513**
	GBP/USD (MSPE)						GBP/USD (Success ratios)					
PCA	0.803*	0.977	1.043	1.045	1.064	1.091	0.593**	0.540	0.513	0.434	0.522	0.575
TPCA	0.796*	0.965	1.023	1.043	1.064	1.091	0.593**	0.540	0.504	0.469	**0.558****	**0.584****
PC-LARS	**0.784***	**0.953**	1.029	1.035	1.073	1.114	**0.611****	**0.593****	0.531	0.531	0.425	0.496
3PRF	0.875	0.953	1.015	1.065	1.057	1.088	0.522	0.584**	**0.575****	0.460	0.549	0.443
MS-3PRF (first pass)	0.908*	0.980	1.010	1.032	1.064	1.094	0.566*	0.531	0.558	0.504	0.478	0.469
MS-3PRF (first and third pass)	0.977	1.083	1.032	1.113	1.128	1.114	0.531	0.549	0.540	0.513	0.416	0.469
MSS-3PRF (first pass)	0.893	0.984	1.037	1.045	1.074	1.119	0.549	0.549	0.566	0.522	0.496	0.460
MSS-3PRF (first and third pass)	1.392	1.523	1.302	1.361	1.397	1.354	0.513	0.478	0.540	**0.531**	0.496	0.496

Note: This table shows the relative mean square forecast error (RMSFE) for selected currency pairs (CAD/USD, EUR/USD, JPY/USD and GBP/USD) using PCA, TPCA, PC-LARS, linear 3PRF, MS-3PRF (first pass), MS-3PRF (first and third passes), MSS-3PRF (first pass) and MSS-3PRF (first and third passes) as forecasting approaches. Entries in bold indicate the best-performing approach for a specific horizon. Statistically significant reductions in the MSFE (or impovements in directional accuracy) relative to the random walk according to the Diebold-Mariano (Pesaran-Timmermann) test are indicated by asterisks (*denotes significance at the 10 per cent level. and ** denotes significance at the 5 per cent level).

62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur
18-23 August 2019

Forecasting household consumption components: A forecast combination approach

Angelia L. Grant, Liyi Pan, Tim Pidhirnyj, Heather Ruberl, Luke Willard

Abstract

This paper outlines a methodology for forecasting the components of household final consumption expenditure, which is necessary in order to forecast revenue collections from a number of different taxes. A forecast combination approach using autoregressive models, regressions on relative prices and the almost ideal demand system developed by Deaton and Muellbauer (1980) is found to offer a more robust forecasting framework than using one of the single models alone. In particular, the combination approach outperforms the almost ideal demand system, which is currently used by the Australian Treasury to forecast the components of consumption. The combination framework takes advantage of models that account for the persistence and longer-term trends experienced in a number of the consumption components, as well as shifts caused by evident relative price changes. A forecast combination framework is shown to be particularly useful when forecasting over a three-year forecasting period.

1. Introduction

Forecasts for each of the expenditure components of nominal GDP are important for forecasting tax revenue collections – different compositions result in different tax revenue forecasts. A particularly important task is the forecasting of the components of household final consumption expenditure. This is because different components of consumption are subject to different taxes. For example, alcohol, tobacco and fuel are subject to excise taxes, while motor vehicles may be subject to the luxury car tax. A number of the components of household final consumption expenditure – durables, other goods, electricity and gas, and other service – are also subject to the goods and services tax.

A wide variety of models can be used to forecast the components of household consumption, with different models using different types of information. Some models are good at accounting for the persistence and longer-term trends experienced in a number of the consumption components, while other models are better at taking into account shifts caused by relative price changes. It is also the case that some models are better at forecasting over shorter time horizons, while others are better over longer time horizons.

Under these circumstances, a forecast combination approach has a number of advantages. It allows the use of information across a number of models and the use of models that perform differently across different time

horizons. It is often the case that when forecasts from a variety of different models are appropriately combined, the forecast combination approach outperforms individual forecasts (see, e.g., Timmermann, 2006; Guidolin and Timmermann, 2009; Rapach et al., 2010).

This paper develops a forecast combination approach for the components of household consumption expenditure using autoregressive models, regressions on relative prices and the almost ideal demand system developed by Deaton and Muellbauer (1980).[1] The autoregressive models capture the persistence and longer-term trends in the consumption components, while the relative price regressions and the almost ideal system capture shifts in consumption components that are driven by relative price changes. At shorter forecasting horizons, models that capture short-run dynamics perform well, while at longer horizons models with trend terms and relative prices generally perform better based on root mean squared forecast errors.

Two forecast combinations are constructed – one based on equal weights and the other weighted based on forecasting performance according to rolling squared forecast errors.[2] The advantage of combining forecasts based on past forecast performance is that the forecast combination is robust to changes in modelling performance. That is, it accounts for the fact that certain models can improve or diminish in performance over particular time periods. However, it is also often found that equal weights perform strongly (see, e.g., Timmermann, 2006).

The forecast combinations generally perform better than the almost ideal demand system, which is the model currently used for estimating the household final consumption components. The fuels and lubricants and the electricity and gas components are the components where the forecast combination performance is closest to that of the almost ideal demand system. The forecast combination based on past forecast performance performs better than the equal weights model for all components.

2. Individual Forecasting Models

The first models considered are autoregressive models. These models take into account the persistence of the shares of consumption components, with the share of consumption on a particular component modelled to be a linear combination of past shares. The implicit assumption within autoregressive models is that consumption patterns tend to depend on those in recent periods, consistent with habit-forming preferences.

[1]There are, of course, other factors that might affect household consumption, such as changes in tax policy, income uncertainty and changes in wealth.

[2]For density forecasts of Australian output growth, inflation and interest rates, Gerard and Nimark (2008) use the predictive likelihood for combining forecasts from different vector autoregression models.

The main advantage of autoregressive models is that they perform well at modelling persistence. On the other hand, the disadvantage is that they do not use any other information, such as relative price movements. Each model is estimated using ordinary least squares regressions.

The autoregressive models do not account for changes in relative prices, which can drive important shifts in the share of each consumption component. As such, the next models considered are regressions on the relative price of the consumption component. The main advantage of relative price regressions is that they take into account information about relative price shifts. A disadvantage is that they do not include dynamics in the form of past consumption shares.

The almost ideal demand system (AIDS) of Deaton and Muellbauer (1980) takes into account that the consumption of a particular good or service depends not only on its own price, but also the relative prices of other goods and services which may be either complements or substitutes. It also takes into account an income effect, with each of the shares depending on total consumption expenditure. The main advantage of the almost ideal demand system is its strong theoretical grounding. But this strong theoretical grounding may mean that the model may be too restricted to fit the data well. In addition, the model has a large number of parameters.

3. Forecast Combination Approach

There are two methods used to construct the forecast combinations. The first method uses equal weights. The simple combination approach is often found to outperform other more sophisticated combination schemes. The second method weights the forecasts using past forecast performance. More specifically, each of the models for each of the household consumption components is weighted using a four-quarter rolling weight of the inverse of the sum of the squared forecast error. The four-quarter rolling weight strikes a balance between having relatively stable weights and weights that quickly adapt when there is a change in performance across models.

The approach of combining forecasts based on past forecast performance accounts for changes in modelling performance. That is, it captures the benefits of different modelling approaches and accounts for the fact that certain models can improve or diminish in performance over particular time periods and at different forecasting horizons.

4. Forecasting Results

This section reports the out-of-sample forecasting results for each of the models and for the forecast combinations. The out-of-sample forecasting is based on 12-step ahead forecasts, which are computed every 2 quarters. The forecast evaluation period is from 2006Q1 to 2016Q1. In assessing the forecast

performance, the benchmark model is the almost ideal demand system given that it is the model currently used for forecasting the household final consumption components.[3]

Individual Models

Five models are estimated for each individual component of household final consumption expenditure. They are the standard AR(2) models, AR(2) models with linear time trends, relative price models, relative price models with linear time trends and the almost ideal demand system (AIDS). The lag length for the autoregressive models is chosen with a view to modelling the persistence in the data, while maintaining a parsimonious specification. The estimated AIDS does not include total consumption expenditure to ensure that the model is not over-parameterised.

Table 1 reports the root mean squared forecast error (RMSFE) relative to that of the almost ideal demand system for each of the models over different forecast horizons. At the one-quarter-ahead forecasting horizon, both sets of AR(2) models significantly outperform the almost ideal demand system for all household consumption components. These models perform particularly well for the components of food, cigarettes and tobacco, durables, other goods and other services. For example, the RMSFE for other goods under the standard AR(2) model is only 17 per cent of that of the almost ideal demand system. The forecasting gains are smaller for the components of fuels and lubricants and electricity and gas, but they continue to be better than the benchmark.

In the case of the relative price models, the model with the linear time trend generally performs much better than the model without the trend. Further, even for the components where the model with the linear time trend does not outperform the almost ideal demand system – alcohol and fuels and lubricants – the performance is not substantially worse than AIDS. The relative prices model with the linear time trend does particularly well at forecasting other services, with the RMSFE being only 17 per cent of that of the almost ideal demand system. The RMSFE for other services under the relative prices model without a time trend is 57 per cent of that of the almost ideal demand system.

As the forecasting horizon increases, the performance of the AR(2) models for some components deteriorates relative to the AIDS. For example, at the

[3]Formal statistical tests could be performed to assess the statistical significance of the results. However, given the relatively short evaluation period it would be difficult to obtain conclusive results.

three-year-ahead horizon, the forecast for the fuels and lubricants component of the benchmark model is 53 per cent better compared to the standard AR(2) model. In contrast, the relative price model with the linear time trend continues to perform relatively well for most of the components. Consequently, it can generally be concluded that, at shorter forecasting time horizons, models that capture short-run dynamics perform well, but that, at longer horizons, models with trend terms and relative prices tend to perform better.

Forecast Combinations

Table 2 reports the root mean squared forecast error relative to that of the almost ideal demand system for both of the forecast combinations over different forecast horizons. At the one-quarter-ahead forecasting horizon, both forecast combinations perform better than the almost ideal demand system, with the exception of the equal weight model for fuels and lubricants. The forecast combinations perform particularly well for the food and other services components. In the case of other services, the root mean squared forecast error for the equal weight model is only 25 per cent of that of the almost ideal demand system and 19 per cent for the model weighted by forecast performance. This is an important result given that the other services category accounts for a large share of consumption subject to the goods and services tax.

Table 1: Root mean squared forecast error relative to almost ideal demand system (values less than 1 indicate better forecast performance than the benchmark).

	Household consumption components*								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
One-quarter-ahead forecast									
AR(2) model	0.21	0.40	0.23	0.23	0.17	0.38	0.82	0.76	0.21
AR(2) model with trend	0.22	0.45	0.22	0.21	0.16	0.36	0.74	0.71	0.16
Relative price model	0.81	0.85	1.30	1.25	1.43	0.97	3.71	1.11	0.57
Relative price model with trend	0.75	1.08	0.32	0.33	0.45	0.81	1.06	0.80	0.17
One-year-ahead forecast									
AR(2) model	0.37	0.87	0.72	0.29	0.34	0.81	1.77	1.56	0.33
AR(2) model with trend	0.45	1.09	0.70	0.30	0.32	0.78	1.58	1.44	0.12
Relative price model	0.66	0.94	1.16	1.16	1.42	0.96	3.36	1.10	0.44

Relative price model with trend	0.75	1.30	0.38	0.34	0.54	0.83	1.04	0.89	0.14
Two-year-ahead forecast									
AR(2) model	0.26	0.90	0.89	0.27	0.54	1.00	1.83	2.58	0.51
AR(2) model with trend	0.43	1.33	0.89	0.34	0.49	0.95	1.33	2.45	0.17
Relative price model	0.46	0.88	1.10	1.11	1.45	0.96	3.23	1.08	0.33
Relative price model with trend	0.75	1.36	0.40	0.32	0.65	0.83	0.96	1.01	0.11
Three-year-ahead forecast									
AR(2) model	0.24	0.86	1.06	0.31	0.68	1.12	2.14	3.20	0.61
AR(2) model with trend	0.47	1.56	1.17	0.40	0.57	1.06	1.04	3.22	0.17
Relative price model	0.18	0.97	1.07	1.03	1.40	0.96	3.16	1.11	0.15
Relative price model with trend	0.74	1.63	0.78	0.34	0.65	0.85	0.87	1.14	0.10

*****The labelling corresponds with Figure 1: (a) food; (b) alcohol; (c) cigarettes and tobacco; (d) durables; (e) other goods; (f) vehicles; (g) fuels and lubricants; (h) electricity and gas; and (i) other services.

It is also the case that, at the one-quarter-ahead forecasting horizon, the forecast combination based on past forecast performance performs better than the equal weighted model. The performance of the forecast combinations does not, however, outperform the standard AR(2) models or the AR(2) models with linear time trends. This reinforces the conclusion that models that capture short-run dynamics perform well at shorter forecasting time horizons.

Figure 2 shows the model weights for the one-quarter-ahead forecast for the food component. For most periods, each of the models have a non-negligible weight in the forecast combination based on past forecast performance. In other words, all models are contributing to produce the final forecast. In addition, the weights are evolving over time. At the beginning of the forecast period, the AR(2) model and the AR(2) model with a linear time trend perform well and account for most of the weight. Over time the weight of the relative prices model increases, illustrating that the forecast performance of this model improves towards the end of the forecast period. As such, combining the forecasts from all of the models using time-varying weights means that the forecast combination can quickly adapt to changes in model performance.

Table 2: Root mean squared forecast error relative to almost ideal demand system (values less than 1 indicate better forecast performance than the benchmark).

	Household consumption components*								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
One-quarter-ahead forecast									
Forecast combination, equal weights	0.38	0.49	0.53	0.48	0.52	0.66	1.15	0.82	0.25
Forecast combination, SSFE weights	0.22	0.45	0.45	0.24	0.36	0.52	0.91	0.80	0.19
One-year-ahead forecast									
Forecast combination, equal weights	0.45	0.76	0.67	0.53	0.55	0.84	1.34	1.13	0.20
Forecast combination, SSFE weights	0.37	0.82	0.64	0.31	0.43	0.83	1.05	1.03	0.12
Two-year-ahead forecast									
Forecast combination, equal weights	0.44	0.76	0.72	0.55	0.59	0.93	1.27	1.56	0.17
Forecast combination, SSFE weights	0.32	0.74	0.70	0.37	0.56	0.92	0.92	1.19	0.16
Three-year-ahead forecast									
Forecast combination, equal weights	0.44	0.81	0.75	0.56	0.58	0.99	1.30	1.84	0.12
Forecast combination, SSFE weights	0.30	0.65	0.73	0.42	0.61	0.97	0.99	1.29	0.12

*****The labelling corresponds with Figure 1: (a) food; (b) alcohol; (c) cigarettes and tobacco; (d) durables; (e) other goods; (f) vehicles; (g) fuels and lubricants; (h) electricity and gas; and (i) other services.

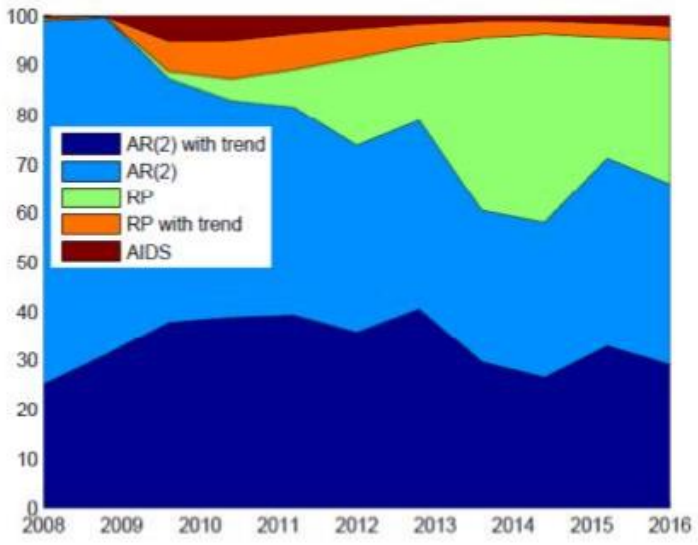


Figure 2: Weights for the one-quarter-ahead forecast for the food component

At the one-year-, two-year- and three-year-ahead forecasting horizons, the forecast combinations also generally perform better than the almost ideal demand system, with fuels and lubricants and electricity and gas being the only components where the forecasting performance is not uniformly better than that of the almost ideal demand system. The forecast combination using past forecasting performance uniformly outperforms the combination based on equal weights at the longer forecasting horizons.

At the longer forecasting horizons, the forecast combinations perform significantly better than the autoregressive models. In contrast, at the three-year-ahead forecasting horizon, the relative prices model with linear time trends performs better than the forecast combination based on past forecasting performance for five out of the nine household consumption components. This shows that models with trend terms and relative prices tend to perform better over longer forecasting horizons, while the autoregressive models are better at forecasting over shorter time horizons.

The varied forecasting performance across the different individual models for the different components of household consumption expenditure and across different forecasting time horizons highlights the benefit of a forecast combination framework. The forecast combination based on forecasting performance takes advantage of models that account for the persistence and longer-term trends in a number of the consumption components, and the shifts caused by relative price changes. Moreover, as a model outperforms its competitors in the recent past, a higher weight is given to that successful model. In this way, the forecast combination

approach quickly adapts to changes in model performance. A forecast combination framework is particularly useful when it is necessary to forecast over a three-year forecasting period.

5. Concluding Remarks

This paper outlines a methodology for forecasting the components of household final consumption expenditure. It uses a forecast combination approach with autoregressive models, regressions on relative prices and the almost ideal demand system developed by Deaton and Muellbauer (1980). The forecast combination that weights the forecasts based on forecasting performance according to rolling squared forecast errors generally performs better than the currently-used almost ideal demand system. The forecast combination takes advantage of the forecasting performance across the different individual models for the different components of consumption expenditure and across different forecasting horizons. The forecast combination is particularly useful when it is necessary to forecast over a three-year forecasting period, given significant differences in forecasting performance of models across different forecasting horizons.

References

1. A. Deaton and J. Muellbauer. An almost ideal demand system. American Economic Review,70(3):312-326, 1980.
2. H. Gerard and K. Nimark. Combining multivariate density forecasts using predictive criteria. RBA Discussion Paper 2008-02, 2008.
3. M. Guidolin and A. Timmermann. Forecasts of US short-term interest rates: A flexible forecast combination approach. Journal of Econometrics, 150(2):297-311, 2009.
4. D.E. Rapach, J.K. Strauss, and G. Zhou. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies, 23(2):821-862, 2010.
5. A. Timmermann. Forecast combination. Handbook of Economic Forecasting, edited by: G. Elliott, C.W.J. Granger and A. Timmermann, North-Holland, 2006.

Bayesian MIDAS penalized regressions: Estimation, selection, and prediction

Matteo Mogliani
Banque de France, International Macroeconomics Division, 31 Rue Croix des Petits Champs, 75049 Paris CEDEX 01, France

Abstract

We propose a new approach to modeling and forecasting with mixed-frequency regressions (MI-DAS) in presence of a large number of predictors. Our approach resorts to penalized regressions such as Group Lasso, allowing for simultaneously selecting the relevant regressors and estimating the non-zero parameters, and Bayesian techniques for estimation. In particular, the penalty hyper-parameters governing the model shrinkage are automatically tuned via an adaptive MCMC algorithm. To achieve sparsity and improve the variable selection ability of the model, we also consider a Group Lasso estimator augmented with a spike-and-slab prior. Simulations show that the proposed models have good in and out of sample performance, even when the design matrix presents high cross-correlation. When applied to a forecasting model of U.S. GDP, the results suggest that high-frequency financial variables may have some, although limited, short-term predictive content.

Keywords

MIDAS regressions; Penalized regressions; Variable selection; Forecasting; Bayesian estimation

1. Introduction

The outstanding increase in the availability of economic data has led econometricians to the development of new regression techniques based on Machine Learning algorithms, such as the family of penalized regressions. This consists in regressions with a modified objective function, such that coefficients estimated close to zero are shrunk to exactly zero, leading to simultaneous selection and estimation of coefficients associated to relevant variables only. While some of these techniques have been successfully applied to multivariate and usually highly parameterized macroeconomic models, such as Vector Autoregressions (Gefang, 2014; Korobilis and Pettenuzzo, in press), only a few contributions in the literature have paid attention to mixed-frequency (MIDAS) regressions. In the classic MIDAS framework (Andreou et al., 2010), the researcher can regress high-frequency variables (e.g. monthly variables such as surveys) directly on low-frequency variables (e.g. quarterly variables such as GDP) by matching the sampling frequency through specific

aggregating (weighting) functions. The inclusion of many high-frequency variables into MIDAS regressions may nevertheless lead to overparameterized models, with poor predictive performance. This happens because the MIDAS regression approach can efficiently address the dimensionality issue arising from the number of high-frequency lags in the model, but not that arising from the number of high-variables. Hence, recent literature has focused on MIDAS penalized regressions, based mainly on the so-called Lasso and Elastic-Net penalizations (Marsilli, 2014; Siliverstovs, 2017; Uematsu and Tanaka, in press).

In the present paper, we propose a similar approach, but we depart from the existing literature on several points. First, we consider MIDAS regressions resorting to Almon lag polynomial weighting schemes, which depend only on a bunch of functional parameters governing the shape of the weighting function and keep linearity in the regression model. Second, we consider a Group Lasso penalty, which operates on distinct groups of regressors, and we set as many groups as the number of high-frequency predictors, allowing each group to include the entire Almon lag polynomial of each predictor. This grouping structure is motivated by the fact that if one high-frequency predictor is irrelevant, it should be expected that zero-coefficients occur in all the parameters of its lag polynomial. Third, we implement Bayesian techniques for the estimation of our penalized MIDAS regressions. The Bayesian approach offers two attractive features in our framework. The first one is the inclusion of spike-and-slab priors that, combined with the penalized likelihood approach, aim at improving the selection ability of the model by adding a probabilistic recovery layer to the hierarchy. The second one is the estimation of the penalty hyper-parameters through an automatic and data-driven approach that does not resort to extremely time-consuming pilot runs. In this paper we consider an algorithm based on stochastic approximations, which consists in approximating the steps necessary to estimate the hyper-parameters in such a way that simple analytic solutions can be used. It turns out that penalty hyper-parameters can be automatically tuned with a small computational effort compared to existing and very popular alternative algorithms.

2. Methodology

Consider the variable y_t, which is observed at discrete times (*i.e.* only once between $t-1$ and t), and suppose that we want to use information stemming from a set of K predictors $x_t^{(m)} = x_{1,t}^{(m)}, \dots, x_{K,t}^{(m)})'$, which are observed m times between $t-1$ and t, for forecasting purposes. The variables y_t and $x_{k,t}^{(m)}$, for $k = 1, \dots, K$ are said to be sampled at different frequencies. For instance, quarterly and monthly frequencies, respectively, in which case $m = 3$. Let us define the high-frequency lag operator $L^{1/m}$, such that $L^{1/m} x_{k,t}^{(m)} = x_{k,t-1/m}^{(m)}$.

Further, let $h = 0, 1/m, 2/m, 3/m, \ldots$ be an (arbitrary) forecast horizon, where $h = 0$ denotes a nowcast with high-frequency information fully matching the low-frequency sample. The MIDAS approach plugs-in the high-frequency lagged structure of predictors $x_{k,t-h}^{(m)}$ in a regression model for the low-frequency response variable y_t as follows:

$$y_t = \alpha + \sum_{K=1}^{K} \beta(L^{1/m}; \theta_k x_{k,t-h}^{(m)} + \varepsilon_t, \quad (1)$$

where ε_t is i.i.d. with mean zero and variance $\sigma^2 < \infty$, and $\beta\left(L^{1/m};\ \theta_k\right) = \sum_{c=0}^{C-1} B(c;\ \theta_k) L^{c/m}$ is a weighting structure which depends on the weighting function $B(c; \theta_k)$, a vector of p + 1 parameters $\theta_k = (\theta_{k,0}, \theta_k, 1, \ldots, \theta_{k,p})$, and a maximum lag length C. In this study, we consider the simply polynomial approximation of $\beta\left(L^{1/m};\ \theta_k\right)$ provided by the Almon lag polynomial $B(c; \theta_k) = \sum_{i=0}^{p} \theta_{k,i} c^i$. Under the so-called "direct method", Equation (1) with Almon lag polynomials can be reparameterized as:

$$y_t = \alpha + \boldsymbol{\theta_1} \boldsymbol{z}_{1,t-h}^{(m)} + \cdots + \boldsymbol{\theta}_k \boldsymbol{z}_{K,t-h}^{(m)} + \varepsilon_t \quad (2)$$

where $\boldsymbol{z}_{k,t}^{(m)}, k = 1, \ldots, K$, is a vector of linear combinations of the observed high-frequency regressors, $\boldsymbol{z}_{k,t}^{(m)} = \boldsymbol{Q} \boldsymbol{x}_{k,t}^{(m)}$, with $\boldsymbol{x}_{k,t}^{(m)}$ a $(C \times 1)$ vector of high-frequency lags and $\boldsymbol{Q}$ a $(p + 1 \times C)$ polynomial weighting matrix. The main advantage of the Almon lag polynomials is that (2) is linear and parsimonious, as it depends only on $K(p + 1)$ parameters and can be estimated consistently and efficiently via standard methods.

Although appealing, the MIDAS regression in (2) may be easily affected by over-parameterization and multicollinearity in presence of a large number of potentially correlated predictors. To achieve variable selection and parameter estimation simultaneously, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (Lasso). In a nutshell, the Lasso is a penalized least squares procedure, in which the loss function $\mathcal{L}_T(\boldsymbol{\theta})$ is minimized after setting a constraint on the ℓ_1 norm of the vector of regression coefficients, where the amount of penalization is controlled by a parameter λ. To achieve the *oracle property*, which guarantees that the estimator performs as well as if the true model had been revealed to the researcher in advance by an oracle, Zou (2006) proposed the Adaptive Lasso (AL), where a different amount of shrinkage (i.e. a different penalty term) is used for each individual regression coefficient. However, the AL may not be suited in the present framework, as lags of high-frequency predictors are by construction highly correlated and hence the Lasso estimator would tend to select randomly only one lag and shrink the remaining polynomial coefficients to zero. The theoretical rationale for a failure in the selection ability of the AL in our mixed-frequency setting is

similar to that pointed out by Zou and Hastie (2005), and it is mostly related to the lack of strict convexity in the Lasso penalty. To address this issue, we propose a solution based on the Adaptive Group Lasso (AGL) estimator (Wang and Leng, 2008). This approach introduces a penalty to a group of regressors, rather than a single regressor, that may lead (if the group structure is carefully set by the researcher) to a finite sample improvement of the AL. In the present framework, it seems reasonable to define a group as each of the k vectors of lag polynomials in the model. This grouping structure is motivated by the fact that if one high-frequency predictor is irrelevant, it should be expected that zero-coefficients occur in all the parameters of its lag polynomial. This strategy should overcome, at least in part, the limitation of the Lasso in presence of strong correlation in the design matrix arising from the correlation among lags of the transformed high-frequency predictors.

Several approaches have been proposed in the literature to estimate penalized regressions. In this paper, we consider a Bayesian hierarchical approach. We then introduce the Bayesian MIDAS Adaptive Group Lasso model (BMIDAS-AGL), based on the Bayesian Group Lasso prior of Kyung et al. (2010), where the conditional prior of $\boldsymbol{\theta}$ can be expressed as a scale mixture of Normals with Gamma hyper-priors. However, an expected feature of this model is that a sparse solution cannot be perfectly achieved, as the Bayesian approach provides a shrinkage of the coefficients towards zero, but usually not exactly to zero. Recent literature has increasingly focused on combining the potential advantages of spike-and-slab methods and the penalized likelihood approach (Roëkováand George, 2018). In the present study, we follow Xu and Ghosh (2015) and we introduce the Bayesian MIDAS Adaptive Group Lasso with spike-and-slab priors (BMIDAS-AGL-SS). This prior provides two shrinkage effects: the point mass at $\mathbf{0}$ (the spike part of the prior), which leads to exact zero coefficients, and the Group Lasso prior on the slab part.

Our hierarchical models treat the penalty parameters $\boldsymbol{\lambda}$ as hyper-parameters, i.e. random variables with gamma prior distributions and gamma posterior distributions. However, the main drawback of this approach is that these posterior distributions can be sensitive to the choice of the prior. An alternative approach resorts to an Empirical Bayes estimation of the hyper-parameters, i.e. using the data to propose an estimate of $\boldsymbol{\lambda}$, which can be obtained through marginal maximum likelihood. For this purpose, a usual choice is the Monte Carlo EM algorithm (MCEM), which complements the Gibbs sampler and provides marginal maximum likelihood estimates of the hyper-parameters. From a computational point of view, the MCEM algorithm may be extremely expensive, as each nth Monte Carlo iteration requires a fully converged Gibbs sampling from the posterior distribution of the parameters. In the present framework, careful attention must be paid to this point, because the computational burden implied by the Group Lasso increases dramatically

as the number of predictors increases. To deal with this issue, in this work we adopt an alternative Empirical Bayes approach that relies on stochastic approximation algorithms to solve maximization problems when the likelihood function is intractable, by mimicking standard iterative methods such as the gradient algorithm. This approach is computationally efficient, because it requires only a single Monte Carlo run. Using a stochastic approximation to solve the maximization problem, we get an approximate EM algorithm, where both E- and M-steps are approximately implemented. Hence, marginal maximum likelihood estimates of the hyper-parameters and draws from the posterior distribution of the parameters are both obtained using a single run of the Gibbs sampler.

3. Results

We evaluate the performance of the proposed models through Monte Carlo experiments. For this purpose, we use a DGP similar to Equation (1) and involving K = {30, 50} predictors sampled at frequency $m = 3$ and $T = 200$ in-sample observations. The predictors follow all the same stationary AR(1) process, but only five are relevant in the model. As for the weighting function B (c; ϑ), we choose an exponential Almon lag function. We investigate three alternative weighting schemes that correspond to fast-decaying weights, slow-decaying weights, and near-flat weights. For ease of analysis we assume $h = 0$. In this specification, the error terms are assumed i.i.d. normally distributed, but the design matrix is allowed to present moderate to extremely high correlation structure.

We compute the average mean squared error (MSE), the average variance (VAR), and the average squared bias (BIAS2) over R Monte Carlo replications. Further, we evaluate the selection ability of the models by computing the True Positive Rate (TPR), the False Positive Rate (FPR), and the Matthews correlation coefficient (MCC). Simulation results point to a number of interesting features. First, the models perform overall quite similarly in terms of MSE, although the BMIDAS-AGL-SS seems to perform somewhat better across DGPs by mainly providing the smallest bias. This leads to highest TPR and lowest FPR for this model, entailing better classification of the active and inactive sets across simulations. Second, the MSE increases substantially with the degree of correlation in the design matrix, but it tends to decrease with more irrelevant predictors. It follows that the performance of the models in selecting and estimating the coefficients of the relevant variables holds the same regardless the increase in the degree of sparsity. This result is confirmed by the TPR, which is relatively high and hovers around 80-90% for moderate correlation, and it's overall stable across the different values of K, suggesting that the models can select the correct sparsity pattern with a high probability even in finite samples. However, it is worth noting that the TPR drops to 30-50% for very

high correlation, while the the FPR remains overall very low. This result is nevertheless not unexpected, as the Group Lasso can address the issue of strong collinearity within the lag polynomials but is not designed to handle strong collinearity between the high-frequency regressors. Finally, looking at the forecasting performance, the results are broadly in line with the in-sample analysis and suggest that the models perform overall quite similarly in terms of point and density forecasts, although the BMIDAS-AGL-SS model seems to perform best overall. The performance of the models deteriorates substantially with higher correlation in the design matrix, but it is relatively stable with K increasing.

We apply the proposed Bayesian MIDAS penalized regressions to US quarterly GDP data. We consider 42 real and financial indicators, sampled at monthly, weekly, and daily frequencies. The data sample starts in 1980Q1, and we set 2000Q1 and 2017Q4 the first and last out-of-sample observations, respectively. Estimates are carried-out recursively using an expanding window, and h-step-ahead posterior predictive densities $(h = 0,1,4)$ are generated through a direct forecast approach. Forecasts are compared to those from a benchmark model represented by a simple random-walk (RW). Point and density forecasts are evaluated by the means of standard criteria. As a robustness check, we further consider forecasts from alternative competing models, such as the AR(1), the combination of K single-indicator Bayesian MIDAS models, the Bayesian model selection (BMS), and the Bayesian model averaging (BMA). Our findings suggest that the penalized BMIDAS models outperform the benchmark RW at all the horizons, whether point or density forecasts are considered. When compared to the set of alternative models, our penalized BMIDAS models display predictive gains at $h = 0$ and $h = 1$. At $h = 4$, the predictive performance of most of the alternative competing models is only slightly superior or inferior to that of our penalized regressions. Looking at the selection of predictors over the pseudo out-of-sample, the results point to systematic inclusion of a few real high-frequency indicators. Further, selection appears more parsimonious and stable for the BMIDAS-AGL-SS model. Virtually no financial indicators are selected by our models at $h = 0$, with real hard- and soft-data conveying all the relevant information. However, this feature tends to attenuate for $h = 1$, where some high-frequency financial indicators are selected. All in all, this result is broadly in line with recent literature (Andreou et al., 2013) and suggests that financial variables may convey some, although limited, short-term leading information which goes beyond the predictive content of real indicators.

4. Discussion and Conclusion

We proposed a new approach to modeling and forecasting mixed-frequency regressions (MIDAS) that addresses the issue of simultaneously estimating and selecting relevant high-frequency predictors in a high-dimensional environment. Our approach is based on MIDAS regressions resorting to Almon lag polynomials and an adaptive penalized regression approach, namely the Group Lasso objective function. The proposed models rely on Bayesian techniques for estimation and inference. In particular, the penalty hyper-parameters driving the model shrinkage are automatically tuned via an Empirical Bayes algorithm based on stochastic approximations. Simulations show that the proposed models present very good in-sample and out-of-sample performance. When applied to a forecasting model of U.S. GDP with high-frequency real and financial predictors, the results suggest that our models produce significant out-of-sample short-term predictive gains compared to several alternative models. Further, our findings are broadly in line with the existing literature, in the extent that high-frequency financial variables have non-zero, although limited, short-term predictive content.

References

1. Andreou, E., Ghysels, E., Kourtellos, A., 2010. Regression models with mixed sampling frequencies. Journal of Econometrics 158 (2), 246–261.
2. Andreou, E., Ghysels, E., Kourtellos, A., 2013. Should macroeconomic forecasters use daily financial data and how? Journal of Business & Economic Statistics 31 (2), 240–251.
3. Gefang, D., 2014. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. International Journal of Forecasting 30 (1), 1–11.
4. Korobilis, D., Pettenuzzo, D., in press. Adaptive hierarchical priors for high-dimensional vector autoregressions. Journal of Econometrics.
5. Kyung, M., Gill, J., Ghosh, M., Casella, G., 2010. Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis 5 (2), 369–412.
6. Marsilli, C., 2014. Variable selection in predictive MIDAS models. Working Paper 520, Banque de France.
7. Park, T., Casella, G., 2008. The bayesian lasso. Journal of the American Statistical Association 103 (482), 681–686.
8. Roˇcková, V., George, E. I., 2018. The spike-and-slab LASSO. Journal of the American Statistical Association 113 (521), 431–444.
9. Siliverstovs, B., 2017. Short-term forecasting with mixed-frequency data: a MIDASSO approach. Applied Economics 49 (13), 1326–1343.
10. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 (1), 267–288.

11. Uematsu, Y., Tanaka, S., in press. High-dimensional macroeconomic forecasting and variable selection via penalized regression. The Econometrics Journal.
12. Wang, H., Leng, C., 2008. A note on adaptive group lasso. Computational Statistics & Data Analysis 52 (12), 5277–5286.
13. Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for group lasso. Bayesian Analysis 10 (4), 909–936.
14. Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.
15. Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301–320.

Combination nowcasts of advance estimates of private consumption of services in the U.S National Accounts

Baoline Chen[*†], Kyle Hood[*]
University of the Philippines

Abstract
This paper evaluates combination forecast methods for nowcasting advance quarterly estimates of private consumption (PCE) of services in the U.S. national accounts using data from 2009 to 2018. In a previous study, we showed that both the bridge-equation and bridging-with-factors frameworks could improve the accuracy of advance estimates of detailed PCE services components by reducing revisions when quarterly data become available. However, degrees of reduction in revisions vary over time and across PCE services categories. Studies have shown that despite unstable performances of individual nowcasting or forecasting models, combination forecasts often improve upon individual nowcasts or forecasts. In this study, we evaluate alternative methods to combine nowcasts from general-bridge-equation (GB) and bridging-with-factors (BF) frameworks. We consider weights for combination based on simple averaging (mean and median), information-criterion averaging, and Bates-Granger averaging with leave-one-out cross-validation (LOOCV) errors. We evaluate the performances of combined forecasts by comparing their root mean squared revisions (RMSR).

Keywords
Nowcasting; forecast combination; model averaging; national accounts

1. Introduction

In the United States, Gross Domestic Product (GDP) is released three times each quarter. The first "advance" estimate is made approximately one month after the quarter's end and is based on the most limited source data, while two more estimates, the "second" and "third," are based on more-detailed or less-preliminary source data. Revisions in GDP result from using these new or revised source data and can be rather large for some of its components. Quarterly GDP estimates are built up from detailed components of the major subaggregates

[*] Bureau of Economic Analysis, Washington, DC. The views expressed herein are those of the authors and do not reflect the views of the Bureau of Economic Analysis or the Department of Commerce.
[†] Corresponding author. Email: baoline.chen@bea.gov.

(consumption, investment, government expenditures, exports and imports), with personal consumption expenditures on services (PCE services) showing particularly large revisions due to lack of availability of source data for the first estimate in most detailed components. In this paper, we focus on these detailed components, using a combination of nowcasting and model-averaging techniques to reduce the size of revisions.

A lack of source data for the first estimate of most PCE services components means that Bureau of Economic Analysis (BEA) must use other indicators for these components in the first estimate. These indicators are chosen to match as well as possible the component being considered. For some estimates, the indicator is growth in population and a component-specific price index, while for other estimates, the indicator is wage and salary growth in the industry associated with the component. Unfortunately, these indicators are often weakly or even negatively correlated with the third-estimate values of components which use the Quarterly Services Survey (QSS) as an indicator for the third estimate. For example, for health care services, correlations between the indicator and third estimate are negative for 4 of 15 detailed components and are above 0.5 in only 2 of 15 cases. Similar patterns hold for other categories.

Lack of a close correlation between indicators and the target series can be alleviated by the addition of other useful data. The indicators used for the first estimate do not incorporate additional information that is available at the time of the first estimate. For example, medium-term movements in the target series could be relevant, as well as long-term trends. In addition, information from more general movements in PCE services could be relevant, but the current method restricts indicators to those matching the specific components.

Because in this paper we are interested in estimating the recent past where data have not become available yet, we exploit the nowcasting literature. Nowcasting is defined as the prediction of the present, the very near future and the very recent past (Giannone, Reichlin and Small, 2008). It is different from forecasting in some specific data-related problems that must be overcome. In this case, we have access to monthly indicators for our quarterly series and we must deal with the fact that our indicators are preliminary data which are likely to be subsequently revised. Other issues specific to nowcasting such as the "ragged edge" problem do not arise (cf. Giannone, Reichlin and Small, 2008). For a survey of relevant techniques, see Foroni and Marcellino (2013).

In this paper, we expand work from an earlier paper (Chen and Hood, 2018) in which we showed that two nowcasting techniques, the General Bridge Equation (GB) (Klein and Sojo, 1989) and Bridging with Factors (BF) models (Giannone, Reichlin and Small, 2008), reduce revisions for many PCE services detailed components. Here, we combine these two methods (in an array of

specifications) using a set of model-averaging algorithms. We have two goals: The first is to identify specific methods that can be used to accurately impute individual detailed PCE services components for the advance GDP estimate, while the second is to look for patterns in the revisions implied by the different classes of model-averaging techniques. These patterns provide useful evidence for the type of model-averaging techniques that are appropriate in similar situations.

The two nowcasting models that we use differ not so much in form, but in the type of information used. In the GB model, quarterly indicators are derived from monthly source data. However, instead of the current method which uses the growth rate of the quarterly indicator as the estimated growth rate, we allow for a long-term trend, lags of the dependent variable, and lags of the indicator. The second model that we use, BF, discards the indicator constructed in this way, using rather a common factor derived from all indicators from the GB model. This collection of indicators yields two factors that appear to accurately capture the movement of many of the detailed PCE services component series.

The reason that model averaging is chosen to combine this information rather than using the more traditional technique of augmenting the model with additional data is that in this application we have only about 34 time periods. The short time-series dimension that we are working with is not compatible with a large number of right-hand-side variables. Model-averaging techniques are designed in part to combat this issue.

We estimate and average 12 versions (6 GB and 6 BF) of the models which differ by how many lags of the dependent and independent variables are included. Five model-averaging techniques are then considered: Two simple techniques (equally-weighted average and median), two information-criteria-based (IC-based) averaging techniques, and Bates-Granger (BG) averaging with leave-one-out cross-validation (LOO-CV). All models and averages are computed on an estimation sample (sometimes called a "training sample") and compared using pseudo-out-of-sample data from the end of the period.

2. Methodology

This section contains a more detailed description of methods. We start by discussing the GB and BF models, then discuss the model-averaging techniques, and finally we detail which GB and BF specifications are averaged using these techniques and describe other details of the algorithm.

Nowcasting models

We consider two types of nowcasting models. In the GB class of models, one or more indicators, lags of these indicators, and lags of the dependent variable are used to nowcast the target variable in a regression framework. Indicators are typically available at a higher frequency than the target variable,

and so in the literature sometimes do not cover the entire period of time being nowcasted. In this case, however, we have access to all three months of data that cover our sample period. We convert these data to a quarterly value and use the quarterly data in the GB model. The regression specification for a single detailed component is

$$y_t = c + \sum_{i=1}^{p} \beta_i y_{t-1} + \sum_{j=1}^{s} \sum_{i=0}^{k} \alpha_{j,i} \bar{x}_{j,t-i} + \varepsilon_t. \quad (1)$$

Here, t is time (quarter) and $j(1, \ldots,)$ indexes indicators if there are more than one. y_t is the dependent variable (the nowcast target), while $\bar{x}_{j,t-1}$ represents the quarterly average of a monthly indicator, and ϵ_t is the error term. $\alpha_{j,i}$ and β_i are collections of parameters on the lag polynomials, and c is a constant. Note that this regression is estimated separately for each detailed component (i.e., there is no pooling of observations across cross-sections). p and k are the number of lags used for the dependent variable and for the indicators and are assumed to be equal to each other in most of the specifications below. Because of its use of indicators, we sometimes refer to this as an "indicator model."

Our BF model is very similar in its basic specification to the GB model, except that the indicators (x's) in Equation (1) are replaced with common factors, denoted with an f (and of which there are r), namely,

$$y_t = c + \sum_{i=1}^{p} \beta_i y_{t-1} + \sum_{j=1}^{s} \sum_{i=0}^{r} \bar{f}_{j,i} \bar{x}_{j,t-i} + \varepsilon_t. \quad (2)$$

These common factors, meant to capture movements in the entire collection of (monthly) indicators, are assumed to be related to the (monthly) indicators *via*

$$x_m = \Lambda f_m + \zeta_m \quad (3)$$

Note that we have omitted the bars above these variables (which indicate quarterly averages of monthly variables) and replaced the time subscript with an $m \in \{1, \ldots, M\}$, indicating that these variables are expressed at a monthly frequency. Here, the Λf_m is known as the "common component" of x_m while the ζ_m represents the "idiosyncratic component" of x_m. Λ is an $n \times r$ matrix of "factor loadings" which is constant over time (and where n is the number of indicators) and f_m is an $r \times 1$ vector of common factors for month m. The factor model is estimated *via* principal components analysis (PCA), which is appropriate if any cross-sectional dependence in x_m coming from the idiosyncratic component (ζ_m) is transitory. The PCA procedure produces a collection of $min\{M,n\}$ mutually-orthogonal monthly factors which can be fed into Equation (2). The Bai-Ng criterion (Bai and Ng, 2002) is used to select the number of factors that

are used in estimation. We also note that while a general version of the bridging-with-factors model can be found in Giannone, Reichlin and Small (2008), our model is simplified in that it makes no assumptions on how factors evolve over time—assumptions that tend to restrain the behavior of the factors to some extent.

Model-averaging algorithms

Five procedures are selected to average among specifications of the above GB and BF models. We define three categories of model averaging: simple averaging, information-criterion-based averaging, and Bates-Granger (BG) averaging with LOOCV.[1] We focus narrowly on averaging the nowcasts associated with each model, but because the models are linear, this is equivalent to averaging the parameters.

In the first category, simple averaging, models are averaged using either equally weighted means or medians of the models, giving us a total of two simple averaging techniques. Simple averaging is typically optimal when short samples hamper precise estimation of the weights (Smith and Wallis, 2009).

In the second, category, IC-based averaging, weights are defined to be proportional to the exponential of the negative of an information criterion, in general,

$$\mathcal{W}_h^{IC} = \frac{\exp\{-IC_h\}}{\sum_{h'=1}^{H}\exp\{-1C_{h'}\}}. \tag{4}$$

Here, *h* indexes the model, ranging from *1* to *H*. IC_h is an information criterion associated with the estimated model *h*. In this paper, we use the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for this purpose, for a total to two IC-based averaging techniques.

The third model-averaging method, BG averaging (Bates and Granger, 1969), has weights that are proportional to the inverse of the sample forecast variance of model *h*, denoted $\hat{\sigma}_h^2$,

$$\mathcal{W}_h^{BG} = \frac{1/\hat{\sigma}_h^2}{\sum_{h'=1}^{H} 1/\hat{\sigma}_{h'}^2}. \tag{5}$$

To ensure that we have a "clean" pseudo-out-of-sample subset with which to compare nowcasts, the inverse forecast variances are computed in-sample using LOO-CV. The LOO-CV algorithm iterates over the in-sample observations, leaving each observation out once. An error is computed for each observation that was left out based on parameters estimated from the

[1] Each of these techniques is discussed by Diks and Vrugt (2017). Our Bates-Granger technique differs slightly in that we are using the leave-one-out cross-validation errors, rather than in-sample residuals.

rest of the sample. Thus, if there are T in-sample observations, the model is computed T times on T -1 observations, and T leave-one-out errors are computed. These errors are then used to compute a sample forecast variance.

Model specifications

We have selected 12 model specifications to be supplied to each of the model-averaging algorithms. These are further grouped into six indicator model specifications and six factor model specifications. Within each grouping, we consider model specifications with lags from 0 (only the contemporaneous indicator or factor) to 4 (including the contemporaneous indicator or factor, lags of the indicator or factor up to and including lag 4, and lags of the third estimate up to and including lag 4). This yields 5 model specifications for each grouping. In addition, we consider the most general model specification in each grouping (the 4-lag model), selecting the best submodel using the small-sample-corrected Akaike Information Criterion (AICC). This provides two additional model specifications, for a total of 12 = 2(5 + 1). Table 1 shows a summary of the number of parameters to be estimated for each of these model specifications.

Other modeling details

To provide an unbiased picture of the improvements in revision performance anticipated for any of the models and model-averaging algorithms, the sample is split into an estimation sample and a test sample. All model parameters and model-averaging weights are computed only on the estimation sample. To compute the complete set of revisions for the test sample, we estimate the models and model-averaging weights on a "rolling" basis, meaning that for each period in the test sample, model parameters and weights are recomputed using all observations prior to the period under consideration (this is also often called "recursive" estimation in the literature).

Table 1. Number of model parameters by specification

Lag	Indicator model	Factor model
(s)	$(c, x_t, x_{t-s}, y_{t-s})$	$(c, f_{1,t}, f_{2,t}, f_{1,t-s}, f_{2,t-s}, .., y_{t-s})$
0	2	1+r[1]
1	4	2+2r
2	6	3+3r
3	8	4+4r
4	10	5+5r
"best" model (AICC)	≤10	≤5+5r

Notes
1. r is the number of factors.

3. Results

There is not space in this paper for a full accounting of the results of the estimation exercise, and so we will provide an overview. We start by discussing

factor number selection, move on to a discussion of revisions by PCE services group, and finally discuss revisions by algorithm type.

Because of the number of parameters to be estimated in the richest lag structures, we set an upper limit of 2 common factors. The Bai-Ng criterion selects 2 as the number of factors, so in all cases, 2 factors are considered.
Table 2 provides a summary of the estimation results. This summary is focused on revision reductions for the individual PCE component series, grouped by algorithm and PCE services grouping. Columns (2) through (9) represent counts of detailed components for which the algorithm performed the best. Columns (3) through (7) show the number of times a model-averaging algorithm was optimal, and columns (8) and (9) show counts of components for which single indicator (GB) and factor (BF) models chosen by AICC were optimal.

For 16 of the 85 series that were nowcasted, no improvement was seen in the out-of-sample validation set relative to the current method. In the remaining 69 series, at least one of the methods showed some improvement over current methods. Improvements were relatively evenly spaced out over the PCE services component groups. Healthcare showed improvement in 15 of 20 series, recreation in 14 of 16 series, communications in 4 of 6 series, professional services in all 5 series, travel and transportation in all 14 series, personal services in 5 of 7 series, and social services in 12 of 17 series. The column on the far right shows the average relative revision (in root mean square terms), compared to the current method. The best improvement in a components series was in motor vehicle rental (in the TRSFTR group), which showed a revision reduction of about 75%. Average reductions by grouping ranged from about half (PRS) to 12.6% (RCA), with an overall average of 26.8%.

As noted above, the forecast combination puzzle asserts that empirically, simple averages often perform better than other methods that derive weights from model performance (Smith and Wallis, 2009). This holds for these results. For 34 of the series (nearly half of the 69 for which a reduction in revision was achieved), simple averages (means or medians) of the 12 forecasts were the best-performing algorithms. For 26 series, the best model was a single model within either the indicator or factor models selected by AICC. Among these, the indicator models was selected more than three times as often as the factor model. In only 9 cases were the other 3 methods optimal, with the AIC-based and the Bates-Granger methods accounting for 8 of the 9.

Table 2. Counts of best model average or selection by PCE services group

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
PCE	No	IC-based averaging		Bates-Granger	Simple averaging		Model selection			Average[3]
Group[1]	Improvement[2]	AIC	BIC	averaging	Equal weights	Median	Indicator	Factor	Total	Relative RMSR
HLC	5	1	0	1	1	6	6	0	20	67.3%
RCA	2	1	0	2	2	1	7	1	16	87.4%
COM	2	0	0	0	1	2	0	1	6	74.2%
PRS	0	1	0	1	1	0	2	0	5	48.3%
TRSFTR	0	1	1	0	5	4	2	1	14	73.4%
PER	2	0	0	0	1	2	2	0	7	74.7%
SOC	5	0	0	0	4	4	1	3	17	73.3%
Total	16	4	1	4	15	19	20	6	85	Overall average
Percentage	18.8%	4.7%	1.2%	4.7%	17.6%	22.4%	23.5%	7.1%	100.0%	73.20%

Notes:
1. HLC: Health care, RCA: Recreation, COM: Communications, PRS: Professional services, TRSFTR: Travel/transportation, PER: Personal services, SOC: Social services
2. Unweighted totals by component within each group (including only components with differing indicators for first and third estimate)
3. Unweighted mean

4. Conclusion

Model averaging has the potential to be a powerful tool in reducing revisions in national economic accounts statistics at the detailed-component level. For some detailed components, these techniques can reduce revisions by half or more. However, because of short time series, the forecast combination puzzle is especially relevant. Simple averages of the models under consideration performed best nearly half of the time. Nevertheless, it does appear that other model-averaging techniques may perform well under certain circumstances. Furthermore, the GB model seems to frequently outperform the BF model when a single model is optimal, suggesting that many of the indicators already used to produce these PCE component estimates contain relevant and useful information on the movements of the underlying series. Over all, it is clear that there is not broad agreement on a single approach or algorithm that can be applied across all of these components. As such, a hybrid approach that uses a specific algorithm for each detailed component may offer the best option.

References

1. Bai, J. and S. Ng, "Determining the number of factors in approximate factor models," Econometrica, 70(1): 191–221.

2. Bates, J. M. and C. W. J. Granger (1969), "The combination of forecasts," Journal of the Operational Research Society, 20(4): 451–468.
3. Diks, C. G. H. and J. A. Vrugt, "Comparison of point forecast accuracy of model averaging methods in hydrologic applications," Stochastic Environmental Research and Risk Assessment, 24(6): 809–820.
4. Foroni, C. and M. J. Marcellino (2013), "A survey of econometric methods for mixed-frequency data," Norges Bank Research Working Paper: 2013–6.
5. Giannone, D., L. Reichlin, and D. Small (2008), "Nowcasting: The real-time informational content of macroeconomic data," Journal of Monetary Economics, 55(4): 665–676.
6. Klein, L. R. and E. Sojo (1989), "Combinations of high and low frequency data in macroeconometric models," pp. 1–13 in Economics in Theory and Practice: An Eclectic Approach, L.R. Klein and Marquez (eds.), Kluwer Academic Publisher. Smith, J. and K. F. Wallis (2009), "A simple explanation of the forecast combination puzzle," Oxford Bulletin of Economics and Statistics, 71(3): 331–355.

Empirical bayes method for modelling of air pollution index

Yousif Alyousifi, Nurulkamal Masseran, Kamarulzaman Ibrahim
Universiti Kebangsaan Malaysia

Abstract

Air pollution is becoming a problem of concern in many parts of the world nowadays. Monitoring the level of air pollution by using air pollution index (API) is commonly practiced in Malaysia. In this study, an empirical Bayes method is applied for estimating the parameters in the Markov transition probability matrix for describing the stochastic behaviour of API data. The study reported in this paper is conducted based on the hourly data collected from the central region in Malaysia for a period of 3 years. The results describe the experience of air pollution for the region whereby the risk of occurrences for unhealthy events is small; however, some areas experienced a longer unhealthy condition and also with a higher probability as compared to the other areas.

Keywords

Empirical Bayes; Markov Chain Modeling; Air Pollution Index

1. Introduction

The problem of air pollution in Malaysia is an important topic that has attracted the concern of many researchers (Azid et al.2014; Latif et al 2014; Masseran et al. 2016; Alyousifi et al 2017; Al-Dhurafi, et al 2018). The urban and industrial areas in Malaysia are considered to be the most affected due to the presence of high density of traffic and manufacturing industries (Azid et al.2014). Based on the studies by (Latif et al. 2014), traffic is known to be one of the major sources of air pollution in the urban areas for most developing countries, including Malaysia. In addition, the open burning and forest fires that often occurred in the neighbouring countries such as Indonesia could be one source of air pollution in Malaysia.

In Malaysia, since 1998 the Department of Environment (DOE) has adopted the Air Pollution Index (API) as an indicator of air quality, for providing the public with information on the quality of air in the environment (DOE 2000; Masseran et al. 2016). The API value is described by the Department of Environment as a simple measure for describing the state of air quality in the environment and providing easily understood information about air pollution. It is determined based on the maximum value of five sub-indices of pollutants, namely, particle matter (PM10), sulphur dioxide (SO2), carbon monoxide (CO),

nitrogen dioxide (NO2) and ozone (O3) (Gass et al. 2015). The US model is used to determine the API data in Malaysian where the concentration of each pollutant is transformed into a numerical scale which ranged between 0 and infinity (DOE 2000). The API value of less than 100 indicates a moderate air quality while the API value that is more than 100 shows a higher level of air pollution.

More specifically, several researchers have also utilized Markov chain in their studies on air pollution. For example, Hoyos et al. (2010) have proposed a finite Markov chain for modeling the levels of air pollution in Mexico City for the purpose of evaluating how far the control policies on air pollution are effective. Rodrigues & Achcar (2012) employed a Markov chain model on ozone air pollution using the daily maximum measurements. They applied the Bayesian method to estimate the parameters of the first order Markov Matrix and found that the Markov model to be adequate to explain the classification of the different level of air pollution. Alyousifi et al. (2017) have employed a Markov chain model to predict and investigate the occurrences of air pollution in Peninsular Malaysia. They have estimated the count matrix using the maximum likelihood method. Although, the Markov model appear to be quite flexible in representing the transitions between the different air pollution states, the resultant Markov matrix is found to include some zero probabilities, indicating no possibility of going from certain state to another. It has been argued by the authors that the results are not surprisingly due to the persistent occurrences of dry months during the period of the study. Meanwhile, the empirical Bayes approach has been suggested by several authors such as (Fienberg & Holland 1973; Meshkani & Billard 1992; Rodrigues & Achcar 2012; Seal & Hossain 2013,2015; Sanusi et al.2013; Agresti & Chuang 1989), which offers a solution for addressing the problem of zero probabilities in the transition matrix of the Markov chain found based on maximum likelihood method. In this study, the empirical Bayesian method is applied for smoothing the zero probability in the transition matrix and to estimate the parameters of Markov chain model based on hourly API data to describe air pollution characteristics of seven cities located in the central region of Peninsular Malaysia. For each particular state of air pollution, the characteristics of interest are air pollution persistence, probability of air pollution and air pollution duration, which will be obtained by determining the mean residence time, the steady-state probability and the mean recurrence time of the air pollution respectively.

2. Study Area and Dataset

2.1 Study Areas

In this study, the hourly data for the period of three years, i.e. 2012 to 2014, is collected from seven air-monitoring stations which are located in areas in the central region of Peninsular Malaysia, as provided by the Department of Environment (DOE). Areas considered for the study are Kuala Lumpur, Klang, Cheras, Petaling Jaya, Banting, Shah Alam and Kuala Selangor. Modelling the behaviour of air pollution and determine the risk of unhealthy states of air pollution are the main interest of this study. Figure 1 shows the location of seven areas involved in this study.

The dataset used in this study included the hourly API data from 1st January 2012 until 31st December 2014. The locations of the study areas that provide the air quality monitoring stations for this research are shown in Figure 1.

3. Methodology

3.1 Air Pollution Index (API)

The air pollution index (API) values at the selected monitoring stations obtained from the Department of Environment (DOE) in Malaysia. The API has been developed based on the Pollutant Standard Index (PSI) and determined by the average indices for five main pollutant variables, which are (PM10, O3, SO2, CO and NO2), and then the maximum value among these five sub-indices is chosen at a particular hour as the API value (DOE 2000).

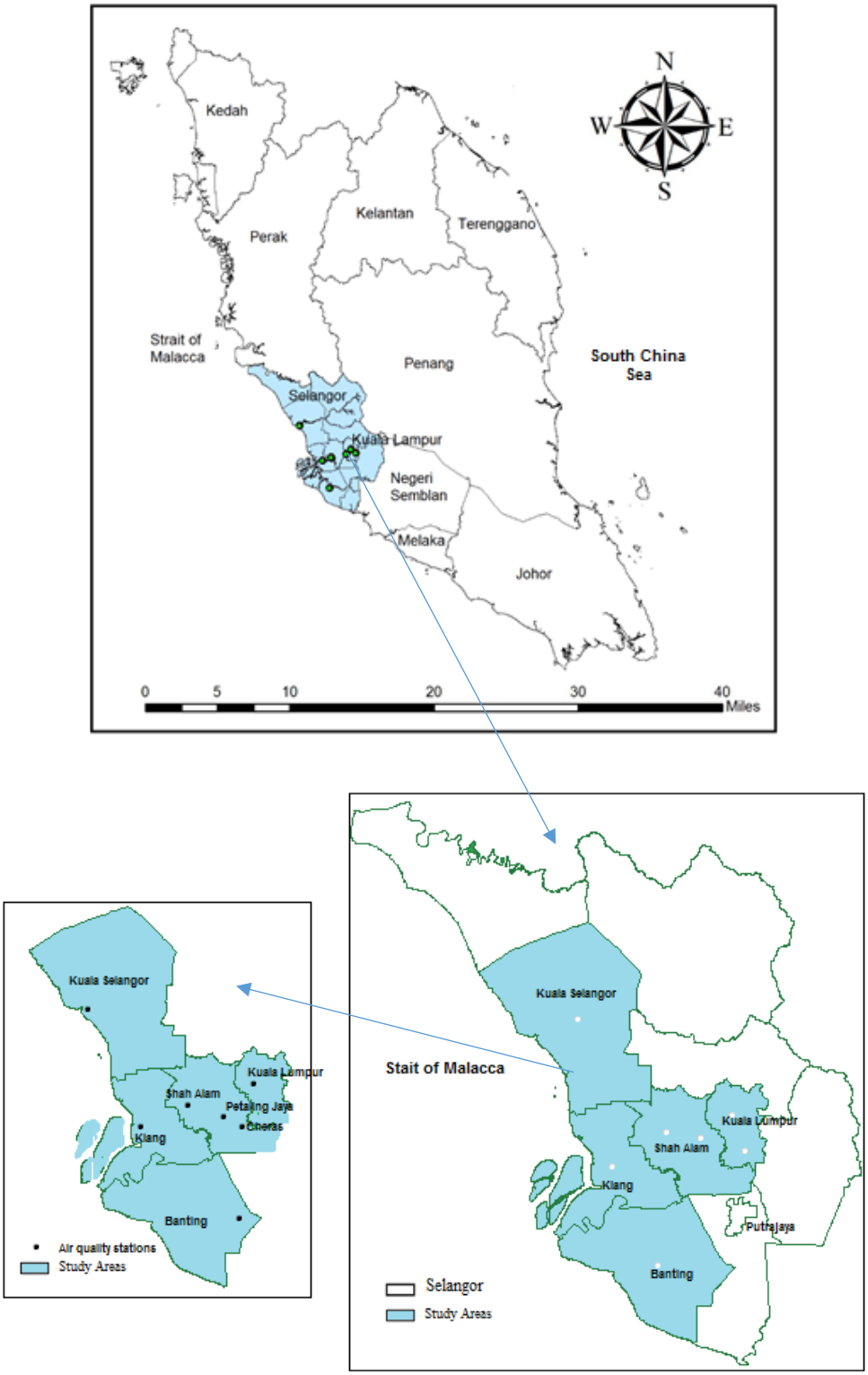


Figure 1. Locations of seven air pollution-monitoring stations in the central region of Peninsular Malaysia

For statistical modelling, the hourly API values are analysed in order to determine the air pollution conditions for seven main cities in the central region of Peninsular Malaysia for a three-year period from 2012 to 2014. In this study, the data of air pollution index is classified into a three-state Markov chain model, namely, $(0 < \text{API} \leq 100)$, $(101 < \text{API} \leq 200)$, $(\text{API} > 200)$, which correspond to moderate, unhealthy and very unhealthy states respectively, representing the three different levels of air quality.

3.2 Empirical Bayesian Estimation of the Transition Probability Matrix

In the analysis of the multinomial data, as in this study, we frequently seek to provide a count matrix with smoothed cell frequencies. In the studies by (Seal &Hossain 2015; Rodrigues & Achcar 2012; Meshkani & Billard 1992), Dirichlet prior has been proposed as a conjugate prior for the parameters of the multinomial distribution. In this section, the determination of the empirical Bayes estimator will be carried out for the multinomial distribution with parameters $\boldsymbol{p_i} = (p_{i1}, p_{i2}, \cdots, p_{ik})$, for all $i = 1, 2, \cdots, k$ under the assumption of conjugate Dirichlet prior. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ denotes the row vector of the transition count matrix $\boldsymbol{y} = [y_{ij}]$, which is an observed random sample that is assumed to follow the multinomial distribution with parameter vector $\boldsymbol{p_i}\ \forall\, i$. The parameter vector $\boldsymbol{p_i}$ is assumed to follow the Dirichlet conjugate prior with parameter $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{ik})$ for all $i = 1, 2, \cdots, k$, which is a natural conjugate prior for the multinomial observations. Accordingly, we have

$$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik}) \sim \text{Multinomial } (y_{i.};\ p_{i1}, p_{i2} \ldots, p_{ik}),$$

where $\boldsymbol{p_i} = (p_{i1}, p_{i2}, \ldots,\ p_{ik}) \sim$ Dirichlet $(\theta_{i1}, \ldots, \theta_{ik})$ for all $i = 1, 2, \ldots, k$. It can be shown that the posterior pdf can be written as

$$f(\boldsymbol{p_i}|\boldsymbol{y_i}, \boldsymbol{\theta_i}) \;=\; \prod_{j=1}^{k} \left[\frac{\Gamma(\sum_{j=1}^{k}(y_{ij} + \theta_{ij})}{\prod_{j=1}^{k}\Gamma\left(y_{ij} + \theta_{ij}\right)}\, p_{ij}^{\,y_{ij} + \theta_{ij} - 1} \right] \qquad (1)$$

3.3 Characteristics of Air pollution

3.3.1 The Mean Residence Time of Air Pollution

Let p_{ij} represent the transition probability of a discrete-time Markov chain $\{X_t, t = 0, 1, 2, \ldots, T\}$ from state i to state j and M_j is the residence time for any state j and h is the number of hours. The probability of observing h hour of residence time is given by

$$Pr(M_j = h) = \left(p_{ij}\right)^{h-1}\left(1 - p_{ij}\right) \qquad (2)$$

Then the mean residence time for any air pollution state j is given by

$$E(M_j|X_t) = \frac{1}{(1-p_{ij})} \quad (3)$$

3.3.2 The Steady-State Probability of Air Pollution

Suppose that $\{X_t, t = 0, 1, 2, \dots, T\}$ be a discrete-time Markov chain with state space S, and let $\boldsymbol{P}$ be the transition probability matrix of the Markov chain. A vector $\boldsymbol{\pi} = [\pi_1 \cdots \pi_s]^t$ is said to be the steady-state probability of air pollution state if elements of $\boldsymbol{\pi}$ are non-negative and satisfy the conditions $\pi_j = \sum_{i=1}^{s} \pi_i \, p_{ij}$ and $\sum_{j=1}^{s} \pi_j = 1$, $j = 1, 2, \dots, s$, indicating that the proportion of time in which the stochastic process stays in a particular state. In a matrix form, the linear equations system of the equation above can be solved based on the following form

$$\boldsymbol{\pi} = [(\boldsymbol{P} - \boldsymbol{I})^t + \boldsymbol{E}]^{-1} \, \boldsymbol{e}, \boldsymbol{P} = [\boldsymbol{p_{ij}}]_{3\times 3} \quad (4)$$

where $\boldsymbol{I}$ is an identity matrix, $\boldsymbol{E}$ is a unit matrix, $\boldsymbol{e}$ is a unit vector and $[(\boldsymbol{P} - \boldsymbol{I})^t + \boldsymbol{E}]$ is a nonsingular matrix (Kulkarni 2011; Tijms 2003; Sanusi et.al 2014). If the value of π_j is high, then the probability of occurrences of state j is high.

4. Results and Conclusion

The mean residence time (MRST), which describes the duration for each state of air pollution event, is determined. It can be seen that most areas experienced MRST for the moderate state to be approximately between 288 to 2359 hours, while for unhealthy and very unhealthy states, the duration of the MRST is shorter, approximately between 4 to 28 and between 2 to 32 hours respectively. This implies that, on the average, the moderate air pollution condition is expected to occur longer than the unhealthy and very unhealthy air pollution conditions in the study areas.

The steady state probability values (SSP) of air quality status for each station are determined to represent the long-term probability of occurrence of a particular air pollution state. It can be seen that the probabilities of the air pollution states (unhealthy and very unhealthy states) slightly varies, with values of from 0.00008 to 0.03026 and for moderate state the average value is 0.975. Therefore, most areas have a high probability of being in the moderate state as opposed to unhealthy and very unhealthy states.

In this study, it is demonstrated that the empirical Bayes method could successfully smoothed out the zero probability in the transition probability matrices. In addition, it also provides a more precise probability values as opposed to those found based on maximum likelihood method.

References

1. Alyousifi, Y., Masseran, N., & Ibrahim, K. (2017). Modeling the stochastic dependence of air pollution index data. Stochastic Environmental Research and Risk Assessment, 1-9.

2. Azid, A., Juahir, H., Aris, A. Z., Toriman, M. E., Latif, M. T., Zain, S. M., ... & Saudi, A. S. M. (2014). Spatial analysis of the air pollutant index in the Southern Region of Peninsular Malaysia using Environmetric Techniques. In From Sources to Solution (pp. 307-312). Springer, Singapore.
3. Hoyos L, Lara P, Ortiz E, Bracho R, Gonza´lez J (2010) Evaluation of air pollution control policies in Mexico city using finite Markov chain observation model. Rev Mat Teor Apl 16(2):255–266
4. Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., Nadzir, M. S. M. (2014). Long Term Assessment of Air Quality from a Background Station on the Malaysian Peninsula. Science of the Total Environment. 482-483: 336-348.
5. Sanusi, W., Jemain, A. A., & Wan Zin, W. Z. (2013). Empirical Bayes estimation for Markov chain models of drought events in Peninsular Malaysia. In AIP Conference Proceedings (Vol. 1571, No. 1, pp. 1082-1089). AIP.
6. Meshkani, M. R., & Billard, L. (1992). Empirical Bayes estimators for a finite Markov chain. Biometrika, 79(1), 185-193.
7. Seal, B., & Hossain, S. J. (2015). Empirical Bayes estimation of parameters in Markov transition probability matrix with computational methods. Journal of Applied Statistics, 42(3), 508-519.
8. Masseran N, Razali A, Ibrahim K, Latif M (2016) Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. Environ Monit Assess 188(1):1–12.
9. Gass K, Klein M, Sarnat S, Winquist A, Darrow L, Flanders W, Chang H, Mulholland J, Tolbert P, Strickland M (2015). Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: a classification and regression tree approach. Environ Health 58:1–14.
10. Seal, B., & Hossain, S. J. (2013). Bayes and Minimax Estimation of Parameters of Markov Transition Matrix. In ProbStat Forum (Vol. 6, pp. 107-115).
11. Rodrigues E, Achcar J (2012) Applications of discrete-time Markov chains and Poisson processes to air pollution modeling and studies. Springer Science and Business Media, New York.
12. DOE (2000) A guide to air pollutant index in Malaysia (API). Department of environment. Ministry of Science, Technology and the Environment, Kuala Lumpur, Malaysia.
13. Tijms, H (2003). A first course in stochastic models. Wiley, England.
14. Sanusi, W., Jemain, A. A., Zin, W. Z. W., & Zahari, M. (2015). The drought characteristics using the first-order homogeneous Markov chain of monthly rainfall data in peninsular Malaysia. Water resources management, 29(5), 1523-1539.

15. Kulkarni V (2011) Introduction to modeling and analysis of stochastic systems, 2nd edn. Springer, New York.
16. Agresti, A., & Chuang, C. (1989). Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. Computational Statistics & Data Analysis, 7(3), 245-258.
17. Fienberg, S. E., & Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. Journal of the American Statistical Association, 68(343), 683-691.
18. AL-Dhurafi, N. A., Masseran, N., & Zamzuri, Z. H. (2018). Compositional time series analysis for Air Pollution Index data. Stochastic Environmental Research and Risk Assessment, 32(10), 2903-2911.

Estimating the proportion of unreported traffic accidents using Bayesian Poisson lognormal model with an adjusted mean

Zamira Hasanah Zamzuri, Nik Sarah Nik Zamri, Kamarulzaman Ibrahim
School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia

Abstract

Typically, traffic accident count data is overdispersed and exhibit the presence of extra zeros. This presence is commonly explained as a result due to the under reporting scenario; in which the accident did occur but not reported. Past research tend to use the zero adjusted or zero inflated models; however these models may not explain the true situation of under reporting. This paper intends to offer an alternative model for traffic accident count data. We propose an adjustment being made to the mean of the Poisson lognormal model by incorporating a parameter to estimate the proportion of unreported accidents. The Poisson lognormal features cater for the overdispersion characteristics whereas the proportion parameter explains the true situation of extra zeros in the data set. Parameters in this model are estimated based on the Bayesian approach. Simulation studies are conducted to compare the performance of the proposed model with existing models in literature. It is expected that this model will offer a satisfactory fit to the data and provides a thorough explanation on the unreported accident count data.

Keywords

traffic accidents; Poisson lognormal; extra zeros

1. Introduction

In order to plan strategies for reducing the risk of traffic accidents, we need to understand factors that contribute to their occurrence. Traffic flow, road condition, weather and geographical location are among factors that potentially influence the occurrence of accidents (Lord et al. 2004). Statistical models have been developed in the traffic accident literature for this purpose. The primary reason for using statistical models in practice is to estimate the safety performance functions (SPFs), which are subsequently used to detect traffic accident black spots, i.e. locations with high frequency of accidents. Among the work that illustrated the use of SPFs in identifying black spots can be found in Lu et al. (2013), Chen (2012) and Sims & Somenahaili (2010).

A groundbreaking work in this field was produced by Maycock & Hall (1984) and Hauer & Lovell (1989), by associating the relationship between the accident rate and explanatory variables, using generalized linear models. The most basic model used is the Poisson regression model (Miao et al. 1992; Miao & Lum 1993). However, traffic accident data are typically overdispersed, hence attention shifted to the negative binomial regression model as can be found in Miao (1994), Vogt (1999), Miao (2001) and Zeeger et al. (2001). The work by Chin & Quddus (2003) and Kweon & Kockelman (2003) show that these univariate fixed effects models are inadequate due to their inability to capture variation caused by unobserved covariates. To cater for this, random effects models that allow the unobserved heterogeneity have been introduced as can be found in Anastasopoulos & Mannering (2009).

Another stream of interest in traffic accident modelling is to handle count data with extra zeros, which is contributed by the underreporting scenario. Underreporting or failure to report the road traffic accidents has frequently increased the attention on the imprecision of data and its impact on road safety policy-making and development. The World Health Report has highlighted the necessity for precise and comprehensive information and scientific methodologies with regard to the prevention and control of road traffic injuries (Peden, 2004). A previous study demonstrated that generally, the exact figure of road crashes is indefinite, and practically entire studies of road crashes comprising greater than single form of data compare only two sources including police and hospital records (Elvick and Mysen.,1999). In this case, the level of comprehensiveness of these datasets is incomplete. The aforementioned two sources fail to include those who do not go to the hospital or to the police, causing in an additional underestimation of underreporting. Evidence has indicated that community-based studies usually provide precise death and injury rates (Sethi et al., 2004).

Commonly to handle the presence of this extra zeros, zero adjusted models are used. According to Winkelmann (2003), when dealing with data in the form of extra zeros, zero inflation model may be used. Overdispersion of zero inflated model is caused by the occurrence of extra zeroes in observed than expected. Zero inflated Poisson (ZIP) model is applied when the count data with extra zeros possess the equality of mean and variance. For data with heavy zeros and long tails, zero inflated negative binomial (ZINB), zero inflated double poisson (ZIDP), and zero inflated generalized Poisson (ZIGP) are suggested (Phang & Loh 2013). A new distribution has very recently been introduced for analyzing data characterized by a large number of zeros. This mixed distribution is known as the NB–Lindley (NBL) distribution which is a mixture of the NB and Lindley distributions. This two-parameters distribution has interesting and sound theoretical properties in which the distribution is characterized by a single long-term mean that is never equal to zero and a

single variance function, similar to the traditional NB distribution (Lord & Geedipally 2011). A discussion on using a mixture of several count distributions to explain the presence of extra zeros can be found in Zamzuri et al. (2018).

Although the zero inflated model give new perspective to statistical modeling of the frequency of accidents, but this model is likely to have difficulty in case of undesirable phenomena (Jang et al 2010). Lord et. al. (2007) say that the zero inflated model has a modeling error in highway accidents. Specifically, this error is defined as that segment is said to be in safe condition although it is likely to be involved in an accident even if no accident occurs for a long period. According to Malyshkina and Mannering (2010), they noticed the phenomenon of the segments based on previous data. However, this data has changed in terms of data collection, the police and the traffic environment, in accordance with the changing times. Hence, they used a Bayesian inference approach since the conventional maximum likelihood estimate (MLE) cannot be used here. This is due to the Markov switching models are difficult to estimate since the state variable is unobservable. The Bayesian approach is a useful statistical approach in which all forms of uncertainty are stated in terms of probability.

This paper intends to introduce a model that capture the presence of these extra zeros based on a different specification. Since that these zeros are hypothesised contributed due to the underreporting scenario, we propose that a proportion parameter is introduced to the accident rate in the Bayesian Poisson Lognormal model.

2. Methodology

Chib and Winkelmann (2002) explained in detail on implementation of the Multivariate Poisson Lognormal (MPL). Park and Lord introduce the usage of this model in traffic accident literature. Several critical elements on fitting the MPL model are also discussed in Zamzuri (2015). The extension of this model with addition of spatio-temporal component can be found in Zamzuri (2018). The specification of this model is presented as an application to traffic accident data, where observations are considered as recorded across intersections, with each intersection having a number of levels of crash severity.

Let i represent the intersection and j the level of crash severity and let $\boldsymbol{y}_i$ be a column vector of reported crash counts for the ith intersection, $\boldsymbol{y}_i = (y_{i1}\ y_{i2} \ldots\ y_{iJ})$. We have N intersections and J levels of severity. Parameter $\boldsymbol{\beta_j} = (\beta_{1j}\ \beta_{2j} \ldots\ \beta_{Kj})$ is a vector of regression coefficients, while $\boldsymbol{X_{ij}}$ is the vector of covariates. Parameter D is the covariance matrix for the random effects vector $b_i = (b_{i1}\ b_{i2} \ldots\ b_{iJ})$. When J=1, this model is reduced to the univariate setting. We called the proposed model as the Adjusted Poisson Lognormal (APL) model. In this model, we add two more parameters, π and τ. Parameter π

measures the proportion of reported accident whereas τ is the true accident rate. Then the specification of this model is

$$Y_{ij}|\beta_j, b_{ij}, \pi \sim Po(\mu_{ij})$$
$$\mu_{ij} = \pi\tau_{ij}$$
$$\mu_{ij} = \exp\left(\beta_{0j} + \beta_{1j}\log(X_{1ij}) + \cdots + \beta_{(K-1)j}\log(X_{(K-1)ij})\right)$$
$$\boldsymbol{b}_i \sim N_J(0, D)$$

The posterior distribution of this proposed model is proportional to

$$\left\{\prod_{j=1}^{J}\phi_K\left(\beta_j|\beta_0, B_0^{-1}\right)\right\} f_w(D^{-1}|\upsilon_0, R_0) f_B(\pi|\alpha_1, \alpha_2) \prod_{i=1}^{N}\phi_J(\boldsymbol{b_i}|D) \prod_{j=1}^{J} f\left(Y_{ijt}|\beta_j, b_{ij}, \pi\right)$$

in which ϕ_K is the k-variate normal distribution, f_w is the Wishart density, $f_B$is the Beta density and f is the Poisson density. Parameter estimation in this model is performed in two phases; first by estimating the proportion of reported accident rate π; then secondly, the estimation of other parameters. Hence, four sampling stages are identified:

1) Sampling π
2) Sampling $\boldsymbol{b}_i$
3) Sampling $\boldsymbol{\beta_j}$
4) Sampling D^{-1}

In Bayesian framework, if the posterior distribution is in the same family as the prior, it is called as conjugate distributions; in which the prior is called as conjugate prior to the likelihood. In cases that the conditional posterior distribution is identified, the Gibbs sampling algorithm can be used. In contrast, when the conditional posterior distribution is unidentified, there is a need for Metropolis Hastings algorithm, in which a sampling density will be proposed.

We present the details of these four sampling stages in the APL model.

1) Sampling π

   We want to sample from a density proportional to $f_B(\pi|\alpha_1, \alpha_2)\ \prod_{j=1}^{J} f\left(Y_{ijt}|\beta_j, b_{ij}, \pi\right)$. Since this is not a recognized distribution, the Metropolis-Hastings algorithm is needed. A technique as suggested in Chib & Winkelmann will be applied here in which we will maximize the log of this density using Newton-Raphson algorithm. Then, the parameter estimates will be sampled from a proposed density.

2) Sampling $\boldsymbol{b}_i$

   We want to sample from a density proportional to $\prod_{i=1}^{N}\phi_J(\boldsymbol{b_i}|D)\ \prod_{j=1}^{J} f\left(Y_{ijt}|\beta_j, b_{ij}, \pi\right)$. Since this is also not a recognized distribution, the Metropolis-Hastings algorithm is needed. The same technique as (1) is used for this stage. Then, we sample from the proposal density, multivariate-t.

   Let

$$\pi_i\left(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},D\right) = \frac{\exp(-0.5\boldsymbol{b}_i{}'D^{-1}\boldsymbol{b}_i)}{|2\pi D|^{0.5}}\prod_{j=1}^{J}\frac{\exp(-\mu_{ij})(\mu_{ij})^{y_{ij}}}{y_{ij}!}$$

$$= k_i\pi_i{}^{+}\left(\boldsymbol{b}_i \mid \boldsymbol{y}_i,\boldsymbol{\beta},D\right)$$

Then, the log of the above function is maximized. The expansion of the function is given as

$$\begin{aligned}&\log(\pi_i{}^{+}\left(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},D\right))\\ =\ &\log\left(\frac{\exp(-0.5\boldsymbol{b}_i{}'D^{-1}\boldsymbol{b}_i)}{|2\pi D|^{0.5}}\prod_{j=1}^{J}\exp(-\mu_{ij})(\mu_{ij})^{y_{ij}}\right)\\ =\ &-0.5\log(|2\pi D|)-0.5(\boldsymbol{b}_i{}'D^{-1}\boldsymbol{b}_i)+\sum_{j=1}^{J}\left(-\mu_{ij}+y_{ij}\log(\mu_{ij})\right).\end{aligned}$$

3) Sampling $\boldsymbol{\beta_j}$

We want to sample from a density proportional to $\left\{\prod_{j=1}^{J}\phi_K\left(\beta_j|\beta_0,B_0^{-1}\right)\right\}\prod_{j=1}^{J}f\left(Y_{ijt}|\beta_j,b_{ij},\pi\right)$. Since this is also not a recognized distribution, the Metropolis-Hastings algorithm is needed. The same technique as (1) is used for this stage. Then, we sample from the proposal density, multivariate-t.

Let

$$\pi_j\left(\boldsymbol{\beta}_j|\boldsymbol{y},b_{ij}\right) = \frac{\exp\left(-0.5(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})'B_0(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})\right)}{2\pi|B_0^{-1}|^{0.5}}\prod_{i=1}^{N}\frac{\exp(-\mu_{ij})(\mu_{ij})^{y_{ij}}}{y_{ij}!}$$

$$= k_j\pi_j{}^{+}\left(\boldsymbol{\beta}_j|\boldsymbol{y},b_{ij}\right)$$

Then, the log of the above function is maximized. The expansion of the function is given as

$$\begin{aligned}&\log\pi_j{}^{+}(\boldsymbol{\beta}_j\,|\boldsymbol{y},b_{ij})\\ =\ &\log\left(\frac{\exp\left(-0.5(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})'B_0(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})\right)}{2\pi|B_0^{-1}|^{0.5}}\prod_{i=1}^{N}\exp(-\mu_{ij})(\mu_{ij})^{y_{ij}}\right)\\ =\ &-0.5\log\left|2\pi B_0^{-1}\right|\ -0.5(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})'B_0(\boldsymbol{\beta}_j-\boldsymbol{\beta_0})+\sum_{i=1}^{N}(-\mu_{ij}+y_{ij}\log(\mu_{ij}))\end{aligned}$$

4) Sampling D^{-1}

We want to sample from from a density proportional to $f_w(D^{-1}|\upsilon_0,R_0)\prod_{i=1}^{N}\phi_J(\boldsymbol{b_i}|D)$. We can see that this function is distributed as Wishart, hence we can sample directly from the Wishart distribution using the Gibbs sampling.

The derivation of the function is given as

$$
\begin{aligned}
& f_w\left(D^{-1}|\upsilon_0, {R_0}^{-1}\right) \prod_{i=1}^{N} \phi_J\left(\boldsymbol{b}_i|0, D\right) \\
= & \frac{|D^{-1}|^{\frac{\upsilon_0-J-1}{2}}}{2^{\frac{J\upsilon_0}{2}}\left|{R_0}^{-1}\right|^{\frac{\upsilon_0}{2}}\Gamma_p\left(\frac{\upsilon_0}{2}\right)}\exp\{-0.5\ Tr\ (R_0\ D^{-1})\}\ \prod_{i=1}^{N}\frac{\exp(-0.5\boldsymbol{b}_i'\boldsymbol{b}_i)}{|2\pi D|^{0.5}} \\
= & \frac{|D^{-1}|^{\frac{\upsilon_0-J-1}{2}}}{2^{\frac{J\upsilon_0}{2}}\left|{R_0}^{-1}\right|^{\frac{\upsilon_0}{2}}\Gamma_p\left(\frac{\upsilon_0}{2}\right)}\exp\{-0.5\ Tr\ (R_0\ D^{-1})\}\ \left\{\frac{1}{|2\pi D|^{0.5N}}\right\} \\
& \exp\{\ -0.5(\boldsymbol{b'}_1\boldsymbol{b}_1+\boldsymbol{b'}_2\boldsymbol{b}_2+\cdots+\boldsymbol{b'}_N\boldsymbol{b}_N)\} \\
= & \frac{|D^{-1}|^{\frac{\upsilon_0-J-1}{2}}}{2^{\frac{J\upsilon_0}{2}}\left|{R_0}^{-1}\right|^{\frac{\upsilon_0}{2}}\Gamma_p\left(\frac{\upsilon_0}{2}\right)|2\pi D|^{0.5N}}\exp\{-0.5\ \ (\mathrm{Tr}\ (R_0D^{-1})\ +\boldsymbol{b}_1'\boldsymbol{b}_1+\boldsymbol{b'}_2\boldsymbol{b}_2+\cdots+\boldsymbol{b'}_N\boldsymbol{b}_N)\} \\
= & f_w\left(D^{-1}|N+\upsilon_0, \left({R_0}^{-1}+\sum_{i=1}^{N}\boldsymbol{b}_i'\boldsymbol{b}_i\right)^{-1}\right)
\end{aligned}
$$

3. Results

We will present the results on fitting the proposed (APL) and zero inflated Poisson (ZIP) models to simulated data sets based on both distributions. Hence, the results will be arranged in this order:

1) Comparison of APL and ZIP models fitted to simulated data based on APL
2) Comparison of APL and ZIP models fitted to simulated data based on ZIP

A detailed result will be included later.

4. Discussion and Conclusion

In this paper, we have introduced a new model for traffic accident count data based on the adjustment to the MPL model in which a proportion parameter is incorporated into the model. The model fitting process consists of two phases: first estimating the proportion parameter and then estimating other parameters in the new model using an MCMC procedure. We expect that the simulation results reveal that this model performs better and able to explain the presence of extra zeros due to the underreporting scenario.

It is important to note that the aim of this paper is to offer an alternative on fitting count models to the data with extra zeros, specifically in traffic accident environment. Furthermore, more information is obtained through the estimation of the proportion of the reported accidents. When the proportion of the reported accidents and reported accidents rate are estimated, this allows us to estimate the true accident rate in which help in terms of estimation accuracy on the actual accident count that happen.

The work here illustrates the need on reporting traffic accidents and a proper data documentation practice. Such issue may need to be handled in the long run, but with the current situation, especially in Malaysia; in which the source

of accident counts is based on police reports of the accidents, development of a model such as proposed in this paper helps addressing the issue in hand.

References

1. D. Lord, S. P. Washington and J. N. Ivan, "Statistical challenges with modelling motor vehicle crashes: Understanding the implications of alternative approaches" in *83rd Annual Meeting of Transportation Research Board*, 2004.
2. J. Lu, A. Gan, K. Haleem, W. Wu, Journal of Transportation Safety & Security **5**, 224-239 (2013).
3. H. Chen, Journal of Geographic Information System **4**, 608-617 (2012).
4. A. Sims, S. V. C Somenahalli, "Hot spot identification using frequency of distinct crash types rather than total crashes". Australian Transport Research Forum, Canberra, Australia, 2010.
5. G. Maycock and R. D. Hall, "Accidents at 4-arm roundabouts". Laboratory Report LR1120, Transport Research Laboratory, Crowthorne, Berks, UK, 1984.
6. E. Hauer, J. C. N. Ng, J. Lovell, Transportation Res. Rec. 1185, 48-61 (1988).
7. S. P. Miaou, P. S. Hu, T. Wright, A. K. Rathi and S. C. Davis, Transportation Research Record **1376**, 10-18 (1992).
8. S. P. Miaou and H. Lum, Accident Analysis & Prevention **25**(6), 689-709 (1993).
9. S. P. Miao, Accident Analysis & Prevention **26**, 471-482 (1994).
10. A. Vogt, "Crash Models for Rural Intersections: Four-lane by Two-lane Stop- controlled and Two-lane by Two-lane Signalized" (FHWA-RD-99-128). McLean, VA: Turner Fairbank Highway Research Center, Federal Highway Administration, 1999.
11. S. P. Miaou, "Estimating Roadside Encroachment Rates with the Combined Strengths of Accident- and Encroachment-Based Approaches" (FHWARD-01-124). Oak Ridge, TN: Oak Ridge National Laboratory, 2001.
12. C. V. Zegeer, J. R. Stewart, H. H. Huang, and P. A. Lagerwey, Transportation Research Record **1773**, 56-68 (2001).
13. H. C. C. Chin and M. A. Quddus, Accident Analysis and Prevention **35**(2), 253-259 (2003).
14. Y. J. Kweon and K. M. Kockelman, Accident Analysis & Prevention **35**(4), 441-450 (2003).
15. P. C. Anastasopoulos and F. L. Mannering, Accident Analysis and Prevention **41**(1), 153-159 (2009).
16. Peden, M., Scurfield, R., Sleet, D., 2004. World Report on Road Traffic Injury Prevention.
17. Y.N. Phang and E.F. Loh, International Scholarly and Scientific Research & Innovation **7**, 817-819 (2013).

18. D. Lord and S.R.Geedipally, Accident Analysis and Prevention **43**, 1738-1742 (2011).
19. D. Lord and Y. J. Park, Accident Analysis and Prevention **40**, 1441-1457 (2008).
20. D. Lord, S.P. Washington and J.N. Ivan, Accident Analysis and Prevention **39**, 53-57 (2007).
21. H. Jang, S. Lee and S.W. Kim, Accident Analysis and Prevention **42**, 540-547 (2010).
22. N.V. Malyshkina and F.L. Mannering, Accident Analysis & Prevention **42**, 122-130 (2010).
23. S. Chib and R. Winkelmann, Markov Chain Monte Carlo Analysis of Correlated Count Data. Journal of Business & Economic Statistics **19**(4), 428–435 (2001).
24. Zamzuri, Z. H., The spatio-temporal multivariate Poisson lognormal model. 28 Jun 2018, *Proceeding of the 25th National Symposium on Mathematical Sciences, SKSM 2017: Mathematical Sciences as the Core of Intellectual Excellence.* American Institute of Physics Inc., Vol. 1974. 020013
25. Zamzuri, Z. H., Sapuan, M. S. & Ibrahim, K., The extra zeros in traffic accident data: A study on the mixture of discrete distributions. 1 Aug 2018, In : Sains Malaysiana. 47, 8, p. 1931-1940 10 p.
26. Zamzuri, Z. H., Critical elements on fitting the Bayesian multivariate Poisson Lognormal model 22 Oct 2015, *22nd National Symposium on Mathematical Sciences, SKSM 2014: Strengthening Research and Collaboration of Mathematical Sciences in Malaysia.* American Institute of Physics Inc., Vol. 1682. 050005

Engineering applications of hierarchical Bayesian modeling

Stephen Wu[1], Panagiotis Angelikopoulos[2], James L. Beck[3], Petros Koumoutsakos[4]

[1]Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Tokyo 190-8562, Japan

[2]D.E. Shaw Research, New York, NY 10036, USA

[3]California Institute of Technology, Pasadena, CA 91125, USA

[4]Computational Science and Engineering Laboratory, ETH-Zurich, CH-8092, Switzerland

Abstract

Bayesian modelling and inference has become a very important method to many modern engineering applications because it allows a unified framework for uncertainty quantification and propagation to various problems, such as model selection and robust prediction. A major trade-off comes from the heavy computation demand, which prohibits the use of the full Bayesian framework to complex simulation models. In particular, hierarchical Bayesian model is a powerful modelling tool that offers great flexibility for uncertainty quantification, yet classical Markov Chain Monte Carlo approach is usually impractical for even a simple ordinary differential equation model. In my study, I begin with a basic illustration of the power of hierarchical Bayesian model, and then continue with a demonstration of its applications to engineering problems by incorporating high performance computing and specifically designed sampling methods. The applications, including pharmacokinetics and molecular dynamics, involve fairly complicated models that classical models used for Bayesian inference often lead to misleading results.

Keywords

Hierarchical Bayesian modeling; uncertainty quantification; importance sampling; complex simulation

1. Introduction

Bayesian modelling and inference has become a very important method to many modern engineering applications because it allows a unified framework for uncertainty quantification and propagation to various problems, such as model selection and robust prediction (Beck, 2010). A major trade-off comes from the heavy computation demand, which prohibits the use of the full Bayesian framework to complex simulation models, for example finite element models of large civil structures and high-resolution fluid dynamics simulations. Advanced Markov Chain Monte Carlo methods and various Bayesian modeling techniques have extended the applications to a

boarder range of engineering problems, such as reliability estimation due to rare events (Au and Beck, 2001). In particular, hierarchical Bayesian model is a powerful modeling tool that offers great flexibility for uncertainty quantification, yet classical Markov Chain Monte Carlo approach is usually impractical for even a simple ordinary differential equation model. This is because of the inherently high dimensional problem setup as explained in this study. Existing approaches for hierarchical Bayesian modeling usually attempt to use simple stochastic models that can lead to analytical results, but usually ends up with impractical assumptions, or approximating the stochastic models with surrogate models that can result in analytical solutions. In my study, I begin with a basic illustration of the power of hierarchical Bayesian model, and then continue with a demonstration of its applications to engineering problems by incorporating high performance computing and specifically designed sampling methods. Our method focuses on the ability to recalculate the problem for many times, especially when there are newly added data. The applications, including pharmacokinetics and molecular dynamics, involve fairly complicated models that classical models used for Bayesian inference often lead to misleading results. Therefore, it is important to use a reliable, yet computationally efficient algorithm for these problems.

2. Methodology

2.1 Problem setup for the general case

Consider the following probability model:

$$y \sim N\left(y \middle| f\left(x, \vec{\theta}\right), \sigma_y\right) \Leftrightarrow y = f\left(x, \vec{\theta}\right) + \epsilon_y, \epsilon_y \sim N\left(\epsilon_y \middle| 0, \sigma_y\right), \tag{1}$$

where $N(z|\mu, \sigma)$ denotes a normal distribution on a 1D variable z with mean μ and standard deviation σ, x and y denotes input and output variable of a model (function) f with model parameters $\vec{\theta}$. A hierarchical Bayesian model has hyperparameters $\vec{\psi}$ for $\vec{\theta}$ to define a probability model for the parameter space. However, one can typically find two different types of such hierarchical Bayesian models in the literature (Wu et al., 2018). Here, I illustrate using a simple linear example.

2.2 Simple example: single parameter with Gaussian prior model

Consider a simple linear function with only one parameter, i.e., $\vec{\theta} = \theta$, and a Gaussian model for the parameter with mean and standard deviation as the hyperparameters, i.e., $\vec{\psi} = \{\mu_\theta, \sigma_\theta\}$:

$$f(x, \theta) = \theta x \tag{2}$$
$$\text{with } \theta = \mu_\theta + \epsilon_\theta, \epsilon_\theta \sim N(\epsilon_\theta|0, \sigma_\theta) \Leftrightarrow \theta \sim N(\theta|\mu_\theta, \sigma_\theta)$$

Let us consider a set of data that contains n data subsets, $\mathcal{D} = \{\mathcal{D}_i | i = 1, \dots, n\}$. When we want to infer θ using $\mathcal{D}$, the two types of hierarhical Bayesian models, named M1 and M2 model, are basically assuming if θ can take only a single value for all data subsets or not, respectively (Fig. 1). Table 1 demonstrate the theory of Bayesian inference for those two cases.

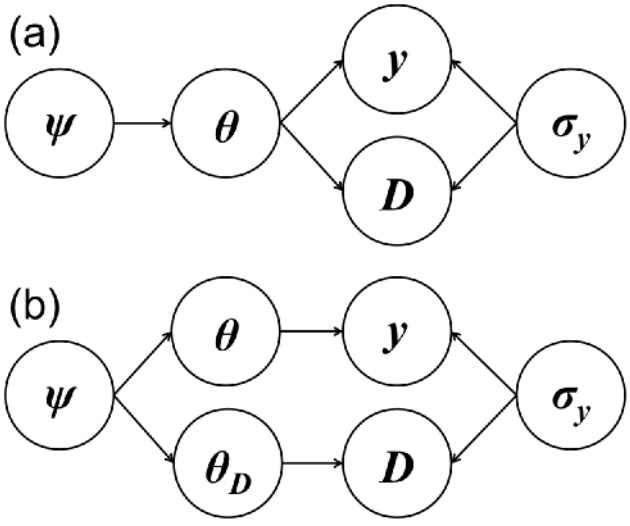


Figure 1: Bayesian network for (a) model of single-data structure (M1), and (b) model of hierarchical data structure (M2).

In most cases, completely analytical solutions are not available for the posterior PDF of the hyperparameters. An estimation method based on importance sampling (IS) can be used. Luckily, in our simple linear problem,

Table 1: Comparison between single and hierarchical data structure

	M2	M1
$p(\theta\|\mathcal{D}) =$	$\iint p(\mathcal{D}\|\sigma_y, \vec{\psi}) \frac{p(\theta\|\vec{\psi}) p(\sigma_y, \vec{\psi})}{p(\mathcal{D})} \mathrm{d}\sigma_y \mathrm{d}\vec{\psi}$	$\iint p(\mathcal{D}\|\theta, \sigma_y) \frac{p(\theta\|\vec{\psi}) p(\sigma_y, \vec{\psi})}{p(\mathcal{D})} \mathrm{d}\sigma_y \mathrm{d}\vec{\psi}$
$p(\mathcal{D}\|\sigma_y, \vec{\psi}) =$	$\prod_{i=1}^{N_D} \int p(D_i\|\theta_{D_i}, \sigma_y) p(\theta_{D_i}\|\vec{\psi}) \mathrm{d}\theta_{D_i}$	$\int \left(\prod_{i=1}^{N_D} p(D_i\|\theta, \sigma_y) \right) p(\theta\|\vec{\psi}) \mathrm{d}\theta$

analytical solutions can be achieved Wu et al. [2018].

Let us consider three different data sets for this case study, with random input x ranging from 0 to 10, $\mu_\theta = 1$, $\mu_y = \mu_\theta = 0.2$ and 20 data set with 50 data point in each set (a total of 1000 data points), i.e., $N_D = 20$ and $N_{D_i} = 50$ for all i.

Data Set 1) All uncertainties are moved to the measurement error, i.e., σ_θ = 0 and σ_y = 0.4. Each data point among the 1000 is generated independently from a random measurement noise.

Data Set 2a) Each data point among the 1000 is generated independently from a random noise for σ_y = 0.2 and a random noise for σ_θ = 0.2.

Data Set 2b) For each data set D_i, one fixed value of $\theta^{(i)}$ is generated based on σ_θ = 0.2 and it is used to generate the 50 data points with independent random noise for σ_y = 0.2.

Figure 2 shows the three data sets generated for this test. You can see a clear distinction between data generated from different presumed stochastic models. Also, Data Set 2b shows more regularity compared to Data Set 2a because it is simulated based on 20 fixed ϑ values. After applying the hierarchical Bayesian modeling analysis, the results turn out to show only the hierarchical can successfully infer ϑ in all three cases correctly. Although this may be a very intuitive result, such a data structure differences are often seen in many practical engineering problems. Hence, we stress the importance of developing an efficient yet reliable algorithm for such kind of hierarchical Bayesian model inference.

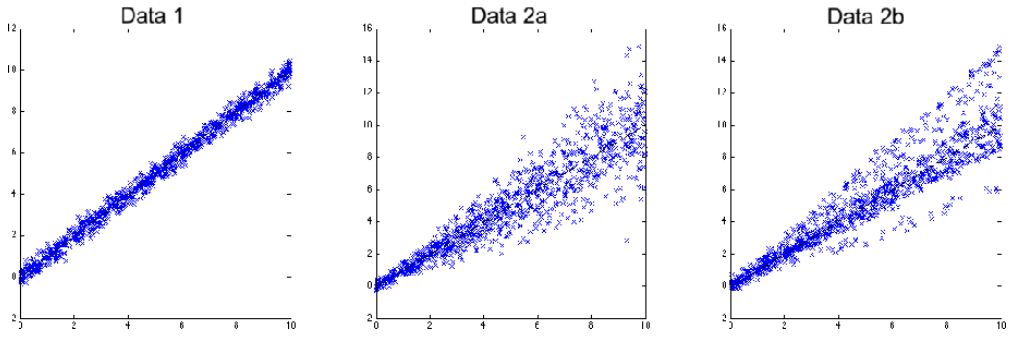


Figure 2: Three sets of data generated for the test.

2.3 Efficient approximation for complex model

Hierarchical Bayesian models require efficient estimation of $p(\mathcal{D}|\sigma_y, \vec{\phi})$ as shown in Table 1. This is a very difficult problem, especially when the likelihood function $p(D_i|\theta_{D_i}, \sigma_y)$ involves evaluating a very computational demanding function. Some of the common solutions include using conjugate pairs to achieve analytical results (Congdon, 2010), using approximation from Laplace Asymptotic Approximation (Wu et al., 2015), or using specially designed Markov Chain Monte Carlo techniques (Nagel and Sudret, 2015). Here, we adopt the post-processing approach proposed in Wu et al. (2018), which was developed to meet many practical constraints.

The key concept of the post-processing method is to pick a general prior proposal $p(\theta_{D_i}|\sigma_y)$ for each likelihood $p(D_i|\theta_{D_i},\sigma_y)$. If we can perform a typical Bayesian inference using these priors for all the corresponding likelihoods, we will be able to obtain the following posterior:

$$p(\theta_{D_i}|D_i,\sigma_y) = \frac{p(D_i|\theta_{D_i},\sigma_y)p(\theta_{D_i}|\sigma_y)}{p(D_i|\sigma_y)} \tag{3}$$

$q_i(\theta_{D_i}^{(j)}) = p(\theta_{D_i}|D_i,\sigma_y)$:

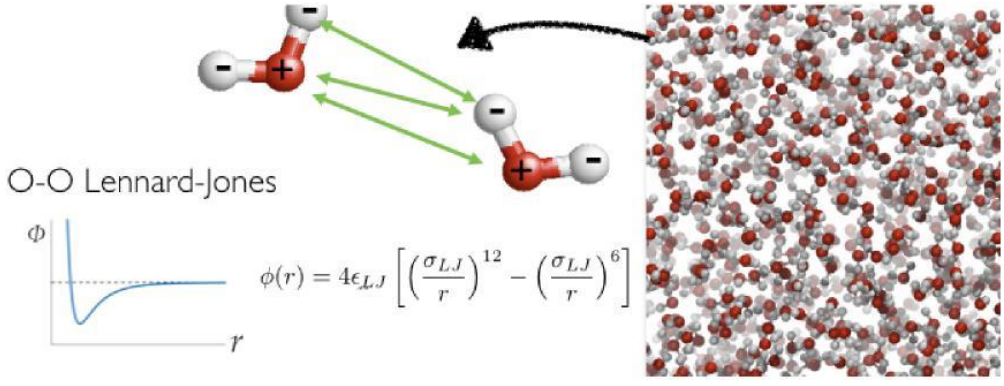
If we can obtain estimation of the evidence term $p(D_i|\sigma_y)$, we will be able to apply the importance sampling method to estimate the integrals, i.e., an estimate of the important term of $p(\mathcal{D}|\sigma_y,\vec{\psi})$ by setting the proposals $q_i\left(\theta_{D_i}^{(j)}\right) = p(\theta_{D_i}|D_i,\sigma_y)$:

$$\begin{aligned} p(\mathcal{D}|\sigma_y,\vec{\psi}) &\approx \prod_{i=1}^{N_D} \frac{1}{N_{s,i}} \sum_{j=1}^{N_{s,i}} \frac{p(D_i|\theta_{D_i}^{(j)},\sigma_y)p(\theta_{D_i}^{(j)}|\psi,\sigma_y)}{q_i(\theta_{D_i}^{(j)})} \\ &= \prod_{i=1}^{N_D} \frac{p(D_i|\sigma_y)}{N_{s,i}} \sum_{j=1}^{N_{s,i}} \frac{p(\theta_{D_i}^{(j)}|\psi,\sigma_y)}{p(\theta_{D_i}^{(j)}|\sigma_y)} \\ &\text{where } \theta_{D_i}^{(j)} \sim p(\theta_{D_i}|D_i,\sigma_y) \end{aligned} \tag{4}$$

Because it is very common to first perform Bayesian inference on each data set D_i to get a rough understanding of the problem, this approximation method for hierarchical Bayesian modeling can be seen as a very efficient post-processing that recycle the samples drawn during those Bayesian inferences.

3. Results

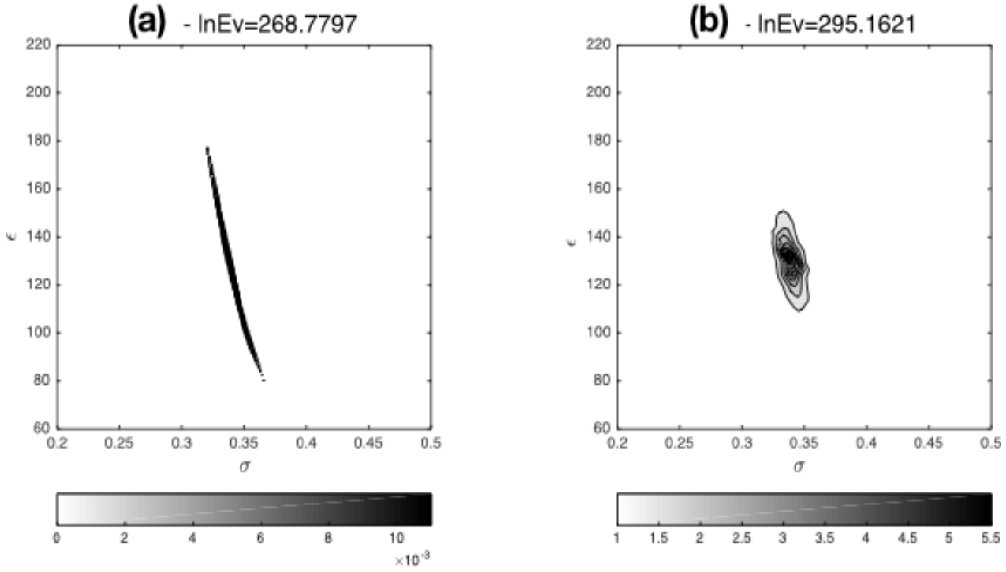
The first example we give is on the calibration of the force fields in Molecular Dynamics (MD) simulations. MD is computational method to simulate the dynamic evolution of molecules under a given environment. The force field that controls the



interaction between molecules is a critical aspect of the predictive capabilities of MD simulations (Fig. 3).

Figure 3: Brief introduction to the key parameters of MD simulation.

Several factors can contribute to such discrepancies, such as the choice of the force field and its calibration, computational errors and experimental uncertainties. Furthermore, the calibration of force fields in MD simulations often relies on experimental data that exhibit a special structure. The experimental data, which is the measurements of some physical quantities, often contains repeated data set using different measurement techniques or under variable environmental conditions. Therefore, this is a perfect example to demonstrate the use of hierarchical Bayesian models (Wu et al., 2015a; Wu et al., 2015b). We can observe from Fig. 4 that non-



hierarchical model tends to under-estimate the uncertainty of the parameters, and instead capturing some strong correlation between the two MD model parameters.

Figure 4: Comparing results of (a) non-hierarhical Bayesian model inference and (b) hierarhical Bayesian model inference.

The second example we give is on the calibration of pharmacokinetics models based on actual clinical data. Once again, these data, which can be coming from different patients or same patient but different period of time, exhibit a similar data structure to our simple linear example. Therefore, this is also a perfect example to demonstrate the use of hierarchical Bayesian models (Wu et al., 2018).

4. Discussion and Conclusion

Hierarchical Bayesian model is an essential tool for many engineering problems, because most of the experiments in practice are performed repeatedly with some inevitable environmental changes. Ignoring such uncertainties will often result in misleading conclusions. It is known that when a likelihood model only consider additive noise, increase of data will lead to shrinking of the model parameters. This is because of the false assumption that all data points are independent, while in fact, there is correlation within each experiment. Therefore, hierarchical modeling is needed to properly capture the model uncertainty, leading to reasonable decision-making. Here, we demonstrate an efficient approximation for the computationally demanding hierarchical model developed under practical concerns. This allows possible application of hierarchical Bayesian model to a wide range of engineering applications with complex models.

References

1. Au, S. and Beck, J.L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. Probabilistic Engineering Mechanics, 16(4):263–277.
2. Beck, J.L. (2010). Bayesian system identification based on probability logic. Structural Control and Health Monitoring, 17(7):825–847.
3. Congdon, P. (2010). Applied Bayesian Hierarchical Methods. CRC Press.
4. Nagel, J.B. and Sudret, B. (2016). A unified framework for multilevel uncertainty quantification in Bayesian inverse problems. Probabilistic Engineering Mechanics, 43:68–84.
5. Wu, S., Angelikopoulos, P., Papadimitriou, C., Moser, R. and Koumoutsakos, P. (2015a). A hierarchical Bayesian framework for force eld selection in molecular dynamics simulations. Phil. Trans. R. Soc. A, 374(2060):20150032.
6. Wu, S., Angelikopoulos, P., Tauriello, G., Papadimitriou, C. and Koumoutsakos, P. (2015b). Fusing heterogeneous data for the calibration of molecular dynamics force fields using hierarchical Bayesian models. The Journal of Chemical Physics, 145(24):244112.
7. Wu, S., Angelikopoulos, P., Beck, J.L. and Koumoutsakos, P. (2018). Hierarchical stochastic model in Bayesian inference for engineering applications: Theoretical implications and efficient approximation. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, 5(1):011006.

Compilation of sectorisation and information granularities on portfolio investment statistics

Norhayati Razi
IIP and EDS, Statistical Services Department Central Bank of Malaysia

Abstract

Given the rapid growth in international financial flows as well as increasing interest by policy makers and market analysts on cross-border portfolio investment, the Central Banks in their dual role as compilers and users of statistics faced numerous challenges in compiling the statistics. For portfolio investment in reporting economies, difficulties to obtain accurate information arises when the securities are not fully settled through domestic custodians. On the other hand, collecting information on portfolio investment abroad could hampered by insufficient details as well as constraints imposed to limit the reporting entities' reporting burden. This presentation is to share experience in portfolio investment data compilation through an integrated approach, particularly the benefits of subscribing to IMF's centralised securities database on International Securities Identification Number (ISIN) information.

Keywords

quarterly IIP survey, portfolio investment statistics, international financial flows, information granularities, reporting burden.

1. Introduction

External sector statistics in Malaysia is compiled jointly by Bank Negara Malaysia (BNM, the Central Bank) and the Department of Statistics, Malaysia (DOSM, the national statistics office), through the formal institutional arrangement and the Memorandum of Understanding (MOU) signed between the two agencies. Each institution carries out its respective roles and responsibilities, guided by the MOU and the legal frameworks, namely the Central Bank of Malaysia Act, 2009 (CBA 2009) and the Statistic Act 1965. These legal provisions provide clear guidance for the compilation and dissemination of external sector statistics. While DOSM focuses on the compilation of Current Account data items, the Central Bank leads on the compilation of Investment Income and Financial Account through the quarterly International Investment Position (IIP) Survey. In terms of dissemination of official statistics, DOSM is the official compiler of the balance of Payments (BOP), IIP, and Coordinated Direct Investment Statistics (CDIS), while the Central Bank is the official compiler of the External Debt Statistics

(EDS), Coordinated Portfolio Investment Statistics (CPIS) and International Banking Statistics (IBS).

2. Methodology

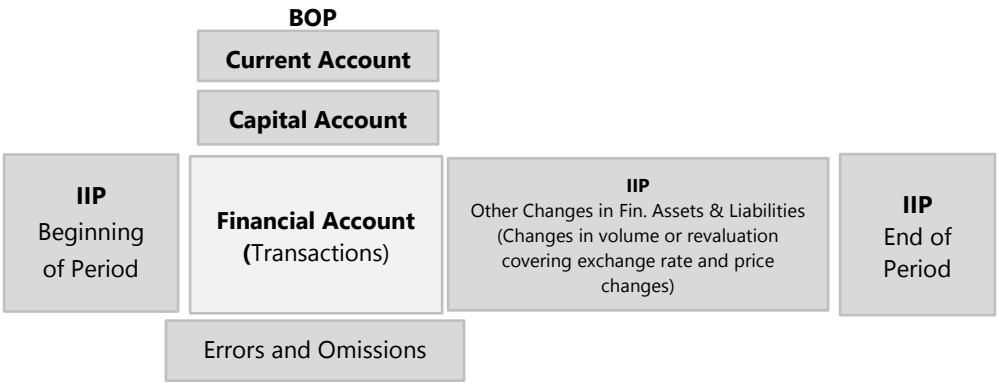
The Portfolio Investment statistics are collected as part of an integrated framework for compiling quarterly BOP Investment Income and Financial Account, EDS, IIP and IBS by the Central Bank. The reporting entities (REs), which comprise custodian agencies and reporting entities that report their owned exposures, record information on transactions, non-transactional and positions of the securities using the ISIN codes and other details required for external sector statistics, such as counter-party and/or issuer name, institutional sector, country, currency, as well as maturity structure where applicable.

2.1 External Sector Compilation in Malaysia

Following the implementation of Balance of Payments and International Investment Position Manual, Sixth Edition (BPM6) in 2009, IMF has strongly recommended on an integrated approach be adopted by compilers to facilitate the production of the BOP and its other related statistics, including IIP, EDS, CPIS, CDIS and IBS, from a single source in order to ensure data quality and consistency throughout these datasets. The compilation approach via the integrated IIP framework is able to fully explain the changes in stock position resulting from financial account transactions, revaluations (separately identifying price and exchange rate changes) and other changes in volume of assets and liabilities, as well as consistent statistics for the Financial Account (FA) of BOP, IIP, EDS and IBS. Data are of intrinsic value in themselves, and to support the assessment of the various measures included in the integrated IIP that facilitate the analysis on factors affecting the movement of the investment position of the country.

The illustration for the Integrated IIP Framework are as follows:

Integrated IIP Framework — Diagram 1



Note: Transactions reported in the integrated IIP framework provides BOP flows, while closing position is the IIP statistics, with EDS being derived from stock position of non-equity liability instruments. CDIS and CPIS refer to stock position of direct and portfolio investment respectively, while IBS is the IIP of banking sector.

The quarterly IIP Survey collects granular data on item-by-item reporting, whereby the items reflect the financial instruments and are reported in the dimensions of the country of the counterparty, currency of transactions and maturity structure for selected financial flows. There are 36 data items that are identified by unique purpose of transactions, namely equity capital, retained earnings, equity securities, debt securities, loans, deposits and others. These items are reported with further details based on each individual counter party, which are crucial identity to determine the types of investment for BOP classification. Additional details, such as name, relationship, percentage holdings of investors and affiliated companies are collected in order to facilitate more in-depth analysis, particularly in the areas of foreign direct investments in Malaysia as well as Malaysia's direct investment abroad.

The main sources of external sector statistics compilation for the IIP Survey on external assets and liabilities (EAL) vis-à-vis non-residents are as follows:

i. Data on EAL of banking institutions;
ii. Data on EAL of selected resident entities, which include direct and portfolio investment abroad, foreign direct investment in Malaysia, credit facilities from non-residents and any other assets and liabilities vis-à-vis non-residents;
iii. Data on portfolio investments reported by custodians, which include holding of Malaysian securities by non-residents and portfolio investments of residents vis-à-vis non-residents;
iv. Data on EAL of entities in Labuan offshore, managed by the Labuan Financial Services Authority; and Data on the government sector, obtained

from the administrative records of the Treasury Department of the Ministry of Finance, Malaysia.

In addition to the above, BNM also submits International Banking Statistics (IBS), in compliance with the Bank for International Settlement (BIS) requirements. Thus far, Malaysia has been submitting the IBS by "Locational" and "Nationality" and targets to submit the "Consolidated Banking Statistics" once the newly implemented compilation system and the statistics collected are credible. Towards this end, BNM also collects the following statistics to ensure the comprehensiveness of the IBS submission to BIS:

i. Data on assets and liabilities (AL) of all banking institutions vis-à-vis residents; and
ii. Data of Malaysian banks' on ultimate risks transfers, foreign branches and subsidiaries as well as consolidated financial position.

The external sector data is collected on a mandatory basis for all financial institutions and identified non-bank entities in Malaysia. All data is to be submitted to the Central Bank by 15 days after end of each reporting period. Penalties and legal actions can be taken against the respondents that do not comply with the reporting requirements, including incidences of non-reporting.

2.2 Reporting of Securities Issued in Malaysia

Reporting entities, mainly the custodians, are required to report security-by-security, based on the ISIN codes provided by the Central Bank of Malaysia, as well as other details on the non-resident holders. The Central Bank maintains the Entity Database that includes the ISIN codes and profiles of the issuers as well as other information required for the statistical compilation purposes, through an integration with various following sources:

i. ISIN codes on local equity securities are obtained regularly from Bursa Malaysia (the Kuala Lumpur Securities Exchange), together with the quarterly position of each individual securities to facilitate the cross-checking of data reported by the custodians in the quarterly IIP Survey;
ii. Integration with the Information and Surveillance System for Debt Securities (INSIDES), the Central Bank's bond data warehouse, which is also a back-end debt securities trading platform of Real-time Electronic Transfer of Funds and Securities System (RENTAS), to obtain both the ISIN for sharing with reporting entities, as well as the stock position of the securities to facilitate the cross-checking of data on debt securities issued in Malaysia being reported by the custodian agencies; and
iii. Integration with the Companies Commission of Malaysia for the purpose of entity validation and to obtain details of entity's information, to facilitate data enrichment in terms of institutional and business sector for the publication of statistics that is aligned with international standards.

This integrated compilation has helped to ease the reporting burden of the reporting entities and to ensure accuracy of the profiles of securities issuers. Listed below are the coverage of the data sources reported by the custodian agencies for the quarterly reporting:

i. Data on *resident holdings of securities issued by non-residents in Malaysia*; and
ii. Data on *non-resident holdings of securities issued by residents in Malaysia*.

The integrated compilation framework classifies the reporting of external financial assets and liabilities by direct investment, portfolio investment, financial derivatives, and other investment, which are being reported under one survey. Guidance notes, which are aligned to BPM6, provide clear instructions on method of reporting to prevent misreporting between direct and portfolio investment by the reporting entities. In addition, the information obtained from direct investment reported by the end-investors are used as a basis for reconciliation to avoid double counting and separating direct investment from portfolio investment.

2.3 Reporting of Securities Issued in International Markets

The reporting of portfolio investment which are classified by institutional sector, facilitates identification of the pattern of investment. Institutional investors tend to have direct holdings abroad, while the mutual funds, unit trusts, and households are more likely to invest through the nominees as reported by the custodians. In general, the reporting entities are required to report all the individual securities based on the name, country and institutional sector of the issuer, as well as currency of the securities issued.

Although highly encouraged, the reporting for securities issued in the international market by ISIN codes is not mandatory in view of the challenges faced by reporting entities, particularly if the securities are entrusted to another non-resident custodians or as part of the pool of funds managed by non-residents custodians. On the other hand, the institutional investors reporting data on their portfolio investment holdings abroad are able to report by ISIN codes and other details, particularly if the decision for the investment are carried out internally. The detailed reporting by ISIN codes and countries of residence issuers has somewhat facilitate the reconciliation if the investments through the non-resident custodians are rechannelled back to home country.

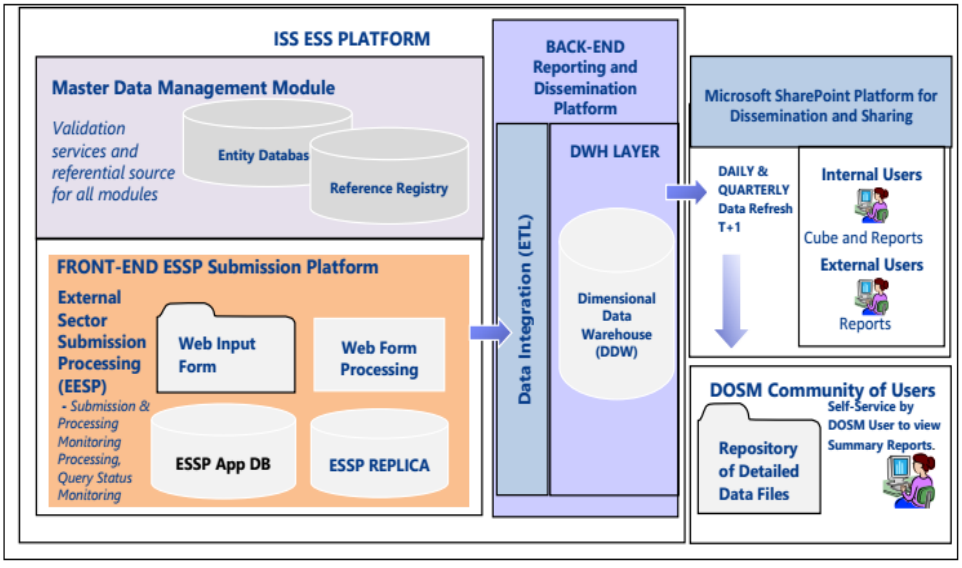
To ensure comprehensive coverage, the compilation guidelines clearly define the reporting methods for the different types of reporting entities. Listed below are the coverage of the data sources for the quarterly reporting:

i. Data on holdings of *securities issued by non-residents in the international market* are obtained from both the end-investors and custodian agencies, with some form of declaration to eliminate double counting.
ii. Data on *debt securities (bonds) issued by residents in foreign currency and sold in the international market* are obtained from the end-investors (the issuer).
iii. Data on *resident holdings of portfolio investment issued by another residents in the international market* are reported by the resident subscriber and are collected for the purpose of eliminating the transactions between residents.

2.4 System Architecture

The Central Bank implemented the Integrated Statistical System for External Sector Statistics (ISS-ESS) in 2017, the online reporting system, replacing the existing system which was in place since year 2002. The new ISS-ESS is envisioned to enable expansion of scope and coverage of the external sector data compilation to facilitate ease of production of various reports, to meet the diverse and changing data demand across internal and external users and as a compilation system that adopts best practices based on national and international reporting standards. The system architecture of the end-to-end solution for the ESS compilation is as illustrated in the diagram below.

ISS-ESS System Architecture — Diagram 2

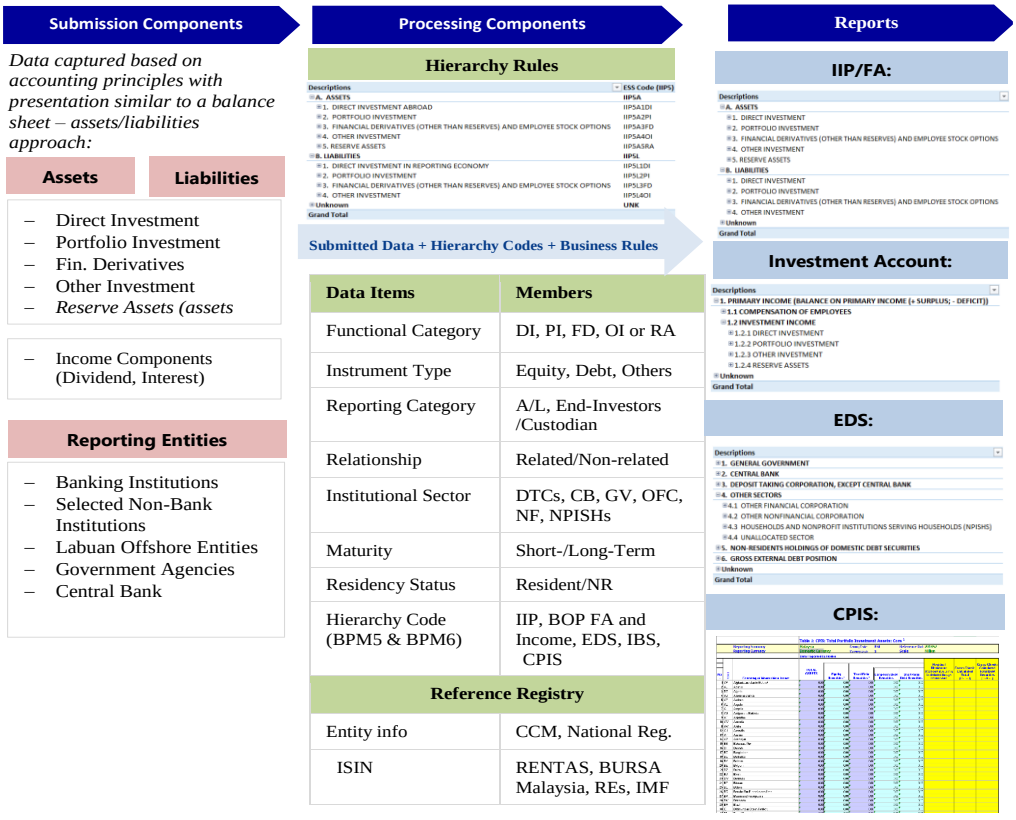


Through the front-end (Submission) modules, the respondents are required to report their external financial assets and liabilities exposures vis-

à-vis their non-resident counterparties following an accounting approach, whereby the reporting is done similarly to a financial statement with further categorisation by direct investment, portfolio investment, financial derivatives and other investment. In addition to the compilation notes and manual, the front-end modules are also equipped with build-in business validation rules to alert and/or prevent on possible misreporting by both the respondents and compilers. A Query Module is also established as a medium between the respondents and compilers to rectify and obtain confirmation on issues relevant to data accuracy.

The front-end modules will process all survey submissions by respondents to prepare data required by the back-end (Dissemination) modules. Once the process is completed, data will be transmitted to the back-end modules for further processing by the end of the day. During the transmission process, data integration programs (Extract, Transform and Load or ETL) are initiated in two stages, with the first stage focusing to harmonise and link the survey data to the Master Data Management (MDM) module (standard codes, reference entity information and ISIN profiles) and second stage mapping the data according to the business rules to produce the required ESS reports. The overall processing workflow are represented in diagram below.

ISS-ESS System Process Flow — Diagram 3



In view of the complex structures of external sector statistics and changing needs users, the MDM is designed to allow maintainability of business rules and centralised storage for reference data, such as standard codes, entity information and ISIN profiles, to cater for any new development in the international reporting standard. Upon completion of ETL linking the survey submissions to the master data, the system will continue the mapping process of transforming the granular data into reports, based on the hierarchy codes which are designed to be aligned with the international Guidelines as follows:

i. Balance of Payments and International Investment Position Manual, Fifth (BPM5) and Sixth (BPM6) Editions for the BOP Income and Financial Account and the IIP;
ii. Coordinated Portfolio Investment Guide for CPIS;
iii. External Debt Statistics: Guide for Compilers and Users for EDS; and
iv. BIS Guidelines for Reporting IBS.

3. Compilation Challenges and Moving Forward

For the compilation of portfolio investment in the domestic market, difficulties arise to obtain accurate information when the change of securities' ownership are not fully settled through domestic custodians. These issues are pertinent, particularly for ringgit debt securities, and pose increasing challenges faced by the compilers in view of more sophisticated players, mainly from the European markets. Nevertheless, as the aggregated stock by custodian banks in RENTAS are transparent to the Central Bank, this has helped the compilers to closely monitor the statistical reporting and ensure the accuracy of the financial flow data for BOP compilation.

On the other hand, collecting information on portfolio investment abroad has been hindered by insufficient details, and constraints imposed to limit the reporting burden faced of the reporting entities. In addressing this issue, the system was designed to provide flexibility to capture maximum information that can be obtained from reporting entities, guided by the reporting Guidelines. In addition, Malaysia has subscribed to the IMF initiatives for sharing of ISIN information, mainly the ISIN codes, Country and the Institutional Sector of the issuer through the secured channel IMF Box. Under this arrangement, which is in pilot run, the participating countries are required to share with IMF the ISIN codes for international securities subscribed by their respective residents. IMF will then collate all ISIN submitted by the participating countries and return these ISIN to the issuing countries to obtain the accurate Institutional Sector. The updated information is stored in the IMF's ISIN database, and subsequently shared with the participating countries of the securities subscribers. This will help all participating countries to enhance the quality of the data on portfolio investment abroad, as well as the CPIS data submission to IMF.

Data accuracy has always been the biggest challenge faced by the compilers. In this regard, significant efforts are required to maximise the accuracy on data compilation by working towards having more granular analysis for data quality checking. In addition, regular engagement with reporting entities are essential to ensure accurate understanding on reporting requirements. This has helped the Central Bank to meet the mandate to deliver the official external sector statistics, including the statistics on portfolio investment, within 7 weeks after end of reporting quarter based on primary data collection with 100% response rate.

In addition, the external sector compilation system requires design that meet the long-term needs in order to continuously comply with the international standard. In this regard, Malaysia adopted the integrated compilation system which provided flexibility in handling data submission, processing, as well as data dissemination through single platform. Of significance, the flexibility provided by the MDM allows the compilers to incorporate changes by or additional requirements from IMF and BIS without having to undertake major enhancement to the compilation system.

4. Conclusion

Malaysia has made a significant progress in the compilation of external sector statistics. The Integrated Compilation System has successfully reduced the reporting burden faced by both the compilers and the reporting entities. In view of the favourable responses, the submission deadline was shortened to 15 days after the end of each reporting quarter, compared with 20 days previously. Currently, Malaysia publishes the official BOP and IIP statistics by DOSM and the EDS and IBS by Central Bank respectively, 7 weeks after the end of each reporting period as compared to 12 weeks previously, with data obtained through a single platform to ensure consistency across all statistics.

Additionally, Malaysia is one of the earliest participants in the CPIS submission to IMF, for both mandatory and encouraged items, as well as the recent initiatives on sharing of ISIN information. At present, Malaysia is submitting the CPIS on a half yearly basis for all tables, covering the granular data, i.e. by instruments types, institutional sector and economy of the issuer/holders for both portfolio investment assets and liabilities respectively.

References

1. CPIS Guide Third Edition
2. External Sector Statistics (ESS) System – Submission of International Transactions and External Position Information

Financial integration and globalization: challenges and opportunities for external financial statistics

Carol C. Bertaut, Beau Bressler, Stephanie Curcuru
Division of International Finance at the Board of Governors of the Federal Reserve System

Abstract

The official framework for external financial statistics is according to the concept of *country of legal residence*. This concept provides increasing challenges for both data compilers and end-users, because the growing use of low-tax jurisdictions as locations for firm headquarters and the proliferation of offshore financing vehicles means there is an increasing disconnect between legal residence and actual economic exposures. The growing size, number, and geographic diversity of multinational firms generates additional challenges, as corporate events involving multinationals can leave large imprints on official statistics. Residence-based statistics can thus distort conclusions drawn about investment motivations and the true extent of external exposures. However, new data sources and "big data" techniques can provide new opportunities for data compilers to combine data sets and generate additional presentations of external statistics in ways that can help us better understand investor exposures and financial linkages in our increasingly interconnected world. *The views expressed are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.*

Keywords

Financial Integration, Multinationals, International Portfolios, Tax Havens

1. Introduction

After the global financial crisis, the G20 supported several efforts to produce improved global capital flow and investment statistics, with the goal of a better understanding of cross-border linkages and investor exposures. With respect to cross-border portfolio investment – cross-border flows and positions in bonds and equity – these initiatives included increased participation in the IMF's Coordinated Portfolio Investment Survey (CPIS), and efforts to increase both frequency and granularity of the CPIS, including providing detail on issuer and investor sectors. However, these efforts to improve the coverage of global portfolio assets and liabilities are not sufficient to gain a thorough understanding of global capital movements due to an ongoing fundamental limitation: These statistics use the official balance of payments (BOP) framework

that collects cross-border flows and positions according to *legal residence.* This legal residence concept is increasingly uninformative in a world of increasing globalization and growing usage of offshore financial centers and low-tax jurisdictions, because there is an increasing disconnect between the legal residence and the economic exposure. Firms issuing securities may not do any business at that legal residence, and thus ownership of such securities may say little about the actual exposures investors face. Thus, even if efforts to improve the country coverage are successful, official statistics will still provide an increasingly distorted view of linkages and economic exposures.

Two main aspects of firm behavior lead to the distortions between country residence and economic exposure that we observe in the cross-border portfolio data. First, when able, multinational firms frequently locate in the jurisdiction with the lowest tax rate; this is especially relevant for firms with substantial intangible and other portable assets.[1] As a result, cross-border statistics show elevated holdings of securities from the Cayman Islands, Ireland, and other low-tax jurisdictions which are associated with neither firm production nor expenses. Such distortions are not new: Schlumberger, long one of the largest 100 global firms, has operated in the U.S. since the 1930's and is headquartered in Houston, Texas, but has been incorporated in Curaçao since 1956.[2] As a result, U.S. cross-border statistics have shown large holdings of Curaçao equity for some time. However, distortions have recently become more pronounced following a wave of cross-border mergers and corporate "inversions", whereby former U.S.-resident firms have become foreign-resident firms after the merger.[3] For example, following recent high-profile inversions in the pharmaceutical industry such as Actavis/Allergan and Medtronic/Covidien, the equity of several major U.S. firms is now considered "Irish" equity according to official statistics. Adding to these distortions is the increasing presence of emerging market economy (EME) firms incorporated in the Caribbean. This is especially notable for some large-cap Chinese firms including Alibaba, Baidu, and Tencent.

A second driver is firms seeking to improve their access to capital markets and the pool of global bond investors. Many firms, particularly those in EMEs, issue corporate bonds using a subsidiary firm located in a market outside their home country. For this debt, the residence-based statistics will attribute transactions and securities holdings to the location of the subsidiary. Factors driving the use of offshore subsidiaries include improved pricing, access to

[1] See for example the survey on the tax competition literature in Keen and Konrad 2013, as well as Desai, Foley, and Hines 2006, Hebous and Johannesen 2016, Pomeroy 2016, Devereaux and Vella 2017.

[2] http://www.fundinguniverse.com/company-histories/schlumberger-limited-history/

[3] "Inversions" refer to M&A activity where the acquiring firm is typically larger than the target firm. After the merger, the combined firm "inverts" to establish its residence in the country of the target firm, which is typically a lower-tax jurisdiction.

foreign investors, and the ability to issue larger, lower-rated or longer-maturity bonds.[4]

The growing importance of mutual funds as a vehicle for cross-border investment provides a third source of distortions in the official statistics. Under international standards, holdings of investment fund shares are classified as equity holdings and are assigned to the country of fund incorporation. These standards apply regardless of the focus of the investment fund in terms of either the type of assets that the fund invests in or country of investment focus.

That traditional residence-based statistics are do not adequately represent exposures is gaining increasing recognition. For example, the Bank for International Settlements (BIS) now publishes its statistics on international debt securities on both a residence and a nationality basis, highlighting the rapid growth of issuance via offshore financial centers. Similarly, the world's largest sovereign wealth fund, the Norwegian Government Pension Fund, lists its roughly $1 billion portfolio holdings on both a country of incorporation (residence) basis and on a country of exposure basis.[5] In the academic community, Lane and Milesi-Ferretti (2017) provide an overview of the distortionary effects of increasing offshore issuance and financial center intermediation on properly assessing external exposures.

2. Methodology

We use the U.S. cross-border portfolio as a case study to document the extent of distortions in traditional residence-based statistics. With cross-border holdings of $12.4 trillion as of end-2017, the United States in aggregate is the single largest cross-border investor, reflecting holdings of a wide and diverse set of investors. For our study, we exploit the underlying security-level data on U.S. cross-border portfolio holdings collected on a legal residence basis for construction of the U.S. balance of payments statistics. Using security-level identifiers as well as modern text matching techniques,[6] we map these holdings, security by security, to the country of exposure for each firm assigned by commercial investment products designed for international investors, thus converting these holdings to a nationality basis. For common stock equity holdings, we rely primarily on MSCI constituent information, supplemented with information on the primary

[4] See for example Black and Munro (2010). Serena and Moreno (2016) identify a pickup in offshore issuance by firms in EMEs following the global financial crisis, which they attribute to declining financing costs and the less developed state of EME financial markets more generally. However, since the Asian Financial Crisis in the late 1990s has been away from offshore issuance, which is generally denominated in hard currencies, toward local-currency issuance in the domestic bond market (Black and Munro 2010, Mizen et al 2012, Hale et al 2016).

[5] https://www.nbim.no/

[6] See Cohen et al. (2018).

location of operations for firms that are not included in the MSCI.[7] For bonds, we additionally rely on Moody's information about the parent company and, for asset-backed securities, about the underlying assets to map holdings of corporate bonds to an "ultimate parent" or nationality basis.[8] Finally, we draw implications for distortions created by U.S. cross-border fund shares using "mirror data" on the portfolio assets of two countries that account for the majority of U.S. cross-border fund share holdings, the Cayman Islands and Luxembourg.[9]

3. Results

Figures 1a and 1b show the evolution of U.S. holdings of foreign common stock according to standard residence-based country attribution (1a) and nationality-based attribution (1b). A growing source of distortion arises from firms that the MSCI classifies as "U.S." but are legally incorporated outside the United States. This share has grown from about 7 percent of foreign common stock held in 2005 to about 13 percent (nearly $1 trillion) by 2017, partly as a result of the corporate inversions noted above. U.S. holdings of EME equity are also considerably larger by MSCI definitions, in large part reflecting the classification to EMEs of large Chinese firms incorporated in offshore centers. Overall, we find that by 2017, roughly $1.8 trillion–nearly a fourth–of U.S. holdings of foreign common stock is attributed by official statistics to a country different from the country assigned by MSCI.

Figures 2a and 2b similarly show the evolution of the U.S. cross-border bond portfolio. On a residence basis (2a), U.S. holdings of foreign-issued debt securities have risen from $557 billion in 2001 to about $2.8 trillion in 2017. By 2017, roughly 30 percent of these (nearly $850 billion) consisted of securities issued out of offshore centers or low-tax jurisdictions, an increase from roughly

[7] We assign the ultimate MSCI designation for securities of companies that have not yet been included in the MSCI. For example, we assign any U.S. holdings of Chinese firms such as Alibaba, Tencent, and Baidu (incorporated in the Cayman Islands) to China for years prior to 2015, although these firms were not included in the MSCI China/Emerging Markets indexes until 2015.

[8] Although sovereign bonds of many countries are issued as international debt securities, their country assignment will not be distorted in residence-based statistics the same way that corporate bonds are, because they are not issued via subsidiaries legally incorporated in offshore financial centers.

[9] Beginning in December 2015, the Cayman Islands submission to the CPIS includes securities holdings of resident funds. Cross-border portfolio holdings of the Cayman Islands were roughly $1.9 trillion as of December 2017, with a little over $1 trillion in debt securities and the remainder in cross-border equity. About 70 percent of these holdings are of U.S. securities, 15 percent are securities issued by other advanced economies, and the residual 15 percent those of all other countries, including EMEs. Similar information on assets of non-monetary Luxembourg funds is available from the Central Bank of Luxembourg. Securities held by these funds were more than $4 trillion at end-2017. Nearly a quarter of the underlying assets held by these funds are U.S. securities and another quarter are securities issued by non-euro area countries including EMEs.

20 percent in the early 2000s. Holdings of EME debt account on average for about 20 percent of U.S. holdings of foreign debt. On a nationality basis (2b), holdings of foreign debt securities have risen less, reaching only about $2.4 trillion in 2017. The lower value largely reflects increased issuance by financing arms of U.S. corporations established in offshore centers. Importantly, these holdings of U.S.-parent bonds include substantial investments in asset-backed securities issued out of Cayman Islands financing vehicles, including securities backed by U.S. mortgages in the run-up to the financial crisis and more recently, CLOs backed by U.S. syndicated loans. On the other hand, U.S. holdings of the debt of some other countries and regions are substantially understated. In particular, EME debt holdings are notably larger on a nationality basis and have grown faster in recent years. By 2017, our estimate of U.S. investment in EME debt securities on an ultimate parent basis is about 20 percent higher than under the residence-based statistics. Overall, we estimate that offshore issuance currently distorts the geography of more than $700 billion in U.S. cross-border debt holdings.

We further estimate that of the roughly $1.2 trillion in U.S. holdings of foreign fund shares, nearly $1 trillion is distorted in either asset type (that is, where underlying securities are bonds or other assets other than equity), or country of exposure, or both. Indeed, we estimate that at least half of these U.S. investor holdings actually reflect exposure to the United States.

Combining our findings for U.S. cross-border investment in bonds, common stock, and fund shares, we estimate that nearly $3.5 trillion of the total $12.4 trillion in foreign portfolio securities held by U.S. investors in 2017 reflects exposures to countries other than as reported in the official U.S. statistics. In contrast, in 2005, only a little over $800 billion of U.S. holdings of foreign securities reflected investment in a different country of exposure.

4. Discussion and Conclusion

Our results can be generalized to draw conclusions about the extent of global distortions. We estimate that roughly $10 trillion – about one-fourth – of the stock of global cross-border portfolio investment is similarly distorted in the current statistics. In particular, we estimate that global holdings of EME bonds and equity in the CPIS are understated by roughly $1.5 trillion, reflecting both corporate bonds issued via offshore financing arms and the growing market cap of emerging market firms incorporated in offshore centers. Global holdings of U.S. securities are also understated, owing to the incorporation of U.S.-based multinationals in low-tax jurisdictions as well as the investments of funds located in offshore centers. Securities holdings of other advanced economies, including Germany, Italy, and Spain are also likely understated, because their firms frequently issue debt securities via Luxembourg and Netherlands financing arms.

These findings have implications for understanding the factors influencing capital flows. For example, there has been much focus on the global impact of the extraordinary policy actions undertaken by advanced economy central banks in the wake of the global financial crisis. Of particular emphasis has been how these monetary policies spill over into emerging markets and how EME asset prices will react when these policies are reversed (Bowman et al 2015, Fratzscher et al 2018, Curcuru et al 2018). Our results showing understated growth in holdings of EME assets also imply mismeasurement of capital flows to EMEs. Overall, flows appear to have been stronger when policy was especially accommodative, which suggests that the spillovers may be understated.

Our results also weaken the argument that capital flows arising from foreign direct investment (FDI) are generally preferable because they are less volatile than portfolio flows, in part because FDI is harder to expropriate (Albuquerque 2003) and is driven by pull rather than push factors (Eichengreen et al 2018). However, these arguments assume that portfolio flows in the BOP accounts fully capture investment in a country's securities. When foreign residents buy bonds issued onshore, these purchases will show up as portfolio investment inflows. When corporations issue bonds via offshore affiliates, however, funds borrowed through the offshore entities are funnelled back to the parent firm in the form of lending or "reverse investment" in the parent firm. These flows, which will appear as FDI inflows, are effectively no different from typical portfolio flows, and can be just as volatile. Growing reliance on offshore financing vehicles for debt issuance can thus confound our understanding of the resilience of different types of cross-border financial flows. Similarly, our results also raise some potential flags for interpreting conclusions on the effectiveness of capital controls in preventing portfolio inflows to emerging markets (Forbes and Warnock 2012; Ahmed and Zlate 2014; Forbes et al. 2014; Forbes et al 2015; Pasricha et al. 2015). Foreign investors may still be able to gain exposures to countries via offshore-issued bonds, which typically are unaffected by controls. But because such purchases are not classified as portfolio inflows to these countries, the effectiveness of the controls may be overstated.

Our results are also relevant to the long-standing Lucas (1990) paradox, which arises from differences between the theoretical prediction of movements between developed and developing countries, and what is observed. Theory predicts that capital should move toward economies with lower levels of capital per worker. Contrary to this theory, most studies find that capital does not flow from more to less developed economies; rather, it flows in the other direction (see Alfaro et al. 2008, among others). Our results suggest that advanced economy exposure to EMEs is larger than previously believed, which resolves some portion of this puzzle. This is perhaps especially evident when we consider the global reaches of large multinational firms. In

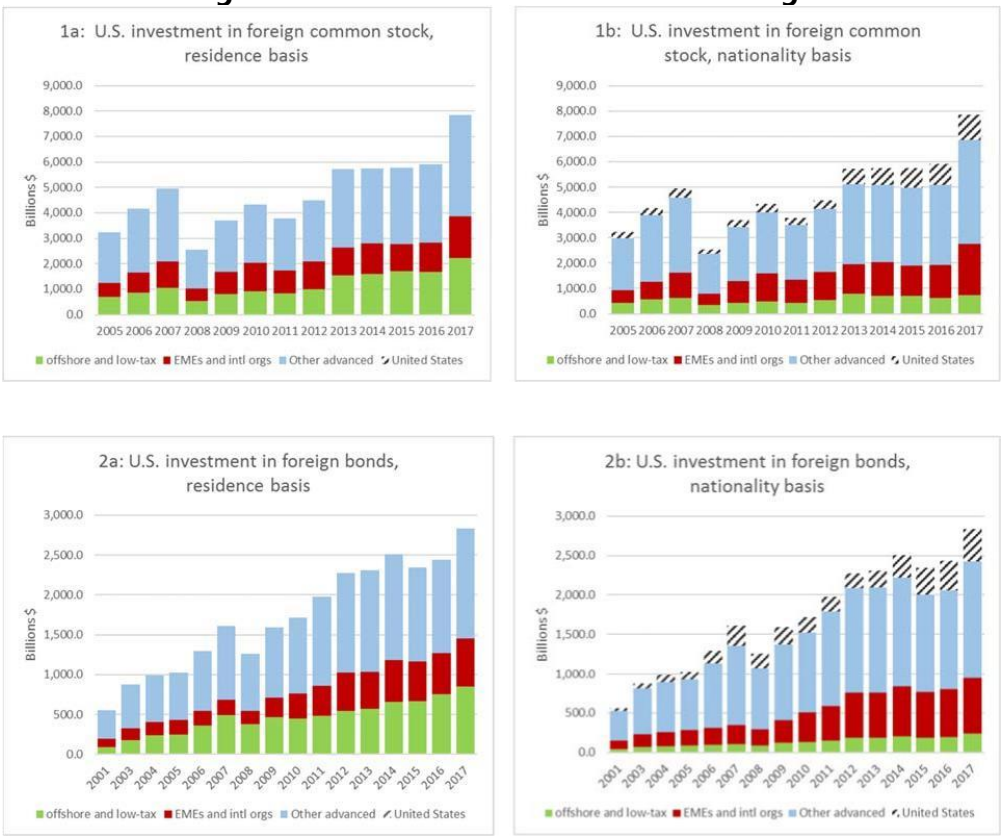
ongoing research, we factor in the global activity of such large firms to measure U.S. portfolio exposures more broadly.

References

1. Ahmed, S. and A. Zlate (2014), "Capital flows to emerging market economies: A brave new world?", Journal of International Money and Finance, Vol. 48(PB), pp 221-248.
2. Albuquerque, R. (2003). The composition of international capital flows: risk sharing through foreign direct investment. *Journal of International Economics*, *61*(2), 353-383.
3. Alfaro, L., Kalemli-Ozcan, S., & Volosovych, V. (2008). Why doesn't capital flow from rich to poor countries? An empirical investigation. *The Review of Economics and Statistics*, *90*(2), 347368.
4. Black, S., & Munro, A. (2010). Why issue bonds offshore? BIS Working Paper No. 334.
5. Bowman, D., Londono, J. M., & Sapriza, H. (2015). US unconventional monetary policy and transmission to emerging market economies. *Journal of International Money and Finance*, *55*, 27-59.
6. Cohen, Gregory J., Melanie Friedrichs, Kamran Gupta, William Hayes, Seung Jung Lee, W. Blake Marsh, Nathan Mislang, Maya Shaton, and Martin Sicilian (2018). "The U.S. Syndicated Loan Market: Matching Data," Finance and Economics Discussion Series 2018-085. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2018.085.
7. Curcuru, S., A. Rosenbaum and C. Scotti (2018). "International Capital Flows and Unconventional Monetary Policy" Working paper.
8. Desai, M. A., Foley, C. F., & Hines Jr, J. R. (2006). The demand for tax haven operations. *Journal of Public economics*, *90*(3), 513-531.
9. Devereux, M. P., & Vella, J. (2017). Implications of Digitalization for International Corporate Tax Reform.
10. Eichengreen, B., Gupta, P., & Masetti, O. (2018). Are capital flows fickle? Increasingly? and does the answer still depend on type?. *Asian Economic Papers*, *17*(1), 22-41.
11. Forbes K., M. Fratzscher, T. Kostka and R. Straub (2016), "Bubble thy neighbour: Portfolio effects and externalities from capital controls", Journal of International Economics, Vol. 99, pp 85-104.
12. Forbes K., M. Fratzscher and R. Straub (2015), "Capital-flow management measures: What are they good for?", Journal of International Economics, Vol. 96, pp S76-S97.
13. Forbes K. and F. Warnock (2012), "Capital flow waves: Surges, stops, flight, and retrenchment", Journal of International Economics, Vol. 88, pp 235–251.

14. Fratzscher, M., Lo Duca, M., & Straub, R. (2018). On the international spillovers of US quantitative easing. *The Economic Journal*, *128*(608), 330-377.
15. Pasricha, G., Falagiarda, M., Bijsterbosch, M., & Aizenman, J. (2015). *Domestic and multilateral effects of capital controls in emerging markets* (No. w20822). National Bureau of Economic Research.
16. Gruić, B., & Wooldridge, P. D. (2012). Enhancements to the BIS debt securities statistics.
17. Hale, G., P. Jones and M. M. Spiegel (2016). "The Rise in Home Currency Issuance", Federal Reserve Bank of San Francisco Working Paper 2014-19
18. Hebous, S., & Johannesen, N. (2015). At your service! The role of tax havens in international trade with services
19. Keen, M., & Konrad, K. A. (2013). The theory of international tax competition and coordination. In *Handbook of public economics* (Vol. 5, pp. 257-328). Elsevier.
20. Lane, M. P. R., & Milesi-Ferretti, M. G. M. (2017). *International financial integration in the aftermath of the global financial crisis*. International Monetary Fund Working Paper No. 17/115.
21. Lucas, R. E. (1990). Why doesn't capital flow from rich to poor countries?. *The American Economic Review*, *80*(2), 92-96.
22. Mizen, P., Packer, F., Remolona, E. M., & Tsoukas, S. (2012). Why do firms issue abroad? Lessons from onshore and offshore corporate bond finance in Asian emerging markets.
23. Pasricha, G., Falagiarda, M., Bijsterbosch, M., & Aizenman, J. (2015). *Domestic and multilateral effects of capital controls in emerging markets* (No. w20822). National Bureau of Economic Research.
24. Pomeroy, James. 2016. "The Rise of the Digital Natives." HSBC report, September.
25. Serena, J. M., & Moreno, R. (2016). Domestic financial markets and offshore bond financing.

Figures: U.S. Cross-border Portfolio Holdings



Source: Authors' estimates based on Treasury International Capital data (https://www.treasury.gov/resource-center/data-chart-center/tic/Pages/fpis.aspx#usclaims).

Offshore and low-tax countries include Bermuda, the British Virgin Islands, the Cayman Islands, Curaçao (Netherlands Antilles until 2013), Guernsey, Ireland, Isle of Man, Jersey, Liberia, Luxembourg, the Netherlands, Malta, the Marshall Islands, Mauritius, Panama, and Switzerland.

Forecasting the recovery rate of non-financial corporations with particular emphasis on sectorial analysis

Natalia Nehrebecka
Narodowy Bank Polski

Abstract

The empirical literature on credit risk is mainly based on modelling the probability of default, omitting the modelling of the loss on default. This paper is aimed to study the recovery rate in theoretical approach - familiarizing with regulatory requirements, and also in practical approach - to predict recovery rates on the rarely applied here nonparametric method of Quantile Regression and Bayesian Model Averaging, developed on the basis of individual prudential and balance of payments data in the 2007–2018. Literature on Losses Given Default focuses on mean predictions, even though losses are extremely skewed and bimodal. The models were created on financial and behavioural data that present the history of the credit relationship of the enterprise with financial institutions. Two approaches are presented in the paper: Point in Time and Through-the-Cycle. Using the estimated risk parameter, the reserves for expected loan losses were also calculated. A correct estimation of LGD parameter affects the appropriate amounts of held reserves, which is crucial for the proper functioning of the bank and not exposing itself to the risk of insolvency if such losses occur.

Keywords

loss given default; recovery rate; regulatory requirements; quantile regression; bayesian model averaging

1. Introduction

Credit risk assessment (in particular ensuring accurate and reliable credit ratings) plays a key role for many market participants. According to the traditional approach the definition of credit risk, it is the risk of loss caused by a debtor's failure to repay a loan, while in the market definition it is the risk of loss driven by a rating downgrade (i.e. an increase in the probability of default) or failure to repay an obligation by a debtor. Basel Committee explains a default event on a debt obligation in the two following ways:

- It is unlikely that the obligor will be able to repay its debt to the bank without giving up any pledged collateral;
- The obligor is more than 90 days past due on a material credit obligation.

Basel II introduced the Internal Ratings-based Approach which enables institutions to provide their own estimates for the Loss Rate Given Default (LGD).

The Basel Committee on Banking Supervision (2005) points to the importance of adequate estimates for economic downturns and unexpected losses. **The purpose of this research is forecasting the recovery rate[1] of non-financial corporations with particular emphasis on sectorial analysis.** The company's industry and its characteristics have an impact on the loss given default. In addition, a research hypothesis has been put forward that assumes that enterprises operating in industries where the market is small have higher losses in the event of the company's insolvency due to the lack of active bidders' market. If the market is not liquid, it is more difficult for creditors to recover the amounts due, and the time may be increased until they are collected.

Research on the Loss Given Default began to gain momentum only in the 21st century. Most empirical work on LGD for loans began to arise after the introduction of the New Capital Accord in 2004. In the first works on modelling losses due to default [Altman, Gande and Saunders 2003; Arner, Cantor, Emery 2004; Cantora and Varmy 2010] linear regression was used. The Basel Committee on Banking Supervision (2005) points to the importance of adequate estimates for economic downturns and unexpected losses. Board of Governors of the Federal Reserve System (2006) proposes the computation of Downturn LGD measures by a linear transformation of means [*Dowturn LGD = 0,08 + 0,092 * E (LGD)*]. Most academic and practical credit risk models focus on mean LGD predictions. However if we consider two loans with different distributions (a uniform and a beta) but the same means values then we have real quantiles and downturns as well as unexpected losses differ. A relatively new method for modelling the loss given default, which was also used in this research, is quantile regression [Somersa and Whittaker 2007; Krüger and Rösch 2017]. While other methods only allow estimating the mean or variance of LGD, quantile regression allows modelling of all quantiles of the dependent variable. In this way, it is easy to obtain measures in the event of a downturn and unexpected losses.

The remainder of this paper is organized as follows. Section 2 presents the methodology. Section 3 presents and discusses the empirical results, while Section 4 concludes the paper.

2. Methodology

In order to calculate the LGD parameter, the Recovery Rate (RR) should be initially estimated. RR is defined as one minus any impairment loss that has occurred on assets dedicated to that contract (see IAS 36, Impairment of Assets) / Exposure at Default. Most LGDs are nearly total losses or total

[1] Recovery Rate given default is the part of the loan liabilities that a creditor can recover from the debtor in the event that the debtor defaults.

recoveries which yields to a strong bimodality. The mean is given by 37% and the median by 24%, i.e., LGDs are highly skewed. Both properties of the distribution may favour the application of quantile regression because most standard methods do not adequately capture bimodality and skewness. Furthermore, many LGDs are lower than 0 and higher than 1 due to administrative, legal and liquidation expenses or financial penalties and high collateral recoveries.
The regression model can be presented as follows (see Nehrebecka, 2019):

$$Recovery\ rate_{i,t} = f\ (Debt\ Characteristics_{i},\ Bank\ Characteristics_{i},\ Firm\ Characteristics_{i,t-1},\ Macroeconomic\ Variables_{t-1})$$

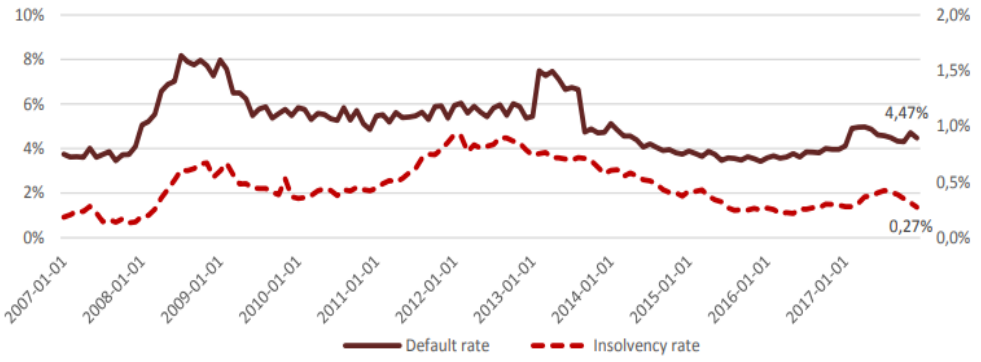
The LGD coefficient - was obtained as a result of nonparametric method of Quantile Regression and Bayesian Model Averaging. Two approaches are presented in the paper: Point in Time and Through the Cycle. The historical loan losses recorded by the National Bank of Poland were used for estimation.

3. Results

The empirical analysis was based on the individual data from different sources (from the years 2007 to 2018), which are:

- Data on banking defaults are drawn from the Prudential Reporting managed by Narodowy Bank Polski. Act of the Board of the Narodowy Bank Polski no.53/2011 dated 22 September 2011 concerning the procedure and detailed principles of handing over by banks to the Narodowy Bank Polski data indispensable for monetary policy, for periodical evaluation of monetary policy, evaluation of the financial situation of banks and bank sector's risks. Large exposures – for a bank that is a joint-stock company, state-run bank and a non-associated cooperative bank – mean exposures towards one enterprise in excess of 2,000,000 PLN.
- Data on insolvencies/bankruptcies come from a database managed by The National Court Register, that is the national network of Business Official Register.
- Financial statement data (source: AMADEUS, NOTORIA, BISNODE, F-02).
- Data on external statistics of enterprises (source: Narodowy Bank Polski).

Figure 1: Insolvency rate and default rate in period 2007-2017



Source: Authors' calculations

During the crisis on global financial markets in 2007-2009 a decline in the GDP growth rate was recorded from 6.6% to 3.2% and the number of declared bankruptcies in the economy increased by 54.6%. At the turn of 2008/2009, the default rate was at the level of 8%. In 2012 the courts declared the bankruptcy of 877 business entities, which was the highest result for 8 years. This state of affairs can be partly explained by the economic downturn in 2012. In the case of the default rate, the second local maximum (7.5%) was noticed. The default rate this year is still declining and is at around 4.5% (Figure 1).

Table 1: Banks' credit exposure to non-financial firms by instrument in 2018

Instrument Type	Exposure in mld PLN	N firms	N banks	HHI by firms	HHI by banks	HHI by sectors
Total explosure	477,4	17 119	44	0,2%	9,7%	16%
Loans	276,0	15 598	44	**0,15%**	**9,2%**	7,6%
Bonds	7,7	84	11	**12,1%**	**32,5%**	**25,9%**
Guarantees	51,1	4 195	32	1,4%	11%	15,4%
Open credit lines	118,9	12 028	35	0,3%	11,1%	7%

Source: Authors' calculations

Table 1 presents the banks' exposure to the non-financial corporation sector in 2018 divided into balance sheet exposure, including: loans and other receivables, debt instruments and off-balance sheet exposure, including: guarantee, open credit lines. The concentration index (HHI) was calculated for each financial instrument, both in terms of lenders and borrowers. Banks' exposure to the corporate sector was highest in loans and other receivables (58% of total exposure). In addition, in the case of loans and other receivables, the concentration index assumed the lowest level in terms of both lenders and

borrowers. In contrast to debt securities that represent concentrated markets, with several banks holding large shares of debt securities. Over half of the off-balance sheet exposure was, however, open credit lines.

For the comparison, I consider two models: (1) Quantile regression (QR) and (2) Bayesian Model Averaging (BMA). Table 2 shows the estimation results of QR for the 5th, 25th, 50th, 75th and 95th per centile and the corresponding BMA estimates.

Table 2: Parameter estimates Quantile Regression and Bayesian Model Averaging

Variables	PIT					
						BMA
	Q(0.05)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.95)	Post Mean (Post SD) PIP
ln_EAD	0.0028*** (0.0002)	0.000192 (0.0003)	-0.0048*** (0.0003)	-0.0061*** (0.0002)	-0.0004*** (0.0000)	-0,0031 (0,0004) 1
Collateral Indicator (No, Yes)	0.0673*** (0.0002)	0.163*** (0.0003)	0.117*** (0.0003)	0.0491*** (0.0002)	0.0011*** (0.0000)	0,1213 (0,0013) 1
Days Past Due (less than 90 days)	-0.335*** (0.0032)	-0.120*** (0.0048)	-0.0720*** (0.0051)	-0.0361*** (0.0031)	-0.0054*** (0.0003)	-0,0699 (0,0053) 1
Days Past Due (more than 90 days)	-0.506*** (0.0019)	-0.343*** (0.0029)	-0.178*** (0.0030)	-0.0595*** (0.0018)	-0.0072*** (0.0002)	-0,224 (0,0032) 1
Time spent in default (months))	-0.0808*** (0.0005)	-0.0114*** (0.0007)	-0.00277*** (0.0008)	0.00175*** (0.0004)	0.0000 (0.0000)	-0,01 (0,0015) 1
	0.00913***	0.0381***	0.0366***	0.0122***	-0.0000	0,0353

Loan type (PLN vs other currency)	(0.00095)	(0.0014)	(0.0014)	(0.0009)	(0.0001)	(0,0011) 1
Guarantee Indicator (No, Yes)	0.00366** (0.0013)	0.0185*** (0.0019)	0.0319*** (0.0020)	0.0183*** (0.0012)	0.0008*** (0.0001)	0,0246 (0,0021) 1
Credit lines (No, Yes)	0.214*** (0.0009)	0.255*** (0.0014)	0.159*** (0.0014)	0.0728*** (0.0008)	0.0013*** (0.0001)	0,1811 (0,0015) 1
Bank firm relationship	0.00334*** (0.0002)	0.0033*** (0.0003)	0.00328*** (0.0003)	0.0026*** (0.0002)	0.0001*** (0.0001)	0,0038 (0,0004) 1
Age of firms	0.00138*** (0.0000)	0.00267*** (0.0000)	0.00241*** (0.0000)	0.000830*** (0.0000)	0.0000* (0.0000)	0,003 (0,0001) 1
Size bank	0.0371*** (0.0018)	0.109*** (0.0026)	0.0720*** (0.0028)	0.0302*** (0.0017)	0.0019*** (0.0001)	0,0769 (0,0024) 1
Intercept	0.515*** (0.0032)	0.693*** (0.0046)	0.907*** (0.0049)	1.054*** (0.0030)	1.0110*** (0.0003)	0,8393 (0,0012) 1

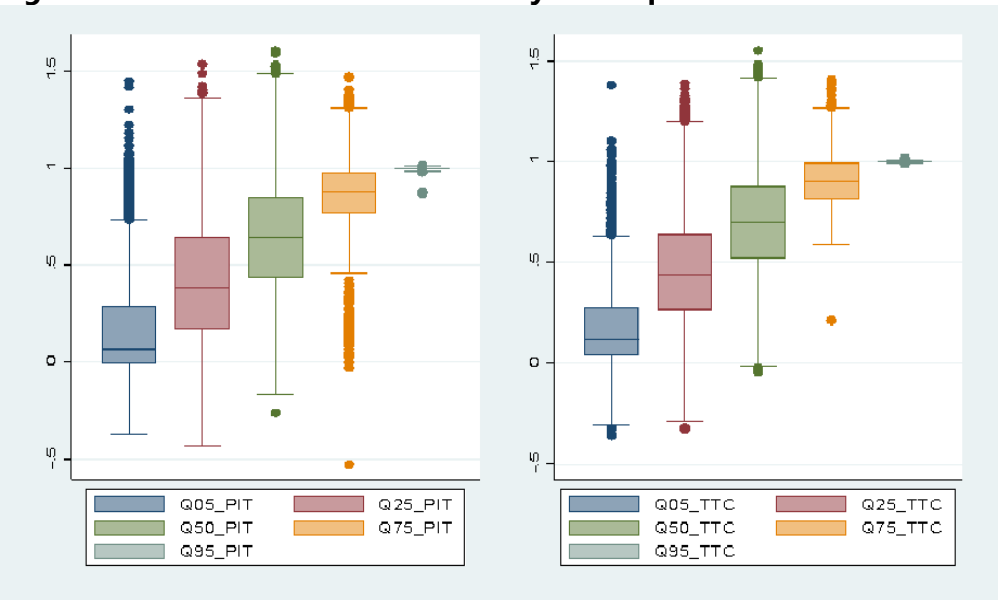
Note: Standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001. Additionally, models for all banks are estimated containing dummy variables for the different banks in addition to the variables mentioned below. Sectors, legal form of companies are also included in models.

Source: Authors' calculations.

At the time of the economic downturn, the chance for the bank to recover completely is decreasing (in the case of LGD = 0), average losses and a chance for a total loss (LGD = 1) increase. Also calculated were 5%, 25%, 50%, 75%, 95% VaR for LGD during the business cycle and during the economic downturn. For the probability of exceeding the loss of 5%, 25%, 50% and 75%,

the VaR value is definitely lower for the Dowturn Recovery Rate than for the Through-the-Cycle Recovery Rate. The LGD estimate was also compared during the economic slowdown proposed by the US central bank and estimated by quantile regression. It can be concluded that Dowturn LGD, calculated as a linear transformation of the average LGD value, is overestimated for low LGD values and underestimated for high LGD values. In summary, the loss calculated due to default in the period of downturn does not catch unexpected risks caused by the bimodal distribution of LGD.

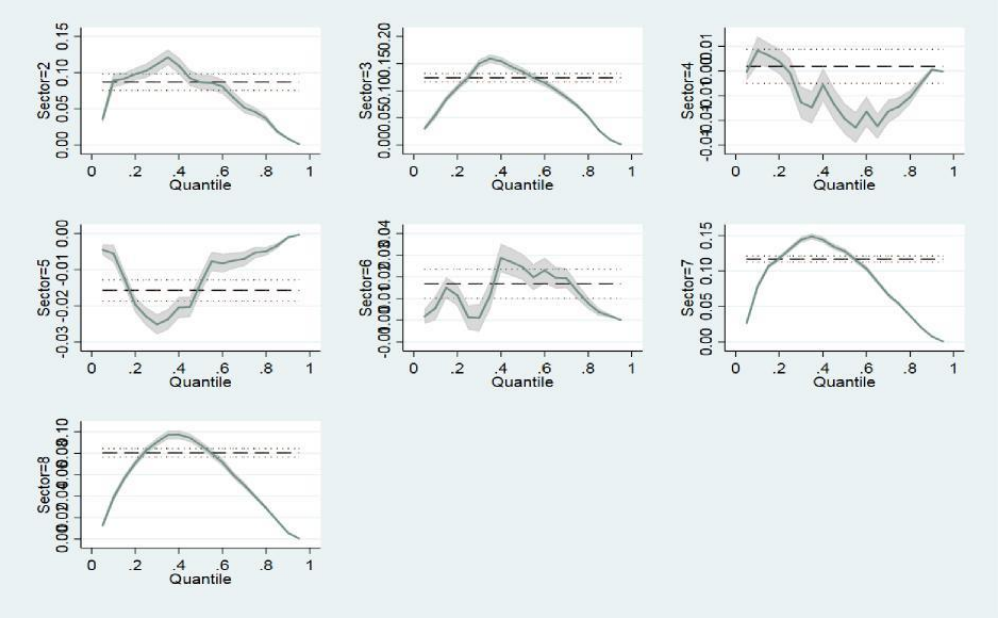
Figure 2: Downturn effects on Recovery Rates quantiles



Source: Authors' calculations

It was observed economic sector effects manly in first quartile (Figure 3). The affiliation may cause a variation up 15 percentage points with lowest Recovery Rate for trade sector (section G) and highest values for real states (section L) as well as energy sectors (section D, E). In contrast, the OLS results are misleading, because the trade sector is not significant. It seems to be the safest economic activity from the creditor perspective of the obligor's repayments. Companies belonging to industries with a high level of concentration and specialization in the event of becoming insolvent may have problems with the sale of their own production assets due to a low liquid market. Undoubtedly, this would have a negative impact on the position of creditors not only by reducing the amount of recovered amounts, but also by postponing the time of their collection. In addition, it is worth noting that if the assets of an insolvent enterprise are so specific that they are not suitable for use in another industry, then the difficulties with their sale at the time of insolvency may increase the poor condition of the enterprise sector.

Figure 3: Estimated coefficients for Recovery Rate using Quantile Regression and OLS along with 95% confidence intervals for economic sectors[2]



Source: Authors' calculations

4. Discussion and Conclusion

In order to estimate the risk parameter LGD in a suitable manner must meet a number of requirements imposed by the regulator. The main aspects are the right approach to the default definition - consistent within all credit risk parameters, creating a reliable reference data set, based on which the LGD is estimated, considering all historical defaults in modelling and selecting the right modelling method. It is necessary to verify and validate the methods of estimated losses due to defaults and to correct any discrepancies. Validation should pay attention to compliance with regulatory requirements as well as the correctness of the estimated parameters and the predictive power of models. Correct estimation of the LGD parameter affects the maintenance of adequate capital for expected credit losses, which is a key element of the bank's operation.

[2] Sector=2: "B" - Mining and quarrying; Sector=3: "DE" - Energy, water and waste; Sector=4: "F" – Construction, Sector =5: "G" - Trade, Sector=6: "H" - Transportation and storage, Sector=7: "L" - Real estate activities, Sector=8: "Others".

References

1. Altman. E., Gande. A., Saunders. A. (2010). Bank Debt Versus Bond Debt: Evidence from Secondary Market Prices. Journal of Money, Credit and Banking. Vol. 42. No. 4. pp. 755-767.
2. Arner. R., Cantor. R., Emery. K. (2004). Recovery Rates on North American Syndicated Bank Loans, 1989-2003. Moody's Special Comments. Available at http://www. moodyskmv.com.
3. Cantor. R., Varma. P. (2004). Determinants of Recovery Rate and Loan for North American Corporate Issuers. The Journal of Fixed Income. Vol. 14. No. 4. pp. 29-44.
4. Krüger. S., Rösch. D. (2017). Downturn LGD modeling using quantile regression. Journal of Banking & Finance. Vol. 79. No. 1. pp. 42–56.
5. Nehrebecka. N. (2016). Approach to the assessment of credit risk for non-financial corporations. Evidence from Poland. IFC Bulletins chapters, in: Bank for International Settlements (ed.), Combining micro and macro data for financial stability analysis, Vol. 41, Bank for International Settlements.

Uses of mirror data: Estimation of household assets with banks abroad

Swapan-Kumar Pradhan[1], João Falcão Silva[2]
[1]Senior Statistical Analyst, Monetary and Economic Department, BIS
[2]Head of Unit Financial account, BoP and IIP Statistics, Statistics Department, Bank of Portugal

Abstract[3]
This paper aims to analyse and estimate cross-border assets of households in the form of bank deposits and bank loans. Such data are scarce and there is no comprehensive system to collect and compile this information directly. The lack of available information combined with a complex delimitation of this institutional sector represent challenging issues to the compilers. The international locational statistics of the Bank for International Settlements (BIS) cover cross-border assets and liabilities of reporting banks broken down by counterparty sector in individual countries around the world. We apply mirror data approach to derive assets of households sector in a given country using source data as the cross-border liabilities of banks to this sector in respective countries. In addition, we apply our method to estimate data backwards for periods when International Banking Statistics (IBS) data for this sector are either limited from 2013Q4 or not available prior to 2013Q4.

Keywords
data gaps; foreign assets/liabilities; households; international banking; mirror data.

1. Introduction

In a more globalized world, the institutional sector "households" is a statistical challenge for the compilers due to non-availability of data or access to accurate data. We address this issue by focusing on cross-border assets/liabilities of this sector vis-à-vis foreign banks using the BIS locational banking statistics (LBS). We apply the mirror data approach which refers to complementary sources that capture similar concepts and is indeed a crucial statistical tool that allows to fill-in data gaps. The mirror data approach involves involve comparison of different statistical data sets that can be

[3] We thank Bruno Tissot and Philip Wooldridge of the BIS and Filipa Lima, Luís Teles Dias and Paula Menezes (Bank of Portugal) for their continued encouragement and support on mirror exercise to enhance the quality and coverage of data. We also thank Patrick McGuire (BIS) and all the contributions from then central banks, particularly colleagues in the Bank of England (Marek Rojicek) and Bank of Finland (Johanna Honkanen) for helpful comments, suggestions and discussions. The views expressed are those of the authors and do not necessarily reflect those of the Bank for International Settlements or the Bank of Portugal.

analysed within one country, or across/between countries aiming to compare the same statistical data under a dual perspective (eg creditors versus debtors). Falcão Silva, João & Pradhan, Swapan-Kumar (2018) demonstrated the importance of mirror data to enhance statistical quality as well as coverage of data across comparable statistical domains.

In the absence of data confidentiality restrictions, mirror data exercise is an important tool to improve the quality of the data, fill-in data gaps and reduce bilateral asymmetries. In the case of households, mirror data exercises can perform better estimates of their financial assets/liabilities because households do not disclose amount/location of their cross-border positions (assets/liabilities) either directly or through a survey[4]. As confidentiality constraints of bilateral data at granular level prevents knowing banks' location (BIS reporting countries), we also provide an estimation method at an aggregate level. Furthermore, we demonstrate that our methods provide better estimates of the households' cross-border positions in a given country, when bilateral data are disclosed for majority of reporting countries. We apply our methodology to the Portuguese data as a country-case example.

Finally, as the financial assets/liabilities of households is considered commonly as a statistical gap in the balance of payments/international investment position (BoP/IIP) and the rest of the world (RoW) accounts compilation, this approach could also support these two statistical domains.

2. Compilation of data for households sector – main challenges

The compilation of data for the household sector raises some difficulties related to the data availability and accuracy. One of the main issues is the accountability because there is no full set of accounts or ability to draw up sets of accounts for household sector[5]. Households' surveys constitute one source to surpass this issue. Nevertheless, some drawbacks are associated with non-responses, estimation or underreporting of their financial assets and income. System of National Accounts (SNA) 2008[6] states an example associated with people earning income arising from illegal activities who may be very reluctant to provide this information and may choose not to participate in the survey. Similarly, it is common for households at the very top/bottom of the distribution to be omitted from the survey either by design or on the grounds of practicality. Low frequency and long lag in data availability are other critical issues. Currently, there exists a lack of regular information on the households' assets and liabilities broken down by financial instrument types such as

[4] Such information from creditor/debtor sources (banks where their deposits are located) would be complementary source for the purpose

[5] that includes also non-profit institutions serving households.

[6] paragraph 24.24.

deposits, loans and securities. Data sources are scarce and obtaining accurate/comparable data is very difficult within each country and across different countries.

The issues relating to demographic changes cannot be ignored. When the population changes, an effect on households' well-being and resources is observable and, consequently, calls for policy-actions. For example, under an ageing population, there is less demand for educational services and more demand for health services[7]. Another concern is whether pension benefits are sufficient to support individuals in the retirement age without any government intervention. According to the SNA, a focus on such issues might suggest sub-sectoring households according to income earner categorization[8]. In addition, demographic patterns will likely put pressure on potential output growth rates, the natural rate of unemployment, and the long-term equilibrium interest rate and on the monetary policy transmission mechanism. The magnitude and the timing are uncertain as they depend on the behaviour of consumers and businesses. Rising fiscal imbalances are projected to lead to higher government debt-to-GDP ratios, potentially putting upward pressure on interest rates, and crowding out productive investment.

Finally, the greater international labour and capital movements, have significant implications on the international connections between resident households and non-residents cannot be ignored.

3. Methodology

We use the LBS as the main data source to perform this empirical exercise. These statistics are consistent with the BoP/IIP methodology, as they correspond to claims/liabilities of residents in one country vis-à-vis those of other countries. In addition, the LBS are best suited for macro analysis of economic and financial stability issues. The linkages with these and other statistical domains cannot be disregarded and should be part of the statistical analysis. The LBS data covers information on the financial instruments (eg loans and deposits and debt securities), currency, counterparty sector (e.g. intragroup, central banks, unrelated banks and non-banks) and counterparties' geographical composition of resident banks' balance sheets. While the LBS data capture the non-bank sector since 1977, the claims/liabilities of banks vis-à-vis subsectors of non-banks (households in particular), are available only from the end-December 2013.

[7] paragraph 24.42, SNA 2008.

[8] If households sector is in work, relevant categorisation of working age but not in work or in retirement (paragraph 24.43

There are some issues associated with the use of LBS as a data source. First, data on banks' liabilities[9] to the households are collected on an encouraged basis only from end-December 2013, which does not ensure full data coverage[10]. Secondly, only 13 countries started reporting these data from the end-December 2013 and another 17 countries started reporting in subsequent quarters[11]. Thirdly, due to confidentiality reasons data for some countries is not published and thus we cannot disclose all the estimates on bilateral data.

The following table shows the mapping of sector hierarchy between the SNA and BIS LBS and thereported liabilities (deposits) in the LBS as of Q4 2018:

LBS sector code/name	Amounts [in $bn]	SNA code
A: All sectors	**20,550**	**S1**
B: Banks, total (sub-sectors exists)	12,688	S121+S122
N: Non-banks, total	7,678	(S123+S124+S125+S126+S127+S128+S129) + (S11+S13+S14+S15)
F: Non-bank financial institutions	4,141	S123+S124+S125+S126+S127+S128+S129
P: Non-financial sectors	3,168	S11+S13+(S14+S15)
C: Non-financial Corporations	1,199	S11
G: General Government	89	S13
H: Households including NPISHs	673	**S14+S15**
K: Non-financial sectors, unallocated	1,208	
X: Non-banks, unallocated	369	
U: Unallocated by sector	184	

Although the counterparty sector breakdown into **F** and **P** (see Table above) are required, as of Q4 2018 about 5% of the total deposit liabilities to the non-bank sector was classified under the category "X: Unallocated non-banks". The further breakdown of **P** into subsectors **C**, **G** and **H** shows that at sub-hierarchy, the share of "K: Unallocated non-financial sector" is about 16%

[9] same for claims on households sectors but we focus on liabilities of banks to estimate assets of this sector with banks abroad.

[10] The non-financial subsectors were introduced in the BIS international banking statistics as part of a series of enhancements to the statistics starting from end-December 2013. National authorities started to report the enhanced data at different times and, as a result, global coverage has been incomplete during the implementation phase. The BIS does not yet publish data for these subsectors, but plans to release these data by end 2019.

[11] Currently 47 countries report LBS data. We thus deal with two types of estimations, one for 17 countries that started after 2013Q4 and another for remaining 17 countries that have not yet started reporting these data (countries with large cross-border positions in the latter group are China, Hong Kong SAR, Japan, Singapore and the United States).

of sector **N,** and 38% of sector **P** (i.e. at the aggregate level 38% of liabilities to sector **P** are not classified to sectors **C**, **G** and **H**)[12].

The interlinkages with other data sources using the mirror data is crucial when analysing the households sector. It offers the use of debtor banks' liabilities to derive assets of the households with banks. We consider BIS reporting banks' cross-border liabilities to the households sector as the measure of assets of this sector with banks abroad in two alternative methods: using aggregate and bilateral data.

We noted that the coverage of reported data differs across counterparty countries. In the overall aggregate as of Q4 2018, 38 of 47 countries reported subsectors **F** (54%) and **P** (41%) of sector **N** with coverage of 95% and remaining 5%, accounted by 9 countries, do not report any sub-sectors of sector **N**[13]. While the estimation of the remaining 5% into subsectors **F** and **P** could be done by various methods, a simple way is to use the proportional approach[14]. If this approach is adopted, the share of **F** and **P** would be 57% and 43% respectively. As shown in Graph 1 we estimate amounts from Q4 2011 (i.e. for period when no information on subsectors was available) for all subsectors of non-banks in all counterparty countries (top right panel) and of which those in Portugal (bottom right panel). These estimates are based on aggregated data of all reporting countries using simple proportional approach, both for the aggregate of all counterparty countries and of which those in Portugal[15].

I – Aggregate level estimations

When confidentiality for bilateral data restrictions arises, mirror data exercises can only be applied at an aggregate level. According to the LBS information each country can estimate its deposits abroad by using the reported information for the aggregate of 'all reporting countries' vis-à-vis sectors **F, P** and subsectors of **P** and in particular sector **H**, with possible alternatives[16]. This procedure provides estimated amounts for a given country

[12] Seventeen countries do not report subsectors of **P**: Bahrain, Brazil, Chile, China, Curacao, Finland, Greece, Hong Kong SAR, Japan, Jersey, Macao SAR, Mexico, Panama, Philippines, Singapore, Turkey and United States.

[13] Of these 9 countries, Singapore comprises about 60%, Jersey and Bahrain each comprise about 15% and rest by another 6 countries (Brazil, Chile, Curacao, Greece, Mexico and Panama).

[14] I.e. allocate 54/95 of 5% to sector **F** and 41/95 of 5% to sector **P.**

[15] This is done in two steps: (1) Sector F and P first estimated by proportional allocation of Sector X amounts (see footnote 13). (2) New unallocated sector **K** amounts, after estimating P, were allocated in the same way proportionally to sectors **C, G** and **H**.

[16] All countries do not yet report full breakdown of non-bank subsectors. Thus the sector-breakdown of aggregate positions by banks in 'all reporting countries' vis-à-vis individual counterparty countries are incomplete. We propose to proportionally allocate residual amounts to reportable subsectors (see footnotes 13 and 14). The average share for each reportable sub-sectors (i.e. after reallocating residual amounts) from the latest quarters are applied to sector **N**

at aggregate level in the presence of confidentiality restrictions. The consequence is that aggregate estimations are broad based, less precise and may not provide good results for distant historical quarters, neither to estimate bilateral positions. The gap in amounts between estimated figure from 'all reporting countries' and that from sum of leaf-level reported plus estimated amounts is narrowed down when the coverage for a counterparty country is high as in the latest quarters for Portugal (coverage more than 95%). One of the main consequences of using aggregate estimations is bilateral asymmetries that will occur as the statistical estimations are not based on a country-by-country data.

II – Bilateral level estimations

In the cases where there is reported data on a bilateral basis by LBS reporting country vis-à-vis the domestic country, mirror data exercises are more effective and precise. The term 'domestic country' used below refers to the country that we wish to estimate households' cross-border deposits. The following three scenarios are defined according to the available information for counterparty sectors **N**, **F**, **P** and **H**[17]:

1. **Deposits placed by households (H) – liabilities of banks vis-à-vis sectors H and N are reported:**

In this first scenario, households' deposits abroad will correspond to the bilateral deposit liabilities of the counterpart reporting country to the domestic country. Such data available only from Q4 2013. If both sectors **P** and **H** are not reported, bilateral positions can be estimated backwards by applying in each quarter the weighted average[18] of the reported households sector (**H**) in the sector **N** to the non-banks sector (**N**) positions.

2. **Deposits placed by households (H) – liabilities of banks vis-à-vis P amounts are reported prior to sector H:**

In this second case, subsector **H** is reported at a later stage than sector **P**, we propose to use information of **H** when available and estimate **H** positions for each of the non-reported quarter using, the average[19] weight of the households sector (**H**) in the total amount of sector **P** (from the quarters with reported data on sector **H**), which is more precise than using sector **N** mentioned above.

amounts for all previous quarters when none of the sub-sectors of **P** are available. An improved alternative is to use moving average or average from the latest 4 or 8 quarters and apply the share to reported sector **N** amounts.

[17] See footnote 13. We propose to apply the same method for individual reporting countries to get better estimates not only for bilateral positions but also for total of 'all reporting countries' when re-aggregated.

[18] Which can be a simple / weighted or moving average.

[19] Which can be a simple / weighted or moving average.

3. Deposits placed by households (H) – liabilities of banks vis-à-vis sector H are not reported for any quarters - we consider two different cases:

3.1. Liabilities of banks vis-à-vis sectors **F** and **P** are (but not subsectors of **P**):

We first examine if deposit liabilities of banks vis-à-vis sector **P** of the domestic country is reported. If reported, we apply estimation using available data of banks in other reporting countries to get an average value for sector **H** vis-à-vis the domestic country (i.e. we propose to use average share of **H/P** from other reporting countries) and also estimate back-quarters as mentioned before in the second method[20]. If sector **P** amounts are not reported vis-à-vis the country, no estimate should be considered for subsector **H**[21].

3.2. Liabilities of banks vis-à-vis sectors **F** and **P are not reported (i.e. only sector N reported)**

There is no mirror information regarding sector **F** and **P**. In this case, if deposit liabilities of banks to sector **N** is reported vis-à-vis the country, we propose to apply average estimation for all quarters using available data of banks in other reporting countries to get value for sector **H** (eg use average share of **H/N**) and this is similar to the first method but uses (aggregated) reported data of banks in other countries[22].

Finally, we do not propose any estimation procedure for the situations where liabilities of banks vis-à-vis sector **N** of the country is not reported (sector **N** value is either zero or missing).

4. Results

According to our findings for Q4 2018(Graph 1b) the non-bank financial institutions represent about 57 % (or $4,350 billion) of the total non-bank amount ($7,678 billion), and the non-financial sectors represent 43% (or $3,328 billion), similar to the reported amount (see Graph 1a or Table on page 3). The estimated breakdown suggests that 34% of total non-financial sectors (sector **P**) are households (sector **H**) whereas sector **C** and sector **G** correspond

[20] For example, US reports sectors **F** and **P** but doesn't report subsectors of **P** for deposit liabilities. In this case, we examine if sector **P** vis-à-vis a given domestic/counterparty country (e.g. PT) is reported. If reported, we apply estimation using reported share from other countries vis-à-vis domestic country to get value for sector **H**. If sector **P** amounts are not reported, we don't estimate amounts for subsector **H**.

[21] If aggregate sector P doesn't exists, subsector **H** can't exist.

[22] SG doesn't report sectors **F** and **P** vis-à-vis any country: If SG reports sector **N** amounts vis-à-vis PT, apply estimation using reported data of other countries to get value for sector **H**). Don´t estimate in any other case.

to 61% and 5% respectively. Graph1b shows that there are no unallocated amounts under the estimated method and that the share of household sector in total of non-banks represents the highest estimated increase of 12.2% approximately (from reported 2.7% to estimated 14.9%) whereas the Non-financial corporations share increase 10.9% (from 15.6% to 26.5%). On the contrary, the government sector decreases from 1.1% reported share to 0.9%.

We elect Portugal as a domestic country to estimate households' cross border positions. We find that as of Q4 2018, 97.7% of deposit liabilities to non-banks ($12.9 billion) are reported with sector breakdown in to sector **F** ($2.6 billion, 20.4%) and sector **P** ($10 billion, 77.3%). In the total reported amount for sector **N**, the share of the unallocated amounts (sector **X** plus sector **K**) correspond to 6.9%. In addition, 72.7% of deposit liabilities to sector **P** is available with subsector breakdown (Graph 1c)[23]. In addition, **H** sector comprise nearly $6.9 billion, 53.6% of reported amount for sector **P**, sector **C** represents 18.8% and sector **G** has almost a null share. The estimated amounts show that Non-financial corporations share (Graph 1d) increases by 1.6% (from 18.8% to 20.5%) and Households from a reported share of 53.6% to an estimated of 58.3%.

Our results show that this estimation method appears consistent not only for counterparties in all countries but also for individual counterparty countries such as Portugal. The estimated shares/amounts for Portugal are expected to be close to the actual amounts because coverage of reporting is as high as 93.1% for non-financial sub-sectors. However, the simple proportional estimates could be improved by estimating sub-sectors for individual reporting countries and then using simple/weighted average of the estimated shares from the reported data to estimate the bilateral positions of countries that do not report these subsectors. Further refinement of estimates backwards could be achieved using a 4 or 8 quarter moving average and also test the robustness of our estimation by comparing with reported data when available.

5. Discussion and Conclusion

We develop this methodological framework on uses of reported data, estimation of non-reported data and data gaps, aiming to provide users with more complete information. The detailed version of this work is in progress (at advanced stage) and will provide much more additional details that we could not include in this version for the conference due to limitation of length. It is needless to mention that the mirror data methodology we propose and have

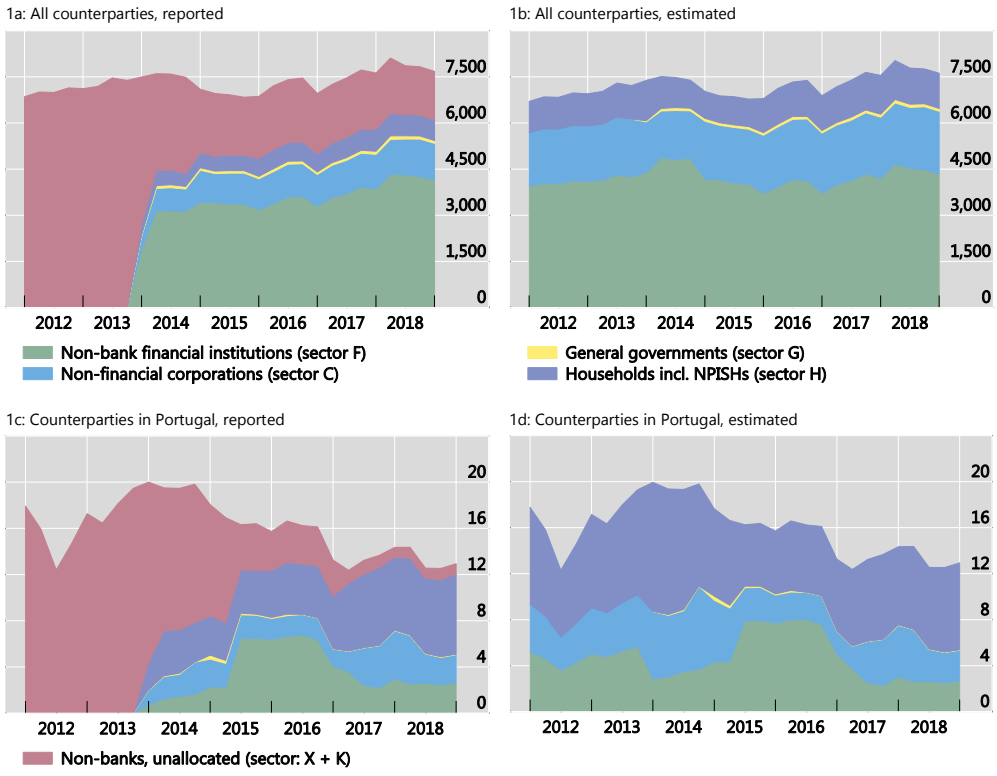
[23] This unallocated share comes to 5.3% of total liabilities to non-banks (aggregate). As of Q4 2018, while 43 countries report liabilities to sector **N** of Portugal, 28 countries report cross-border deposit liabilities to sector **H** and the share of unallocated non-bank is only 7% of non-bank.

explored so far permits the validation of common data items across statistical domains and provides additional information. The effectiveness of these exercises depend on the availability of the information. We discuss methodological aspects and offer guidance on derivation of household assets abroad for individual countries. The Portuguese country-case would be beneficial for other reporting countries (some countries have already started adopting our approach that we shared).

Mirror data analysis ensures consistency and enhances statistical quality standards, which is crucial for economists, analysts and policy makers who explore this information. When these estimates are considered for both BoP/IIP and RoW financial accounts purposes, there is also need to consider impacts on the other flows (exchange rate, price and volume changes). Furthermore, this approach can be seen as generic as it can be extended to the other sectors (in fact applied the estimation to sectors **F**, **P**, **C**, **G** and **H**).

Assets (claims) of non-bank subsectors with banks abroad[1]

Amounts outstanding, in USD billion — Graph 1



[1] Use of mirror data on liabilities of banks to non-bank subsectors; see methodology on estimation and allocation of non-available breakdown including unallocated sector amounts.

Sources: BIS locational banking statistics (by residence); authors' calculations.

References

1. *Committee on the Global Financial System (2012):* "*Improving the BIS international banking statistics*", CGFS Papers, no 47, BIS, November.
2. *Avdjiev, S., McGuire P. and Wooldridge P. (2015):* "*Enhanced data to analyse international banking*", BIS Quarterly Review, September, pp 53–68.
3. *Bank for International Settlements (2016)*: "*Recent enhancements to the BIS statistics*", BIS Quarterly Review, September, pp 35–44.
4. *Eurostat (2017)*: *Balance of Payments Vademecum*, European Commission, February.
5. *Pradhan S. and Silva, J. Falcão(2019)*: "*Uses of mirror data: examples from the BIS international banking statistics and other external statistics*", IFC Bulletin, January, No 49.
6. *International Monetary Fund (2018)*: "*Sectoral classification of international organizations*", Clarification note, IMF Committee on Balance of Payments Statistics.

Statistical formulation of scenarios under method II

The following scenarios are defined according to the available/reported information on counterparty sector **N**, **F**, **P** and **H**[24]**:**

1. **Deposits placed with banks by households** (H) **of a country** – liabilities of banks vis-à-vis sectors **H** and **N** are reported **[**available since Q4 2013]: We use reported data without any change[25]. If banks in country i report **H** sector vis-à-vis country j, country j should use directly the information reported from country i for its sector **H** for all reported periods:

   $H_t^{j,i}(\text{asset}) = H_t^{i,j}(liabilities)$; $t = 1,2,\dots,n$ stands for periods on a quarterly basis (1 is the latest period and n the first reporting quarter).

   For the non-reported periods (prior to Q4 2013), the estimation procedure follows two steps:

   a. Calculate the (simple/weighted/moving) average of the Households weight in the Non-banks sector (**N**) for the available periods ($\bar{H}_{n,N}$):

[24] In cases when subsectors are partially reported, we proportionally allocate residual amounts to reportable subsectors (eg N= F+P+X where F>0, P>0 and X>0, amount in X are proportionally allocated to F and P. The same rule is adopted for residual amounts in sector K)

[25] For example, Austria reports deposit liabilities vis-a-vis sector H of Portugal (PT) since Q4 2013.

$$\overline{H}_{n,N}^{i,j} = \frac{\sum_{t=1}^{n} \frac{H_t^{i,j}}{N_t^{i,j}}}{n}$$

b. Apply the average in a. to the Non-banks sector **N** for periods where no information is available ($\widehat{H}_{i,t+k}$):

$\widehat{H}_{t+k}^{j,i}(asset) = N_{t+k}^{i,j} \times \overline{H}_{n,N}^{i,j}(liabilities)$; $t = n$ and $k = 1, 2, \dots, m$.

2. **Deposits placed with banks by households** (H) **of a country** – liabilities of banks vis-à-vis **P** amounts are reported prior to sector **H** [ie sector H available from quarter later than Q4 2013]: We use reported data from available quarters and apply estimation method such as 1a. (simple/weighted/moving average from reported data of sector **H**) to earlier periods back to Q4 2013[26] . For quarters prior to Q4 2013, apply 1b. including estimated data up to Q4 2013.

However, if for certain periods country i reports sector **P** but no subsectors of **P** vis-à-vis country$j$, we consider the following approach:

a. Calculate the average over reported 'n' quarters of the households (**H**) weight to sector **P** for the reporting country 'i' vis-à-vis country j ($\overline{H}_{n,P}^{i,j}$)

$$\overline{H}_{n,P}^{i,j} = \frac{\sum_{c=1}^{n} \frac{H_c^{i,j}}{P_c^{i,j}}}{n}$$

b. Apply $\overline{H}_{n,P}^{i,j}$ to the reported amount of Non-financial sector **P** for c periods:

$\overline{\widehat{H}}_t^{j,i}(asset) = P_t \; X \;\; \overline{H}_{n,P}^{i,j}$ (liabilities); $t = 1,2, \dots, c$ stands for periods with sector **P** without its subsector.

c. For the non-reported periods, we adopt 1a and 1b. In other words, calculate the average for the estimated periods of the weight of the estimated Household amounts on the Non-bank sector, 'N', ($\overline{\widehat{H}}_N$):

$$\overline{\widehat{H}}_N = \frac{\sum_{t=1}^{n} \frac{\widehat{H}_t}{N_t}}{n}$$

d. Apply the average in c. to the Non-banks sector **N** for the periods where no information is available ($\overline{\widehat{H}}_{t+k}$):

[26] For example, Spain reports deposits liabilities vis-à-vis sector H of PT from Q1 2017.

$\bar{\bar{H}}_{t+k}(assets) = \bar{\bar{H}}_N \times N_{t+k}$ (liabilities); where $t = n$ and $k = 1, 2, \ldots, m$.

3. **Deposits placed with banks by households** (H) **of a country** – liabilities of banks vis-à-vis sector **H** are not reported for any quarters **-** we consider two different cases:

   a. **Case 1**: A country reports sectors **F** and **P** but does not at all report subsectors of **P** for deposits liabilities. We first examine if sector **P** vis-à-vis a given counterparty country is reported. If it is reported, we apply estimation (simple/weighted/moving average) using available data of other reporting countries to get an average value for sector **H** (e.g. use share **H/P**, from other countries including estimated data proposed in 2a above)[27]. If sector **P** amounts are not reported vis-à-vis a given counterparty country, don't estimate amounts for subsector **H**[28].

   b. **Case 2**: A country does not report sector **F** and **P** vis-à-vis any country. In this case, if it reports sector **N** amounts vis-à-vis the counterparty country, we propose to apply estimation (simple/weighted/moving average) for all quarters using available data of other reporting countries to get value for sector **H** (e.g. use share **H/N** from other countries vis-à-vis given counterparty country including estimated data proposed in 1a and 1b). If it is not the case, don't estimate[29].

Finally, we do not propose any estimation procedure for the situations where countries do not report aggregate the **N** sector vis-à-vis country i.

[27] For example, US reports sectors **F** and **P** but doesn't report subsectors of P (C, G and H) for deposit liabilities. In this case, we examine if sector **P** vis-à-vis a given domestic/counterparty country (e.g. PT) is reported. If US reported, we apply estimation using reported data of other countries to get value for sector **H**. If sector **P** amounts are not reported, we don't estimate amounts for subsector **H**.

[28] If aggregate sector **P** doesn't exists, subsector **H** can't exists.

[29] SG doesn't report sectors **F** and **P** vis-à-vis any country: If SG reports sector **N** amounts, say vis-à-vis PT, apply estimation using reported data of other countries vis-à-vis PT to get value for sector **H**). If SG doesn't report sector N amounts vis-à-vis PT, sector H is null or zero.

Robust estimators for some piecewise-deterministic markov Processes

Patrice Bertail[1] ,Gabriela Ciolek[2] ,Charles Tillier[2]
[1] Universit´e Paris Nanterre
[2]TelecomParisTech

Abstract
This talk is devoted to extending the notion of robustness to Markov chains with applications to PDMP, based on their (pseudo-) regenerative properties. Precisely, it is shown how it is possible to define the "influence function" in this framework, so as to measure the impact of (pseudo-) regeneration data blocks on the statistic of interest. We establish some asymptotic results for robust estimators of some usual functionals of the stationary measure. We also define the concept of regeneration-based signed linear rank statistic and L-statistics, as specific functionals of the regeneration blocks, which can be made robust against outliers in this sense. Indeed, even the usual quantile is not robust in this framework. We essentially apply these notions to reservoir models in insurance and hydrology. We obtain robust estimators of probability of ruins and quantiles (value at risk).

Keywords
Markov Chains; Piecewise Deterministic Markov processes; influence function

1. Regenerative Markov chains: Notations and context

Denote by $X = (X_n)_{n\in N}$ a positive recurrent Markov chain on a countably generated state space (E, ε) with transition probability Π and initial probability ν. For any $B \in E$ and $n \in N$, we have

$$X_0 \sim \nu \text{ and } P(X_{n+1} \in B|X_0, \cdots, X_n) = \Pi(X_n, B) \text{ a.s.}$$

In the following, P_x (resp. P_ν) designates the probability measure such that $X_0 = x$ and $X_0 \in E$ (resp. $X0 \sim \nu$), and $E_x(\cdot)$ is the P_x-expectation (resp. $E_\nu(\cdot)$ is the P_ν-expectation). All along this paper, we suppose that X is ψ -irreducible and aperiodic Markov chain. We are particularly interested in the atomic structure of Markov chains as in [1].

Definition 1. *Suppose that X is aperiodic and ψ-irreducible. We say that a set $A \in \varepsilon$ is an accessible atom if for all $x, y \in A$ we have $(x, \cdot) = \Pi(y, \cdot)$ and $\psi(A) > 0$. In that case we call X atomic.*

Roughly speaking, an atom is a set from which all the transition probabilities are the same. Suppose that X possesses an accessible atom. We define the sequence of regeneration times $(\tau_A(j))_{j\geq 1}$ i.e.

$$\tau_A = \tau_A(1) = inf\{n \geq 1 : X_n \in A\}$$

is the first time when the chain hits the regeneration set A and

$$\tau_A(j) = inf\{n > \tau_A(j-1), X_n \in A\} \text{ for } j \geq 2$$

is the j-th visit of the chain to the atom A.

By the strong Markov property, given any initial law ν, the sample paths can be divided into i.i.d. segments corresponding to the consecutive visits of the chain to regeneration set A. The blocks of data are of the form:

$$B_j = (X_{1+\tau_{A(j)}}, \cdots, X_{\tau_{A(j+1)}}), j \geq 1$$

and take values in the torus $\cup_{k=1}^{\infty} E^k$.

In the following, we are interested in steady-state analysis of Markov chains. More specifically, for a positive recurrent Markov chain if $E_A(\tau_A) < \infty$, then the unique invariant probability distribution μ is the Pitman's occupation measure

$$\mu(B) = \frac{1}{\mathbb{E}_A(\tau_A)} \left(\sum_{i=1}^{\tau_A} \mathbb{I}\{X_i \in B\} \right), \forall B \in \mathcal{E}.$$

We introduce few more pieces of notation: $l_n = \sum_{i=1}^{n} II\{X_i \in A\}$ designates the total number of consecutive visits of the chain to the atom A, thus we observe $l_n + 1$ data segments. We make the $B_{l_n}^{(n)} = \emptyset$ convention that when $\tau_A(l_n) = n$. We denote by $l(B_j) = \tau_A(j+1) - \tau_A(j), j \geq 1$ the length of regeneration blocks. From Kac's theorem it follows that

$$\mathbb{E}(l(B_j)) = \mathbb{E}_A(\tau_A) = \frac{1}{\mu(A)}.$$

General Harris Markov chains and the splitting technique

In this framework, we also consider more general classes of Markov chains, namely positive recurrent Harris Markov chains.

Definition 2. *Suppose that X is a ψ-irreducible Markov chain. We say that X is Harris recurrent iff, starting from any point $x \in E$ and any set such that $\psi(A) > 0$, we have*

$$\mathbb{P}_x(\tau_A < +\infty) = 1.$$

In short, Harris recurrence property ensures that X visits set A infinitely often a.s.. It is well-known that in Harris recurrent case it is also possible to recover the regeneration properties via *splitting technique* introduced in [5].

Definition 3. *We say that a set $S \in E$ is small if there exists a parameter $\delta > 0$, a positive probability measure Φ supported by S and an integer $m \in N^*$ such that*

$$\forall x \in S,\ A \in \mathcal{E}\ \ \Pi^m(x, A) \geq \delta\, \Phi(A),$$

where Π^m denotes the m-th iterate of the transition probability Π.

The inequality (1) from above definition is called minorization condition $M(m, S, \delta, \psi)$ and gives a uniform bound from below on the transition probabilities. Note that the parameter δ controls how fast our chain X forgets its past.

It is assumed throughout the rest of this paper that the minorization condition M is satisfied with $m = 1$. The family of the conditional distributions $\{\Pi(x, dy)\}_{x \in E}$ and the initial distribution ν are dominated by a σ-finite measure λ of reference, so that $\nu(dy) = f(y)\lambda(dy)$ and $\Pi(x, dy) = p(x, y)\lambda(dy)$, for all $x \in E$ and that $p(x, y) \geq \delta\varphi(y)$, $\lambda(dy)$ a.s. for any $x \in S$, with $\Phi(dy) = \varphi(y)dy$. We then split the series as in the same way as it was originally done in [1].

Algorithm (Approximate regeneration blocks construction)

1. Construct an estimator (may be on part of the data) $p_n(x, y)$ of the transition density using sample X_{n+1}. An estimator p_n must satisfy the following conditions

$$p_n(x, y) \geq \delta\gamma(y),\ \lambda(dy)\ a.s.\ and\ p_n(X_i, X_{i+1}) > 0,\ 1 \leq i \leq n.$$

2. Conditioned on X_{n+1} draw $\hat{Y}$'s only at those time points when $X_i \in S$. That is because only then the split chain can regenerate. At such time point i, draw $\hat{Y}i$ from the Bernoulli distribution with parameter $\delta\gamma(X_{i+1})/p_n(X_i, X_{i+1})$.
3. Count the number of visits $\hat{l}_n = \sum_{i=1}^{n} II\ "\{Xi \in S, \hat{Y}i = 1)$ to the atom $S_1 = S \times \{1\}$ up to time n.
4. Divide the trajectory X_{n+1} *into* $\hat{l}_n + 1$ approximate regeneration blocks according to the consecutive visits of $(X, \hat{Y})$ to S_1. Approximated blocks are of the form

$$\hat{\mathcal{B}}_0 = (X_1, \cdots, X_{\hat{\tau}_{S_1}(1)}), \cdots, \hat{\mathcal{B}}_j = (X_{\hat{\tau}_{S_1}(j)+1}, \cdots, X_{\hat{\tau}_{S_1}(j+1)}), \cdots,$$
$$\hat{\mathcal{B}}_{\hat{l}_n - 1} = (X_{\hat{\tau}_{S_1}(\hat{l}_n - 1)+1}, \cdots, X_{\hat{\tau}_{S_1}(\hat{l}_n)}), \hat{\mathcal{B}}_{\hat{l}_n}^{(n)} = (X_{\hat{\tau}_{S_1}(\hat{l}_n)+1}, \cdots, X_{n+1}),$$

where

$$\hat{\tau}_{S_1}(1) = \inf\{n \geq 1, X_n \in S, \hat{Y}_n = 1\}$$

and

$$\hat{\tau}_{S_1}(j+1) = \inf\{n > \hat{\tau}_{S_1}(j), X_n \in S, \hat{Y}_n = 1\} \quad \text{for } j \geq 1.$$

5. Drop the first block ˆB0 and the last one $\hat{\mathcal{B}}_{\hat{l}_n}^{(n)}$ if $\hat{\tau}_{S_1}(\hat{l}_n) < n$.

2. Robust functional parameter estimation for Markov Chains

The concepts of *influence function* and/or *robustness* in the i.i.d. setting provide tools to detect outliers among the data or influential observations. Extending the notion of *influence function* and/or *robustness* to the general time series framework is a difficult task; see [3] or [4]. Alternatively, the regenerative approach offers the opportunity of extending much more naturally an extension of the influence function based on the (approximate) regeneration blocks construction.

The influence function on the torus

Just like the stationary probability distribution $\mu(dx)$, most parameters of interest related to Harris positive chains are functionals of the distribution L_A of the regenerative blocks on the torus $T = U_{n \geq} E^n$, namely the distribution of $(X_1,..., X_{\tau A})$ conditioned on $X_0 \, E \, A$ when the chain possesses an accessible atom A, or the distribution of $(X_1,..., X_{\tau AM})$ conditioned on $(X_0, Y_0) \in A_M$ in the general case when one considers the split chain. For simplicity, we shall omit the subscript M and make no notational distinction between the regenerative and pseudo-regenerative cases. Indeed, the probability distribution P_ν of the Markov chain X starting from ν can be factorized as follows:

$$\mathbb{P}_\nu((X_n)_{n \geqslant 1}) = \mathcal{L}_\nu((X_1, ..., X_{\tau_{A(1)}})) \prod_{k=1}^{\infty} \mathcal{L}_A((X_{1+\tau_A(k)}, ..., X_{\tau_A(k+1)})),$$

where L_ν means the conditional distribution of $(X_1, \ldots, X_{\tau A})$ given that $X_0 \sim \nu$. Any functional of the law of the discrete-time process $(X_n)_{n \geq 1}$ can be thus expressed as a functional of the pair (L_ν, L_A). In the time-series asymptotic framework, since the distribution of L_ν cannot be estimated in general, only functionals of L_A are of practical interest. We propose a notion of influence function for such statistics. Let P_T denote the set of all probability measures on the torus T and for any $b \in T$, set $L(b) = k$ if $b \in E^k$, $k > 1$. We then have the following natural definition, which straightforwardly extends the classical notion of influence function in the i.i.d. case, with the important novelty that distributions on the torus are considered here.

Definition 4. (INFLUENCE FUNCTION ON THE TORUS) *Let* $(V, 11.11)$ *be a separable Banach space. Let* $T : P_T \to V$ *be a functional on PT. If, for all L in PT,* $t^{-1}(T((1-t)L + \delta_b) - T(L))$ *has a finite limit as* $t \to 0$ *for any* $b \in T$, *the influence function* $T^{(1)} : PT \to V$ *of the functional T is then said to be well-defined, and, by definition, one set for all b in T,*

$$T^{(1)}(b,\ \mathcal{L}) = \lim_{t\to 0} \frac{T((1-t)\mathcal{L} + t\delta_b) - T(\mathcal{L})}{t}. \qquad (2)$$

Definition 5. (GROSS-ERROR SENSIVITY) *A functional T is said to be Markov-robust iff its influence function $T^{(1)}(b, L)$ is bounded on the torus T. The gross-error sensitivity to block contamination is then defined as*

$$\gamma^*(T, \mathcal{L}) = \sup_{b\in\mathbb{T}} \|T^{(1)}(b, \mathcal{L})\|.$$

These quantities may be estimated either with the true blocks (in the atomic cases) or with the approximated ones in the general Harris recurrent case.

It is now easy to see how it is possible to derive functional central limit theorems for Fréchet differentiable functionals in a Markovian setting. for plug in estimators base either on (in the regenerative case)

$$\mu_n = \frac{1}{n_A} \sum_{i=1}^{l_n - 1} f(B_i)$$

or (in the Harris general case)

$$\hat{\mu}_n = \frac{1}{\hat{n}_{A_{\mathcal{M}}}} \sum_{i=1}^{\hat{l}_n - 1} f(\hat{B}_i),$$

where $\hat{B}_i$, $i = 1, \cdots, \hat{l}_n - 1$ are pseudo-regeneration blocks and $\hat{n}_{AM} = \hat{\tau}_{AM}(\hat{l}_n) - \hat{\tau}_{AM}(1) = \sum_{i=1}^{\hat{l}_n - 1} l(\hat{B}_j)$ is the total number of observations after the first and before the last pseudo-regeneration times.

3. Examples

Example 1: Sample means. Suppose that X is positive recurrent with stationary distribution μ. Let $f: E \to R$ be μ-integrable and consider the parameter μ(f) def = E_μ[f (X)], f is a real function. Denote by B a r.v. valued in T with distribution L_A and observe that

$$\mu(f) = \mathbb{E}\left[\mathbf{f}(\mathcal{B})\right] / \mathbb{E}_{\mathcal{L}_A}\left[L(\mathcal{B})\right] = T(\mathcal{L}_A),$$

with the notation f(b):= $\sum_{i=1}^{L(b)} f(b_i)$ for any b = (b_1, . .., $b_{L(b)}$) ∈ T. A classical calculation for the influence function of ratios yields

$$T^{(1)}(b, \mathcal{L}_A) = \frac{d}{dt}(T((1-t)\mathcal{L}_A + tb)|_{t=0} = \frac{\mathbf{f}(b) - \mu(f)L(b)}{\mathbb{E}_{\mathcal{L}_A}\left[L(\mathcal{B})\right]}.$$

Notice that E_{LA} [$T^{(1)}$(B, L_A)] = 0.

In the i.i.d. setting it is known that, if f is bounded by some constant $M_f < +\infty$, the corresponding functional is robust and may be simply estimated by its empirical counterpart. In the Markovian situation, even in the bounded case, $T^{(1)}(b, L_A)$ is generally not bounded and $\gamma^*(T, L_A) = \infty$. This point has also been stressed in [4], with a different definition of the influence function however. A robustified version of this parameter then can be defined as

$$\widetilde{T}_M(\mathcal{L}_A) = \frac{\mathbb{E}_{\mathcal{L}_A}\left[\mathbf{f}(\mathcal{B})\mathbb{I}\{L(\mathcal{B}) \leq M\}\right]}{\mathbb{E}_{\mathcal{L}_A}\left[L(\mathcal{B})\mathbb{I}\{L(\mathcal{B}) \leq M\}\right]},$$

and the plug-in estimator becomes

$$\frac{\sum_i^{l_n-1} \mathbf{f}(\mathcal{B}_i)\mathbb{I}\{L(\mathcal{B}_i) \leq M\}}{\sum_{i=1}^{l_n-1} L(\mathcal{B}_i)\mathbb{I}\{L(\mathcal{B}_i) \leq M\}}.$$

This simply consists in getting rid of the blocks (or the pseudo-blocks) whose lengths are too large compared to M. This applies in particular to the estimation of the stationary measure μ when f is an indicator function $\sum_{i=1}^{TA} 1\{Xi \leq x\}$ leading to an estimator $\tilde{F}_{LA,M,n}$ of the cdf of the stationary measures.

Example 2: M-estimators. Suppose that $E \subset R$ for simplicity. Let ϑ be the unique solution of the equation:

$$\mathbb{E}_\mu\left[g(X, \theta)\right] = 0,$$

where $g : R^2 \to R$ is of class C^2. Equipped with the notation $g(b, \theta) := \sum_{i=1}^{L(b)} g(bi, \theta)$ for all $b \in T$, the score equation is equivalent to $E_{LA}[g(B, \theta)] = 0$. A calculation entirely similar to that carried out in the i.i.d. setting (provided that differentiating inside the expectation is authorized) gives

$$T_\psi^{(1)}(b, \mathcal{L}_A) = -\frac{\mathbf{g}(b,\ \theta)}{\mathbb{E}_{\mathcal{L}_A}\left[\frac{\partial \mathbf{g}(\mathcal{B},\theta)}{\partial\theta}\right]},$$

where $\partial g(b, \theta)/\partial\theta = \sum_{i=1}^{L(b)} \partial g(b, \theta)/\partial\theta$. By definition of θ, we naturally have $\mathbb{E}_{\mathcal{L}_A}[T_\psi^{(1)}(\mathcal{B}, \mathcal{L}_A)] = 0$.

Example 2: Quantiles. We place ourselves in the case $E \subset R$. Assume that the stationary distribution has a continuous cdf $F_\mu(x) = \mu(]-\infty, x])$ and density $f_\mu(x)$. Consider the α−quantile $\tilde{T}_\alpha(\mu) = F_\mu^{-1}(\alpha)$. This parameter can also be viewed as a functional of L_A, $T_\alpha(L_A)$ say, it is the unique solution of the equation

$$\mathbb{E}_{\mathcal{L}_A}\left[\sum_{i=1}^{L(b)}\{\mathbb{I}\{b_i \leq \theta\} - \alpha\}\right] = 0.$$

A straighforward computation following in the footsteps of those carried out in the i.i.d. case (see [5] for further details) shows that, if $f_\mu(T_\alpha(\mu)) \neq 0$, the influence function is given here by

$$T_\alpha^{(1)}(b, \mathcal{L}_A) = \frac{\sum_{i=1}^{L(b)}(\alpha - \mathbb{I}\{b_i \leq T_\alpha(\mu)\})}{\mathbb{E}_{\mathcal{L}_A}[L(\mathcal{B})] f_\mu(\widetilde{T}_\alpha(\mu))}.$$

It follows that the gross-error sensivity of a quantile in a dependent framework is $\gamma^*(T_\alpha(\mu), L_A) = \infty$: an empirical quantile is generally not robust in the Markovian framework. As in example 1, one has to get rid of large blocks to robustify the estimator of the quantiles.

Example 3: The KDEM model.

The KDEM for Kinetic Dietary Exposure Model is a stochastic process that aims at representing the evolution of a contaminant in the human body through time or the amount of water in a tank (with some elimination *after* each rains). It has been proposed few years ago in [2]. In this context of dietary risk assessment, for $i \geq 0$,

- W_i's are random variables called *intakes.* They correspond to the intake of a contaminated food and occur at times T_i's, called *intake instants.*
- ΔT_i's, called inter-arrivals, are the durations between the (i − 1)-th and the i-th intake and are defined for $i \geq 0$ by $\Delta T_i = Ti - T_i - 1$.
- N(t) is a counting process that counts the number of intakes that occured until time $t \geq 0$.

In the sequel, we denote $X(t)$, the total body burden of a chemical at the instant $t \geq 0$. Following [2], between two intakes, we consider that the exposure process $X = (X(t))_{t\geq 0}$ moves in a deterministic way according to the first order differential equation

$$dX(t) = [\omega \times X(t)]dt, \tag{3}$$

with $\omega > 0$ a fixed parameter, called elimination rate, that describes the metabolism in regards to the chemical elimination. For $t \geq 0$, let $A(t) = t - T_{N(t)}$, be the *backward recurrence time*, which is the duration between the present time t and the lastest intake instant $T_{N(t)}$. Then, the bivariate process $\{(X(t), A(t))\}_{t\geq 0}$ is a PDMP. By solving (3), one may straightforwardly see that the exposure process can be written for any $t \geq 0$ as

$$X(t) = X(T_{N(t)}) \times e^{-\omega A(t)}. \quad (4)$$

The embedded chain of X, denoted $\tilde{X} = (X(T_n))_{n\in N} := (X_n)_{n\in N}$, which is the process on the intake instants $T_0, T_1, \ldots$ plays a leading role in the analysis of X and describes the exposure process immediately after each intake. It is defined by the following stochastic recurrence equation

$$X_{n+1} = X_n \times e^{-\omega \Delta T n+1} + W_{n+1}, \quad n \geq 0. \quad (5)$$

Equation (5) is an autoregressive process with random coefficient. Under additional assumptions [2] have related the continuous-time process X with the embedded chain $\tilde{X}$. Denote, $\mu(dx) = g(x)dx$ *(resp.* $\tilde{\mu}(dx) = \tilde{g}(x)dx$*).* They show that the limiting distribution μ and $\tilde{\mu}$ are linked by the following equation

$$\mu([u,\infty[) = \lambda^{-1} \int_u^\infty \int_0^\infty t \wedge \omega^{-1} log(x/u) \tilde{\mu}(dx) H(dt), \quad u > 0, \quad (6)$$

in such a way that $\frac{1}{T}\int_0^T II\{x(t), u\}dt \rightarrow \mu(]0,\mu])$ when $t \rightarrow \infty$. Let $\tilde{F}_{LA,M,n}$ be a robust estimator of $\mu(]-\infty, x])$ as in example 1, then the robust estimator is given by

$$\widehat{\mu}_n([u,\infty[) = \lambda^{-1} \int_u^\infty \int_0^\infty t \wedge \omega^{-1} log(x/u) \tilde{F}_{\mathcal{L}_A,M,n}(dx) H(dt), \quad u > 0,$$

$$= \frac{\sum_i^{l_n-1} \sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} \mathbb{I}\{\tau_A(i+1) - \tau_A(i) \leq M\} 1_{\{X_j \geq u\}} \int_0^\infty t \wedge \omega^{-1} log(X_j/u) H(dt)}{\lambda \sum_{i=1}^{l_n-1} (\tau_A(i+1) - \tau_A(i)) \mathbb{I}\{\tau_A(i+1) - \tau_A(i) \leq M\}}$$

Similarly to the Sparre-Andersen case, in the exponential inter-arrival case, we have the expression

$$\int_0^\infty t \wedge \omega^{-1} log(X_j/u) H(dt) = \lambda(1 - (X_j/u)^{-1/(\omega\lambda)})$$

Notice that in that case, the (non-robust) plug-in estimator of $\mu([, \infty[)$ has the nice expression

$$\mu_n([u,\infty[) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_j \geq u\}} (1 - (X_j/u)^{-1/(\omega\lambda)}).$$

The robust estimator is simply the version of this mean only over the the X_i's which do not belong to large blocks:

$$\widehat{\mu}_n([u,\infty[) = \frac{\sum_{i=1}^{l_n-1}\sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} \mathbb{I}\{\tau_A(i+1)-\tau_A(i)\leq M\}1_{\{X_j\geq u\}}(1-(X_j/u)^{-1/(\omega\lambda)})}{\sum_{i=1}^{l_n-1}(\tau_A(i+1)-\tau_A(i))\mathbb{I}\{\tau_A(i+1)-\tau_A(i)\leq M\}}.$$

References

1. Bertail, P., Clemencon S. and J. Tressou. A storage model with random release rate for modelling exposure to food contaminants. Mathematical Bioscience and Engineering, vol 60, p. 5-35, 2008.
2. Bertail, P., Clemencon S. Regeneration-based statistics for Harris recurrent Markov chains, Bernoulli vol. 187, p. 3?54, 2006.
3. H. Kunsch. Infinitesimal robustness for autoregressive processes. Ann. Statist., 12, 843-863, 1984.
4. R.D. Martin and V.J. Yohai. Influence functionals for time series. Ann. Stat.,14,781-818, 1986.
5. E. Nummelin. A splitting technique for Harris recurrent chains. Z. Wahrsch. Verw. Gebiete, 43, 309-318, 1978.
6. H. Rieder. Robust asymptotic statistics. Springer verlag, N.Y. 1994.

Generalized partially linear spatial probit models and applications

Mohamed Salem Ahmed[1], Sophie Dabo-Niang[2,3], Michael Genin[1]
[1]University of Lille, laboratory CERIM, EA 2694, France
[2]Laboratory LEM-CNRS 9221, University of Lille, Villeneuve d'Ascq, France
[3]INRIA Lille Nord-Europe, MODAL-Team

Abstract
Generalized partially linear probit regression model for spatially dependent data is considered. Conditional heteroscedasticity and non-identically distributed observations and a linear process for disturbances are assumed allowing various spatial dependencies. The estimation procedure proposed combines a weighted likelihood and a generalized method of moments. Consistency of the parametric and non-parametric components estimators and asymptotic normality results are established under sufficient conditions. Numerical experiments including real economic and environmental data applications to investigate the finite sample performance of the estimators are given.

Keywords
spatial data; probit models; generalized partially linear models; non-parametric

1. Introduction

Agriculture, economics, environmental sciences, urban systems, epidemiology activities are of-ten located in space. Therefore, modeling such activities requires to find a kind of correlation between some random variables in one location with others at neighboring locations, see for instance Pinkse and Slade (1998). This is a significant feature of spatial data analysis. Spatial econometrics/statistics provides tools to solve such modeling. A lot of studies on spatial effects in statistics and econometrics in many divers models have been published; see Anselin (1988), Cressie (1993) and Arbia (2006) for a review.

Two main ways of incorporating the spatial dependence structure can be distinguished basically for geostatistics and lattice data. In the domain of geostatistics, the spatial location is valued in a continuous set of R^N, $N \geq 2$. However, for many activities, the spatial index or location does not vary continuously and may be of the lattice type, the baseline of this current work. This is, for instance, the case in a number of problems. In images analysis, remote sensing form satellites, agriculture and so one, data are often received as regular lattice and identified as the centroids of square pixels, whereas a

mapping forms often an irregular lattice. Basically, statistical models for lattice data are linked to nearest neighbors to express the fact that data are nearby. Two popular spatial dependence models have received a lot of attention in lattice data: the spatial autoregressive (SAR) dependent variable model and the spatial autoregressive error model (SAE, where the model error is a SAR), which extend regression in time series to spatial data.

In a in theoretical point of view, various linear spatial regression SAR and SAE models, their identification and estimation methods by the two stage least squares (2SLS), the three stage least squares (3SLS), the maximum likelihood (ML) or quasi-maximum likelihood (QML) and the generalized method of moments (GMM) methods have been developed and summarized by many authors, such as Anselin (1988), Kelejian and Prucha (1999), Conley (1999), Lee (2004), Garthoff and Otto (2017). Nonlinearity into the field of spatial linear lattice models have less attention, see for instance Robinson (2011) who generalized the kernel regression estimation to spatial lattice data. Su (2012) proposed a semiparametric GMM estimation for some semiparametric SAR models. Extending these models and methods to discrete choice spatial models have less attention, only a few number of papers were concerned in recent years. This may be, as pointed out by Fleming (2004), due to the "added complexity that spatial dependence introduces into discrete choice models". Estimating the model parameters with a full ML approach in spatial discrete choice models, often requires solving a very computationally demanding problem of n-dimensional integration, where n is the sample size.

As for linear models many discrete choice models are fully linear and make use of a continuous latent variable, see for instance Smirnov (2010) and Wang et al. (2013) that proposed pseudo ML methods and Pinkse and Slade (1998) who studied a method based on GMM approach.

When the relationship between the discrete choice variable and some explanatory variables is not linear, then a semiparametric model may be an alternative to fully parametric models. This kind of models is known in literature as partially linear choice spatial models and is the baseline of this current work. When the data are independent, these choice models can be viewed as particular cases of the famous generalized additive models (Hastie and Tibshirani, 1990) and have received a lot of attention in the literature, various methods of estimation have been explored (see for instance Severini and Staniswalis, 1994; Carroll et al., 1997).

To the best of our knowledge, semiparametric spatial choice models, have not yet been investigated in a theoretical point of view. To fill in this gap, this work addresses a SAE spatial probit model when the spatial dependence structure is integrated in a disturbance term of the studied model. We propose a semiparametric estimation method combining the GMM approach and the weighted likelihood method. It consists to first fix the parametric components

of the model and estimate nonparametrically the nonlinear component by weighted likelihood. The obtained estimator depending on the values at which the parametric components were fixed is used to construct a GMM estimator (Pinkse and Slade, 1998) of these components.

2. Model

We consider that at n spatial locations $\{s_1, s_2, \dots, s_n\}$ satisfying $\|s_i - s_j\| > p$ with $p > 0$, observations of a random vector (Y, X, Z) are available. Assume that these observations are considered as triangular arrays (Robinson, 2011) and follow the partially linear model of a latent dependent variable Y^*:

$$Y_{in}^* = X_{in}^T \beta_0 + g_0(Z_{in}) + U_{in}, \qquad 1 \le i \le n,\ n = 1, 2, \dots \tag{1}$$

with

$$Y_{in} = \mathbb{I}\left(Y_{in}^* \ge 0\right), \qquad 1 \le i \le n,\ n = 1, 2, \dots \tag{2}$$

where X and Z are explanatory random variables taking values in two compacts subsets $\mathcal{X} \subset \mathbb{R}^p (p \ge 1)$ and $\mathcal{Z} \subset \mathbb{R}^d (d \ge 1)$ respectively. The parameter β_0 is an unknown $p \times 1$ vector that belongs to a compact subset $\Theta_\beta \subset \mathbb{R}^p$, $g_0(\cdot)$ is an unknown smooth function valued in the space of functions $\mathcal{G} = \{g \in C^2(\mathcal{Z}): \|g\| = sup_{z \in Z} |g(z)| < C\}$ with $C^2(\mathcal{Z})$ the space of twice differentiable functions from $\mathcal{Z}$ to $\mathbb{R}$, C a positive constant. In model (1), β_0 and $g_0(.)$ are constant over i (and n). Assume that the term of disturbance U_{in} in (1) is modeled by the following spatial autoregressive process (SAR):

$$U_{in} = \lambda_0 \sum_{j=1}^{n} w_{ijn} U_{jn} + \varepsilon_{in}, \qquad 1 \le i \le n,\ n = 1, 2, \dots \tag{3}$$

where λ_0 is the autoregressive parameter, valued in the compact subset $\Theta_\lambda \subset \mathbb{R}$, $w_{ijn}, j = 1, \dots, n$ are the elements in the i–th row of a non-stochastic $n \times n$ spatial weights matrix W_n, that contains the information on the spatial relationship between observations. This spatial weight matrix is usually constructed as a function of the distances (with respect to some metric) between locations, see Pinkse and Slade (1998) for more of details. The $n \times n$ matrix $(I_n - \lambda_0 W_n)$ is assumed to be nonsingular for all n where I_n denotes the $n \times n$ identity matrix, and $\{\varepsilon_{in}, 1 \le i \le n\}$ are assumed to be independent random Gaussian variables; $E(\varepsilon_{in}) = 0$ and $E\left(\varepsilon_{in}^2\right) = 1$ for $i = 1, \dots, n$ $n = 1,2, \dots$. Note that one can rewrite (3) as:

$$U_n = (I_n - \lambda_0 W_n)^{-1} \varepsilon_n, \quad n = 1, 2, \dots$$

where $U_n = (U_{n1}, \dots, U_{nn})^T$ and $\varepsilon_n = (\varepsilon_{n1}, \dots, \varepsilon_{nn})^T$ .Therefore, the variance-covariance matrix of U_n is

$$V_n(\lambda_0) = \mathrm{Var}(U_n) = (I_n - \lambda_0 W_n)^{-1} \left\{ (I_n - \lambda_0 W_n)^T \right\}^{-1}, \quad n = 1, 2, \dots$$

This matrix allows to describe the cross-sectional spatial dependencies between the n observations. Furthermore, the fact that the diagonal elements of $V_n(\lambda_0)$ depend on λ_0 and particularly on i and n allows some spatial heteroscedasticity. These spatial dependence and heretoscedasticity depend on the neighborhood structure established by the spatial weights matrix W_n. Before going further, let us give some particular cases of the model.

If one consider i.i.d observations, that is $V_n(\lambda_0) = \sigma^2 I_n$, with σ depending on λ_0, the obtained model may be seen as a particularly case of the classical generalized partially linear models (e.g Severini and Staniswalis, 1994) or the classical generalized additive model (Hastie and Tibshi-rani, 1990). Several approaches of estimating this particular model have been developed, among others, we cite that of Severini and Staniswalis (1994), based on the concept of generalized profile likelihood (e.g Severini and Wong, 1992). This approach consists to first fix the parametric parameter β and estimate nonparametrically $g_0(\cdot) = 0$ by using the weighted likelihood method. This last estimate is then used to construct a profile likelihood to estimate β_0.

When $g_0(\cdot) = 0$ (or is an affine function), that is without a nonparametric component, several approaches have been developed to estimate the parameters β_0 and λ_0. The basic difficulty encountered is that the likelihood function of this model involve a n dimensional normal integral, thus when n is high, the computation or asymptotic properties of the estimates may be difficult (e.g Poirier and Ruud, 1988). Various approaches have been proposed to address this difficulty, among these we cite:

- Feasible Maximum Likelihood approach: it consists of replacing the true likelihood function by a pseudo likelihood function constructed via marginal likelihood functions. Smirnov (2010) proposes a pseudo likelihood function obtained by replacing $V_n(\lambda_0)$ by some diagonal matrix by the diagonal elements of $V_n(\lambda_0)$. Alternatively, Wang et al. (2013) proposed to divide the observations by pairwise groups where the latter are assumed to be independent with bivariate normal distribution in each group and estimate β_0 and λ_0 by maximizing the likelihood of these groups.
- GMM approach used by Pinkse and Slade (1998). These authors used the generalized residuals defined by $\widetilde{U}_{in}(\beta, \lambda) = E(U_{in} | Y_{in}, \beta, \lambda), i = 1, \dots, n,\ n = 1,2, \dots$, with some instrumentals variables to construct moments equations to define GMM estimators of β_0 and λ_0.

In what follows, using the n observations (X_{in}, Y_{in}, Z_{in}) we propose parametric estimators of β_0, λ_0 and a nonparametric estimator of the smooth function $g_0(\cdot)$.

To this aim, assume that, for all $n = 1,2, \dots, \{\varepsilon_{in}, i = 1, \dots, n\}$ is independent of

$\{X_{in},\ i=1,\dots,n\}$ and $\{Z_{in},\ i=1,\dots,n\}$ and $\{X_{in},\ i=1,\dots,n\}$ is independent of $\{Z_{in},\ i=1,\dots,n\}$ We give asymptotic results according to *Increasing domain* asymptotic.

2.1 Estimation procedure

We propose an estimation procedure based on a combination of a weighted likelihood method and a generalized method of moments. We first fix the parametric components β and λ of the model and estimate the nonparametric component using a weighted likelihood. The obtained estimate is then used to construct generalized residuals where the latter are combined to instrumentals variables to propose GMM parametric estimates. This approach will be described as follow. By equation (2) we have

$$E_0\left(Y_{in}|X_{in},Z_{in}\right)=\Phi\left(\left(v_{in}(\lambda_0)\right)^{-1}\left(X_{in}^T\beta_0+g_0(Z_{in})\right)\right),\ i=1,\dots,n,\ n=1,2,\dots \qquad (4)$$

where E_0 denotes the expectation under the true parameters (i.e β_0, λ_0 and $g_0(\cdot)$), $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, and $(v_{in}(\lambda_0))^2=V_{iin}(\lambda_0), 1\le i\le n, n=1,2,\dots$ are the diagonal elements of $V_n(\lambda_0)$.

For each $\beta\in\Theta_\beta,\lambda\in\Theta_\lambda,z\in Z$ and $\eta\in\mathbb{R}$, we define the conditional expectation on Z_{in} of the log-Likelihood of Y_{in} given (X_{in},Y_{in}) for $1\le i\le n, N=1,2,\dots,$

$$H(\eta;\beta,\lambda,z)=E_0\left(\mathcal{L}\left(\Phi\left(\left(v_{in}(\lambda)\right)^{-1}\left(\eta+X_{in}^T\beta\right)\right);Y_{in}\right)\middle|\,Z_{in}=z\right),$$

with $\mathcal{L}(u;v)=\log(u^v(1-u)^{1-v})$. Note that $H(\eta;\beta,\lambda,z)$ is assumed to be constant over i (and n). For each fixed $\beta\in\Theta_\beta,\lambda\in\Theta_\lambda$ and $z\in Z, g_{\beta,\lambda(z)}$ denotes the solution in η of

$$\frac{\partial}{\partial\eta}H(\eta;\beta,\lambda,z)=0. \qquad (5)$$

Then, we have $g_{\beta_0,\lambda_0}(z)=g_0(z)$ for all z ∈ Z.

Now using $g_{\beta,\lambda}(\cdot)$, we construct GMM estimates of β_0 and λ_0 as Pinkse and Slade (1998). For that, we define the generalized residuals, replacing $g_0(Z_{in})$ in (1) by $g_{\beta,\lambda}(Z_{in})$;

$$\begin{aligned}\tilde{U}_{in}(\beta,\lambda,g_{\beta,\lambda}) &= E\left(U_{in}|Y_{in},\beta,\lambda\right)\\ &= \frac{\phi\left(G_{in}(\beta,\lambda,g_{\beta,\lambda})\right)\left(Y_{in}-\Phi\left(G_{in}(\beta,\lambda,g_{\beta,\lambda})\right)\right)}{\Phi\left(G_{in}(\beta,\lambda,g_{\beta,\lambda})\right)\left(1-\Phi\left(G_{in}(\beta,\lambda,g_{\beta,\lambda})\right)\right)},\end{aligned} \qquad (6)$$

where $\emptyset(\cdot)$ is the density of the standard normal distribution and $G_{in}(\beta,\lambda,g_{\beta,\lambda})=(v_{ni}(\lambda))^{-1}\left(X_{in}^T\beta+g_{\beta,\lambda}(Z_{in})\right)$.

For simplicity of notation, we write when it is possible $\theta = (\beta^T, \lambda)^T \in \Theta = \Theta_\beta \times \Theta_\lambda$.

Note that in (6), the generalized residual $\widetilde{U}_{in}(\cdot,\cdot)$ is calculated by conditioning only on Y_{in} not on the entire sample $\{Y_{in},\ i = 1,2,\dots,n,\ n = 1,\dots\}$ or a subset of it. This of course will influence the efficiency of the estimators of θ obtained by these generalized residuals, but it allows to avoid a complex computation. To address this loss of efficiency, let us follow Pinkse and Slade (1998)'s procedure that consists of employing some instrumentals variables in order to create some moments conditions, and use some random matrix to define a criterion function. Both the instrumentals variables and the random matrix permit to take into account more informations about the spatial dependence and heteroscedasticity in the dataset. Let us now detail the estimation procedure.

Let

$$S_n(\theta, g_\theta) = n^{-1}\xi_n^T \tilde{U}_n(\theta, g_\theta), \tag{7}$$

where $\widetilde{U}_{in}(\theta, g_\theta)$ is the $n \times 1$ vector, composed of $\widetilde{U}_{in}(\theta, g_\theta), 1 \le i \le n$ and ξ_n is a $n \times q$ matrix of instrumentals variables whose ith row is denoted by the $n \times 1$ random vector ξ_{in}. The latter may depend on $g_0(\cdot)$ and θ. We assume that ξ_{in} is $\sigma(X_{in}, Z_{in})$ measurable for each $i = 1,\dots,n, n = 1,2,\dots$. We suppress the possible dependence of the instrumentals variables on the parameters for notational simplicity. The GMM approach consists to minimize the following sample criterion function,

$$Q_n(\theta, g_\theta) = S_n^T(\theta, g_\theta) M_n S_n(\theta, g_\theta), \tag{8}$$

where M_n is some positive-definite $q \times q$ weight matrix that may depend on sample information. The choice of the instrumentals variables and weight matrix characterizes the difference between GMM estimator and all pseudo maximum likelihood estimators. For instance, if one takes

$$\xi_{in}(\theta, g_\theta) = \frac{\partial G_{in}(\theta, \eta_i)}{\partial \theta} + \frac{\partial G_{in}(\theta, \eta_i)}{\partial \eta} \frac{\partial g_\theta}{\partial \theta}(Z_{in}), \tag{9}$$

with $\eta_i = g_0(Z_{in}), G_{in}(\theta, \eta_i) = (v_{in}(\lambda))^{-1}\left(X_{in}^T\beta + \eta_i\right)$, $M_n = I_q$ with $q = p + 1$, then the GMM estimator of θ is equal to a pseudo maximum profile likelihood estimator of θ, accounting only the spatial heteroscedasticity.

Now, let

$$S(\theta, g_\theta) = \lim_{n\to\infty} E_0\left(S_n(\theta, g_\theta)\right), \tag{10}$$

and

$$Q(\theta, g_\theta) = S^T(\theta, g_\theta) M S(\theta, g_\theta),$$

where M, the limit of the sequence M_n, is a nonrandom positive definite matrix. The functions $S_n(\cdot,\cdot)$ and $Q_n(\cdot,\cdot)$ are viewed as empirical counterparts of $S(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ respectively. It is clear that $g_0(\cdot)$ is not available in practice. However, we need to estimate it, particularly by an asymptotically efficient estimate. By (5) and for fixed $\theta^T = (\beta^T, \lambda) \in \Theta$ an estimator of $g_0(z)$ for z ∈ Z can be given by $\hat{g}_\theta(z)$ which denotes the solution in η of

$$\sum_{i=1}^{n} \frac{\partial}{\partial \eta} \mathcal{L}\left(\Phi\left(G_{in}(\theta, \eta)\right); Y_{in}\right) K\left(\frac{z - Z_{in}}{b_n}\right) = 0, \tag{11}$$

where $K(\cdot)$ is a kernel from $\mathbb{R}^d$ to $\mathbb{R}_+$ and b_n is a bandwidth depending on n.

Now, replacing $g_0(\cdot)$ in (8) by the estimator $\hat{g}_\theta(\cdot)$ permits to obtain the GMM estimator $\hat{\theta}$ of θ as

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} Q_n(\theta, \hat{g}_\theta). \tag{12}$$

A classical inconvenience of the estimator $\hat{g}_\theta(z)$ proposed in (11) is that the bias of $\hat{g}_\theta(z)$ is high for z near the boundary of Z. Of course, this bias will effect the estimator of θ given in (12) when some of observations Z_{in} are near the boundary of Z. Local linear method, or more generally, the local polynomial method can be used to reduce this bias. Another alternative is to use *trimming* (Severini and Staniswalis, 1994) in which the function $S_n(\theta, g_0)$ is computed by using only observations associated to Z_{in} that are away from the boundary. The advantage of this approach is that the theoretical results can be presented in a clear form but it is less tractable from a practical point of view in particular for low sample sizes.

With some assumptions in place, we give weak consistencies and asymptotic normality results of the proposed estimators. Numerical experiments with Monte-Carlo simulations and real data application are given.

References

1. Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer Science & Business Media.
2. Arbia, G. (2006). *Spatial econometrics: statistical foundations and applications to regional convergence*. Springer Science & Business Media.

3. Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.
4. Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45.
5. Cressie, N. A. (1993). *Statistics for spatial data*. Wiley.
6. Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In *Advances in spatial econometrics*, pages 145–168. Springer.
7. Garthoff, R. and Otto, P. (2017). Control charts for multivariate spatial autoregressive models. *AStA Adv. Stat. Anal.*, 101(1):67–94.
8. Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.
9. Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *Internat. Econom. Rev.*, 40(2):509–533.
10. Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autore-gressive models. *Econometrica*, 72(6):1899–1925.
11. Pinkse, J. and Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85(1):125–154.
12. Poirier, D. J. and Ruud, P. A. (1988). Probit with dependent observations. *The Review of Economic Studies*, 55(4):593–614.
13. Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1):5–19.
14. Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American statistical Association*, 89(426):501–511.
15. Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of statistics*, pages 1768–1802.
16. Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional science and urban economics*, 40(5):292–298.
17. Su, L. (2012). Semiparametric gmm estimation of spatial autoregressive models. *Journal of Econometrics*, 167(2):543–560.
18. Wang, H., Iglesias, E. M., and Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics*, 172(1):77–89.

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 978-967-2000-65-5

#ISIWSC2019