

PROCEEDING

SPECIAL TOPIC SESSION

VOLUME 2




**62nd ISI WORLD
STATISTICS
CONGRESS 2019**

18 - 23 August 2019, Kuala Lumpur
Come | Connect | Create

PROCEEDING

**ISI WORLD STATISTICS
CONGRESS 2019**

**SPECIAL TOPIC SESSION
(VOLUME 2)**



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

Copyright of individual papers resides with the authors.

Suggested citation:

Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Special Topic Session: Volume 2, 2019. 413 pages

Disclaimer:

The views expressed in this proceeding are those of the author(s) of the respective paper and do not necessarily represent the opinions of other participants of the congress, nor the views or policy of the Department of Statistics Malaysia.

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

Preface	i
Scientific Programme Committee	ii
Local Programme Committee	iii
STS451: A Comprehensive Statistical Analysis on A Large-Scaled Scientific Citation Database - Web of Science		
Quantitatively analyze the capability of the organization:	1
Estimating the capability to induce innovation based on co-author information of articles		
ST452: Input-Output Analysis		
Economic diversification and sustainable development – A new assessment with input-output data	11
Value-added exports and trade-embodied carbon emission of China's industrial sectors with heterogeneity of firm size	21
Extended input-output model for demographic change - Preliminary application to the Chinese urbanisation	30
Using input-output tables to study trade and international production sharing arrangements	39
STS459: Statistical Methods and Applications		
The effects of macroeconomic variables on future credit ratings	48
Probability distribution model for predicting ozone (O3) exceedances at two air quality monitoring sites in Malaysia during dry season	56
STS461: Nordic Experiences in Coordination of Global SDG Indicator Compilation and Establishing Database Reporting Platforms		
Collaboration on SDG-data between national stakeholder groups	64
The development of national reporting platform for global SDG indicators in Finland	73
Coordinated communication for better follow-up of the 2030 Agenda in Sweden	81
STS463: Materializing Labour Account: Ways and Challenges		
The need to develop labour accounts in Malaysia: An assessment	88

Labour supply statistics: Challenges and way forward	99
STS466: Intelligent Data Warehouse Enrich Smart Statistics		
Managing unstructured data through Big Data Analytics towards intelligent insights	110
MyHarmony: Generating statistics from clinical text for monitoring clinical quality indicators	118
STS474: Spatial and Spatio-Temporal Analysis in Social Science		
A copula approach to spatial econometrics with applications to finance	124
Spatial extension of GARCH Models for high-dimensional financial time series	129
Analysis of regional economic growth against crisis: US-Japan statistical comparative study before-after Lehman's shock	137
On Gaussian semiparametric estimation for two-dimensional intrinsic stationary random fields	146
STS479: Leveraging South-South Cooperation in the Modernization of Official Statistics Towards Supporting the Production of SDG Indicators		
Prioritisation of Sustainable Development Goals and efforts of SESRIC to support statistical modernisation in OIC member countries	153
Morocco's experience in South/South cooperation and statistical capacity building	162
Modernization of the Tunisian Statistical System and its impact on statistical production for SDGs	168
Modernization and monitoring SDGs in area of conflicts or in fragile conditions: Case study, "The Palestinian Central Bureau of Statistics"	177
Modernisation of statistical systems - Experiences of GCC-Stat	185
STS480: Unleash the Value of Advanced Analytics in Insurance		
The influence of telematics device on driving behaviour of commercial vehicles across long and short haul drivers	194
STS486: Environmental Statistics and Climate Change		
Inspecting ecological communities structure via FDA	200

Recent advances in ecological networks: Regularized grouped Dirichlet-multinomial regression	208
Change detection and harmonisation of atmospheric large spatiotemporal series	217
STS489: Advances in Bayesian Spatio-Temporal Modelling of Disease Risk Based Complex Household Surveys in Sub-Saharan Africa (SSA)	
Geographic variation, trends and determinants of hypertension in South African adult population, 2008 -2017	222
Modelling and mapping prevalence of Female Genital Mutilation/C (FGM/C) among 0-14 years old girls in Kenya, Nigeria and Senegal	231
Spatial heterogeneity of childhood anaemia in four Sub-Saharan African countries	239
STS490: Setting Up Collaborative Support Systems Between Academic Institutions to Enhance Delivery on Industry-Integrated Skills Development Projects	
Professional statisticians/ data scientists: Who are they and how do we train them?	248
STS493: Vision on Future Data Collection for Official Statistics	
Advanced data collection – An outlook to the future	253
Modernizing data collection in Canada	263
Sensor data at the heart of innovation in official statistics	272
STS496: Safeguarding The Professional independence of Statisticians; The international Experience	
INEGI's Statistical Autonomy: Institutional Governance and Some Ever-Present Risks	284
The requirements for a well-functioning statistical system in a modern democratic society	293
Statisticians misbehaving: The ethical dimensions of an essential profession	300
Safeguards for the professional independence of Statisticians in Europe	309
STS497: Digital Economy: The Development of Industry 4.0	
Measuring the digital economy: Malaysia's experience	316

Record linkage for statistical business register data	324
The insights of e-commerce in Malaysia	332
STS498: Supervised and Unsupervised Learning for Modern Data Sets		
The GUIDE approach	340
STS500: Big Data in Official Statistics: A New Dimension for Operational Offices		
Profiling the internet economy in Singapore	346
Use of web-scraping for the compilation of Consumer Price Index: Malaysia's experience	351
Research on deepening the application of big data in government statistics	360
STS506: Bayesian Modelling of Public Health Data in the Presence of Spatial or Temporal Dependence		
On the use of surrogate models to speed the ABC inference for epidemic models in networks	368
STS507: Rising to the Challenges of a Changing Official Statistics World		
Adding statistical value in a rapidly changing government data "Eco-system"	377
STS508: The Affordable Living in Kuala Lumpur City-Region: Accelerating the Implementation of the New Urban Agenda		
Challenges in implementing a new imputation method into production in the 2017 Economic Census or what to do when the research approach oversimplifies the problem	386
The need for granular data in evidence-based policies: the case of housing affordability in Malaysia	396
Index	403



Quantitatively analyze the capability of the organization: Estimating the capability to induce innovation based on co-author information of articles



Yuji Mizukami¹, Keisuke Honda², Frederick Kin Hing Phoa³, Junji Nakano^{2,4}

¹ Nihon University, Chiba Japan

² The Institute of Statistical Mathematics, Tokyo Japan

³ Academia Sinica, Taipei Taiwan

⁴ Chuo University, Tokyo Japan

Abstract

Innovation is the act of creating new value by using "new connection", "new point of view", "new way of thinking", "new usage method". In recent years, the promotion of the Innovation has been strongly encouraged. In the field of research, attempts are also being made to create new value through connection between those fields. Moreover, along with the move to promote integration among these research fields, research is being conducted to grasp and promote the degree of them.

In this research, for the purpose of providing indices for measuring the degree of them, we show indices quantitatively indicating the degree of fusion in different fields. As an example of its application, we showed the degree of integration of different fields of the big data related article data from 2016 is utilized, which included 2544 articles from top 10 countries/regions by using the dissertation database Web of Science.

Keywords

Research Metrix; Institute Research; Network Theory; Co-author analysis; Innovation

1. Introduction

In 2011, the German Engineering Academy and the German Federal Ministry of Education and Science announced the framework of Industry 4.0. Industry 4.0 aims to improve the efficiency of factory production activities by promoting the spread of the Smart Factory [1]¹ based on the concept of the Cyber Physical System [2]². Studies have been conducted earlier to increase the efficiency of factory production activities. The advantage of Industry 4.0 is "predictive maintenance," which forecasts and repairs equipment failures and

¹ It implements "visualization" and "comprehensive management" of information by attaching sensors to every device in the factory and collecting information such as the quality and condition through the Internet.

² It aims to control the knowledge that is understood only by "experience and intuition" based on the quantitative measurement of the various states of the controlled object in the real world.

anomalies in advance. The Internet of Things (IoT), Big Data, and artificial intelligence (AI) technologies have been attracting attention as techniques for creating such advantages.

Different approaches are envisaged in research on the same subject. In this research, focusing on Big-data, we show what fields of researchers in different countries contribute to research in that field, And for the purpose of providing indices for measuring the degree of the Big Data. We showed the degree of integration of different fields of the Big Data related article data from 2016 is utilized, which included 2544 articles from top 10 countries/regions by using the dissertation database Web of Science.

The analysis is performed in two steps. First, the contribution of each field was qualitatively shown using MM-Index [3], which is a method of co-author analysis. Then, using the multiple comparison method presented in this paper, the contribution was quantitatively shown and international comparison was performed.

2. Literature Review and Background

2.1. Classification of research evaluation

The classification of institutional research (IR) was organized by Negishi et al. [4], Cho et al. [5], and Wagner et al. [6]. They divided IR into the following two categories: the number of relevant published articles, and the number of citations within and between these articles. This classification system is shown in Table 1. The purpose of analysing the number of articles is to determine the productivity index of each article and to obtain a scale index of the research activities. For this, a simple tabulation is primarily used. By doing this before analysing the citation statistics, it allows us to determine the consumption index of each article and the quality index of each of the research activities (Group 1). The consumption index is indicative of the degree of utilization, and to obtain this, a citation analysis or a co-author analysis is used [4]. An analysis of the acknowledgments, the co-words, or the attendant classification is also used for minor cases [5]. Furthermore, a simple tabulation is also primarily used. By doing this before analysing the specialized field analysis and diversity analysis, it allows us to determine the specialized field index of the article/author and the diversity index of each article/author (Group 2). Finally, the results of Group 1 and Group 2 allow us to determine the different field fusion indices of the articles/authors [6].

When analysing the number of articles, the betweenness and the following article analytics will be evaluated: field, year, country of origin, and affiliation. In addition, correlation analyses have been performed for other economic statistical indices [4]. For example, from an international comparison of article productivity in microbiology for the years 1995–

2003, Vergidis et al. [7] discovered that the productivity is highest in Western Europe, followed by North America, and that productivity in Asia, Central and South America, and Eastern Europe saw the highest growth rates during this period.

Citation analyses summarize the number of citations per article by the distinction of the academic journal and the country. The impact factor (IF) has been used to gauge the distinction of academic journals, and is derived from the Science Citation Index database of Thomson Reuters. The IF is based on the total number of articles published within the last two years and the number of times that those articles were cited in the selected year [4]. Vergidis et al. [7] performed an international comparison of the average IF in microbiology during the years 1995–2003. Their results showed that journals in North America had the highest IF (3.4), followed by Western Europe (2.8), and the average IF of other journals was 2.4. The IF is typically used for purchasing decisions in libraries, but it is considered unsuitable for evaluating personal achievements [4]. On the other hand, co-author analysis is appropriate for assessing the performance of an individual, and it can be used to evaluate the progress in collaboration between researchers at a given institution [5].

Diversity analyses have been suggested by Rafols and Meyer [8], Stirling [9], and Porter and Rafols [10]. Rafols and Meyer [8] introduced diversity indicators to describe the diversity of a bibliometric set viewed from pre-defined categories. This is using a structural approach that locates each of the elements (e.g. papers) of the set on one of the categories of the global map of science.

Table 1 Classification of institutional research evaluation

Category	Purpose	Analytical method
Analysis of the number of articles	<ul style="list-style-type: none"> - Productivity index of the article - Scale index of research activities 	- Simple tabulation
Analysis of the number of citations within and between these articles	Group 1 -Consumption index of the article -Quality index of research activities	- Citation analysis - Co-author analysis - (Acknowledgment analysis) - (Co-word analysis) - (Attendant classification analysis)
	Group 2 - Specialized field index of the article/author - Diversity index of the article/author	
	Different field fusion index of article/author [Groups 1 and 2]	- Specialized field analysis - Diversity analysis

Boldface indicates the subjects of this study. Parentheses are used to denote minor cases.

2.2. MM-Index

Mizukami et al. (2017) [1] aimed at objectively defining the specialization of the researcher, and defined the derivation method of the specialization based on co-author analysis.

2.2.1. Visualization of researchers' specialization fields and applied fields

Traditionally, the specialization fields chosen by researchers are based on each individual's offer; therefore, can be said to be a subjective definition. Also, because it is a subjective definition, in one case, the achievement in the specialized field was not accompanied, and the significance of the special field of researchers was not clear [2]. In response to this problem, Mizukami et al. [2] aimed at objectively defining the expertise of researchers, and they defined a method for deriving fields of specialization based on co-authored analysis.

Figure 1 (a) shows an example of the specialization field and the application fields of researchers. Researcher A's published papers include two articles in the mathematics field (Field 12), one article in the clinical medicine field (Field 4), one article in the economics and business fields (Field 6), and one article in the comprehensive field (Field 15). The field of specialization of researcher A is the mathematical field, and its concentration rate is assumed to be 40%. If the degree of concentration rate is high, the researcher is thought to concentrate substantially on research in that specialized field. However, if the degree of concentration rate is low, the researcher seems to have applied the research results of the specialization field to other fields (i.e., application fields).

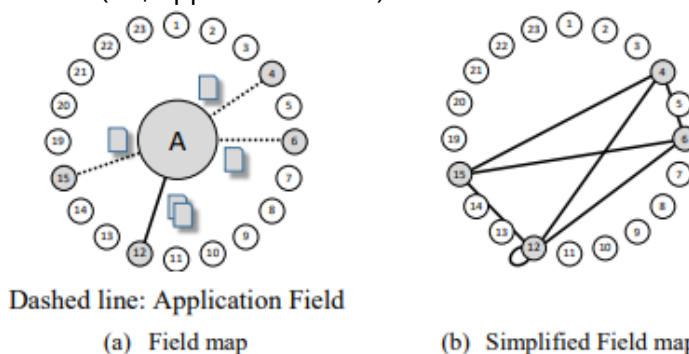


Figure 1 Specialized field and application fields of the researcher

Table 2 shows the classification of research fields used in this article. This classification is based on ESISA [11] posted in the Web of Science.

#	Subject Area	#	Subject Area	#	Subject Area
1	Agricultural Sciences	9	Geosciences	17	Pharmacology & Toxicology
2	Biology	10	Immunology	18	Physics
3	Chemistry	11	Materials Sciences	19	Plant & Animal Science
4	Clinical Medicine	12	Mathematics	20	Psychology/Psychiatry
5	Computer Science	13	Microbiology	21	Social Sciences, general
6	Economics & Business	14	Molecular Biology & Genetics	22	Space Science
7	Engineering	15	Multidisciplinary	23	Arts & Humanities
8	Environment/Ecology	16	Neuroscience & Behavior		

This classification is based on Essential Science Indicators Subject Areas of Web of Science [11].

2.2.2. Visualizing the distribution of an organization's research field

We propose methods for measuring the organizational research ability, and we also propose visualization methods for the organizational integration rates for different fields. To calculate the organizational research ability and the organizational integration rates of different fields, we superimpose information about researchers belonging to the organization, based on information about the specialization and application field of each researcher (see Figure 1 (a)). However, the information of each researcher shown in Figure 1 (a) does not show the connection between each research field unless it passes through the researcher located at the centre of the figure, and the linkage is not clear. Therefore, in this method, we used a simplified indication method for reconstructing the information about each researcher into the information between the fields. Figure 1 (b) shows an example of the simplified indication methods of researchers. In Figure 1 (b), the connection between each field is clarified.

3. Methodology

In this paper, the ratio difference in all the combinations of the fields is examined. Repetition of similar tests in this way is called multiple comparison and it is generally said that the probability of Type 1 error increases. For instance, if a test at the significance level 0.05 is performed 20 times, it will be rejected once potentially even those is in the adopted area. Therefore, a test is conducted by using multiple comparison method based on Bonferroni's

inequality. In other words, it is a method to correct the significance level by the number of connections between fields.

Since there are 23 research fields, there are 253 combinations between fields if there is no direction for connection. In this paper, 276 combinations are conducted, including 23 connections between the same fields. The connection between the same fields is used as the number of papers in the field.

The method of multiple comparison between organizations of MM-Index is shown. Formula 1 is the number of cross-disciplinary links with different dominance of ratios between organizations at significance level 0.05. The mining of the smaller value of RNSD is that the structure of the comparative organizations are resembles. In this formula, the maximum number of connections is calculated from the number of researchers in each specialized field, and convert the actual number of connections into a ratio.

$$RNSD_{a,b} = \frac{d_{a,b}}{\binom{n^2-n}{2}+n} = \frac{2d_{a,b}}{n^2+n} \dots\dots\dots (1)$$

RNSD_{a,b}: Ratio of significant differences between *a* and *b*

a: Organization

b: Organization

d_{a,b}: Significant differences between *a* and *b*

n: Number of research fields (In our case, it is 23.)

$$d_{a,b} = \sum_{i_s=1}^n \sum_{i_t=1}^n \begin{cases} 1, & \text{if } a > b \text{ and } i_s \leq i_t \text{ and } p(a,b,i_s,i_t) \leq s_b \\ 0, & \text{if others} \end{cases} \dots\dots\dots (2)$$

i_s: Research field

i_t: Research field

p(a, b, i_s, i_t): Test result between *i_s* and *i_t* (Research fields) of between *a* and *b* (Organizations)

s_b: Bonferroni's Significance level

$$s_b = \frac{s}{\binom{n^2-n}{2}+n} = \frac{2s}{n^2+n} \dots\dots\dots (3)$$

s: Significance level (In our case, it is 0.05)

$$p(a, b, i_s, i_t) = \text{prop. test} \left(\frac{l(a,i_s,i_t)}{c(a,i_s)c(a,i_t)}, \frac{l(b,i_s,i_t)}{c(b,i_s)c(b,i_t)} \right) \dots\dots\dots (4)$$

l: Number of links (Organization, Research field, Research field)

c: Number of articles/researchers (Organization, Research field)

4. Data source

The big data related article data from 2016 is utilized, which included 2544 articles from top 10 countries/regions by using the dissertation database Web

of Science. The country rankings based on the number of articles are shown in **Table 3**.

Table 3 Top 10 countries ranked according to the number of big data articles (2016)

Rank	Country/Reason	Number of Articles	Rank	Country/Reason	Number of Articles
1	USA	894	6	CANADA	127
2	CHINA	625	7	SOUTH KOREA	118
3	UK*1	242	8	ITALY	95
4	AUSTRALIA	140	9	SPAIN	95
5	GERMANY	137	10	JAPAN	71
			-	Top 10 Total	2544

*1: UK consists of England, Scotland, and Northern Ireland is a member of the EU.

*2: NA & SA are abbreviations of North America and South America.

5. Result

In this section, the quantitatively analyze the degree of country-specific approximation of MM-Index using the data of Big data. The new knowledge added is the introduction of multiple comparison method based on Bonferroni's inequality, and its related method. Figure 2 shown Big data's individual cooperation patterns between subject areas (2016).

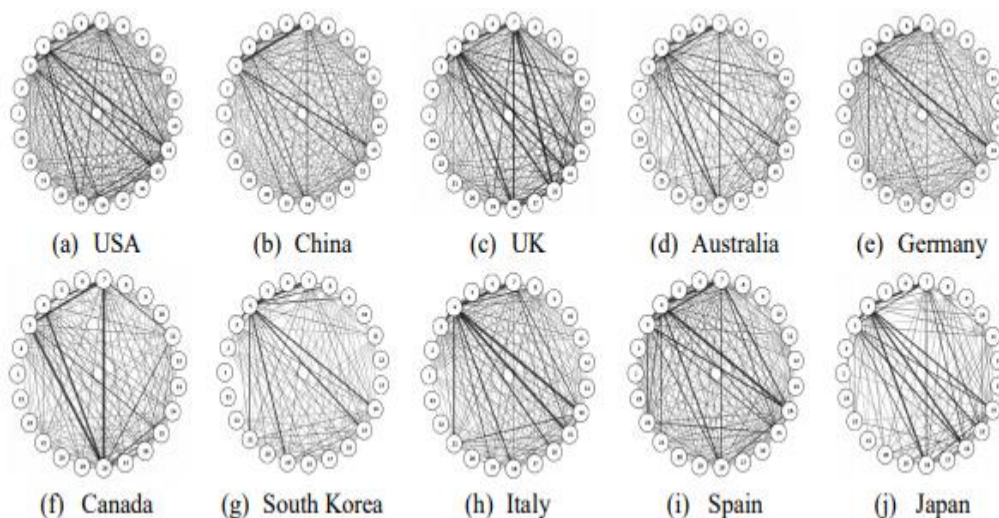
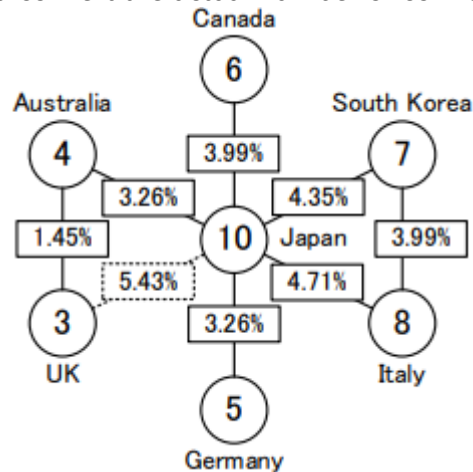


Figure 2 Big data's individual cooperation patterns between subject areas (2016)

The feature of this figure lies in determining the thickness of each connection line based on the maximum value in the connection between specialized fields. On the other hand, since there are different number of researchers in each specialized field, when evaluating the strength of the

connection the comparison is required considering both the number of researchers in each specialized field and the number of links between specialized fields.

Formula 1 is able to solve this problem. In this formula, the maximum number of connections is calculated from the number of researchers in each specialized field, and convert the actual number of connections into a ratio.



Attention: There was no significant result among the other countries at the significance level 0.05.

Figure 3 Approximation degree by country in Big Data

The analysis result is shown in Figure 3. The international comparison of related fields of Big data was conducted, showing that the research co-author relationships in Japan, Australia, Canada, Korea, Italy, Germany are very similar

6. Discussion and Conclusion

Different approaches are envisaged in research on the same subject. In this research, focusing on Big-data, we show what fields of researchers in different countries contribute to research in that field, And for the purpose of providing indices for measuring the degree of the Big Data. We showed the degree of integration of different fields of the Big Data related article data from 2016 is utilized, which included 2544 articles from top 10 countries/regions by using the dissertation database Web of Science.

The analysis is performed in two steps. First, the contribution of each field was qualitatively shown using MM-Index [3], which is a method of co-author analysis. Then, using the multiple comparison method presented in this paper, the contribution was quantitatively shown and international comparison was performed. The international comparison of related fields of Big data was conducted, showing that the research co-author relationships in Japan, Australia, Canada, Korea, Italy, Germany are very similar.

Using the multiple comparison method presented in this paper, we were able to quantitatively show the contribution of each field in co-author analysis.

There are two directions for future research. First, increase the number of analysis cases and examine the effects of these methods. Next is to broaden the scope of application of these techniques.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP17K04710 and the Institute of Statistical Mathematics joint research program (2019-ISMCRP-1026). The data were provided by Clarivate Analytics.

References

1. Kagermann, H., W. Wahlster and J. Helbig, eds., 2013: Recommendations for implementing the strategic initiative Industrie 4.0: Final report of the Industrie 4.0 Working Group
2. Yuji Mizukami, Keisuke Honda and Junji Nakano, Study on Research Trends on the Internet of Things Using Network Analysis, International Journal of the Japan Association for Management Systems, Vol.10, No.1, pp.37-45, 2018
3. Mizukami, Y., Mizutani, Y., Honda, K., Suzuki, S., Nakano, J. (2017). An International Research Comparative Study of the Degree of Cooperation between disciplines within mathematics and mathematical sciences, Behaviormetrika, Vol. 1, 19 pages, On-line.
4. Negishi, M., Yamazaki, S., Sun, Y., & Nishizawa, M., "Research Evaluation," Maruzen, ISBN-10: 4621048902, 2001 (in Japanese)
5. Cho, M., Fujigaki, Y., Hirakawa, H., Tomizawa, H., Hayashi, T., & Makino, J., "Scientometrics Introduction for Research Evaluation and Science Studies," Maruzen, ISBN-10: 4621073990, 2004 (in Japanese)
6. Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J. & Börner, K., "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature", Journal of Informetrics, Vol. 5, No. 1, pp. 14-26, 2011
7. Vergidis, P., Karavasiou, A., Paraschakis, K., Bliziotis, I., & Falagas, M., "Bibliometric analysis of global trends for research productivity in microbiology," European Journal of Clinical Microbiology and Infectious Diseases, Vol. 24, No. 5, pp. 342–346, 2005
8. Rafols, I., and Meyer, M., "Diversity measures and network centralities as indicators of interdisciplinarity: Case studies in bionanoscience", Scientometrics, Vol. 82, pp. 263–287, 2010
9. Stirling, A., "A general framework for analyzing diversity in science, technology and society", Journal of the Royal Society, Vol. 4, pp. 707–719, 2007

10. Porter, A. L., and Rafols, I., "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time", *Scientometrics*, Vol. 81, pp. 719–745, 2009
11. Clarivate Analytics, Essential Science Indicators Subject Areas <https://clarivate.com/products/essential-science-indicators/> (2019, 1 April)



Economic diversification and sustainable development – A new assessment with input-output data



Joerg Beutel

Konstanz University of Applied Sciences, Konstanz, Germany

Abstract

For decades, exports and imports of most countries have grown more rapidly than domestic production. This is a strong indication that, besides foreign trade in final products, trade in intermediates is becoming increasingly important. Globalization in production is changing the way in which nations interact, and any analysis of diversification should therefore also encompass the worldwide exchange of intermediates in production. For this reason, an input-output approach, which accounts for the role of intermediates, is more appropriate for any analysis of diversification than a traditional approach based purely on macroeconomic data.

This article analyses the main trends in foreign trade, value added chains and economic diversification for the ten largest economies (G10) and eight ASEAN Countries using data from input-output tables and national accounts.

It also assesses the relative progress on sustainable development of these countries using the measure 'adjusted net savings' of the World Bank. It measures the true rate of savings in an economy after accounting for investments in man-made physical and human capital, depletion of natural resources, and damage from environmental pollution. This view of sustainable development requires that the country pass on an aggregate stock of physical, human, and natural capital to the next generation that is not smaller than the one that currently exists. This requires that the loss of depleting resources and environmental damage be offset by increasing the stock of physical and human capital.

The article concludes that economic diversification of 8 ASEAN countries reached high levels during the last 10 years. Only the diversification of Brunei Darussalam and Singapore showed volatile behaviour towards less concentration of industries. The test for sustainable development of ASEAN countries for the period 1995-2016 showed mixed results. For Laos the test even failed.

Keywords

Supply, use and input-output tables; trade in intermediates, structural change, global value chains, adjusted net savings

1. Trends in Production

Since decades exports, imports and intermediates are growing more rapidly in most countries than domestic production (GDP). Exchange of intermediates in production becomes more and more important. Globalisation in production is changing the way in which nations interact. The traditional analysis of diversification is mainly based on the structural change of value added by industries. The analysis of diversification should also encompass the worldwide exchange of intermediates in production. An input-output approach is more appropriate for the analysis of diversification than the traditional approach with macroeconomic data.

If exports and imports are growing faster than gross domestic product (GDP), the shares of exports and imports in GDP are also increasing. We explored UNdata for the 10 largest economies of the world (G10) and 10 ASEAN countries during the last 22 years.

For the G10 during 1995-2017, the increase of the export share in GDP was 5.3 per cent and of the import share in GDP 5.8 percent.

The export share in GDP increased in percent in Germany by 25.0, Japan 8.7, France 8.3, India 7.9, Italy 6.4, Brazil 5.2, United Kingdom 5.1, USA 1.4 and China 0.6. Decreases were only observed in Canada -5.1. The import share in GDP increased in percent in Germany by 17.9, France 11.0, India 9.7, Japan 9.0, Italy 7.1, United Kingdom 6.5, USA 3.2, Brazil 2.1, China 0.5 and Canada 0.04.

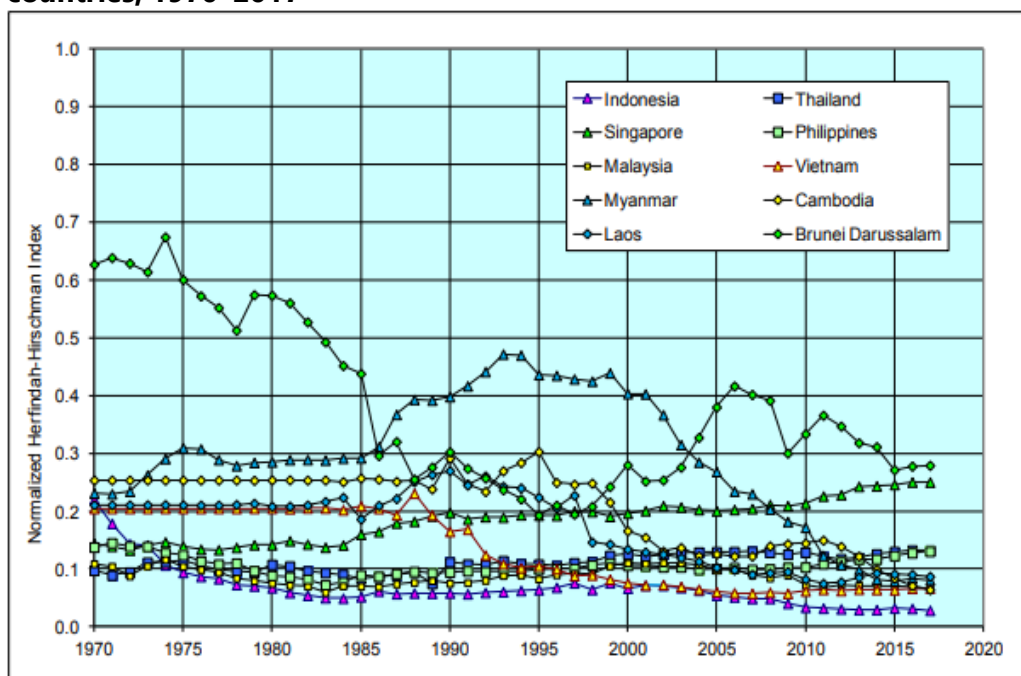
During the last 20 years, economic globalization has increased worldwide interdependencies in production, leading to the more intermediate consumption of goods and services in the international chain of value added. If the consumption of intermediate products is growing above its GDP growth rate, an economy is moving towards more complex participation in inter-industrial production. We extracted the information on intermediates, value added and output from the input-output database of OECD (2018).

In 1995–2015, the share of intermediates in total output for the G10 countries increased by 5.9 per cent—in other words, their production processes became more complex and more interdependent. The share increased in India in percent by 7.7, China 6.7, Germany 3.4, Brazil 2.0, Japan 1.6, Canada 1.2, France 0.8 and Italy 0.8. At the same time a decline of the share in percent was observed in United Kingdom -3.6 and USA -2.7.

2. Economic Diversification of ASEAN Countries

Sustainable development of nations involves economic, social and environmental changes. Within this process, diversification and structural change of production and demand are closely related to many areas of the economy and society. For income per capita to converge: Countries must move towards more diversified, more complex production structures with more technology and knowledge.

Herfindahl-Hirschman Index of industry diversification for ASEAN countries, 1970–2017



Data from: National Accounts Estimates of Main Aggregates—United Nations Statistics Division.

Note: Index values were calculated based on data from the following seven industries: agriculture, hunting, forestry and fishing; mining and utilities; manufacturing; construction; wholesale and retail trade, restaurants and hotels; transport, storage, and communication; and other activities.

Economic diversification means the diversification of exports, imports, and domestic production away from extreme dependence on a single dominant industry or a few natural-resource-based products, as well as a change toward increased complexity and quality of output (Beutel, 2012).

The Herfindahl-Hirschman index is the most-widely used measure to evaluate market concentration and diversification of an economy. The index is the sum of squared shares of the various industries in gross value added. In the normalized form, the index varies from 0 to 1. In case of a low value, the economy has a large number of industries with similar value added shares indicating high diversity. If the index reaches 1, only one sector accounts for all value added and a high concentration of economic activity is in place. The decline of the index signifies less concentration in the dominant industry and greater diversification of other industries. The figure below plots the index for recent decades for ASEAN countries.

During the last 47 Years all ASEAN countries except Singapore increased the diversification of industries. From 1990 onwards high and stable levels of diversification are reported for four countries: Indonesia, Malaysia, Philippines

and Thailand. From 1995 onwards, Cambodia steadily reduced the concentration of industries. Large and volatile fluctuations of diversification are observed for Brunei Darussalam and Myanmar. In Vietnam and Laos the concentration of industries decreased since 1990, while in Singapore the concentration of industries steadily increased since 1970.

3. New Assessment with Input-Output Data

The best way of measuring the relationship between intermediates, gross value added and final demand is through the use of supply, use and input-output tables. These tables have received much attention in recent years. They offer new opportunities to fully understand the 'global value chains' and their impact on production, consumption, investment, employment and environment.

Supply and use tables are an integral part of the System of National Accounts 2008. They mainly serve statistical purposes. The system of supply and use tables ensures the consistency of data obtained from different kind of statistical sources. Input-output tables are derived from supply and use serving as a well-established tool for various analytical purposes related to production and as database for macroeconomic models.

While supply and use tables are data-oriented in nature, the symmetric input-output tables are always constructed from having made certain analytical assumptions from existing supply and use tables. The best way to measure the relationship between intermediate consumption, gross value added, and final demand is through the use of input-output tables, which are derived from supply and use tables that are an integral part of the System of National accounts (Beutel, 2017).

In May 2018 the United Nations Statistics Division published the final draft of the Handbook on Supply, Use and Input-Output Tables with Extensions and Applications (United Nations, 2018). The new Handbook explains in great detail how supply, use and input-output tables can be compiled from the main statistical sources. I was a member of the Editorial Board and drafted several chapters of the Handbook.

Extended Input-Output Tables provide the information which is required to assess diversification and sustainable development. They comprise useful information of satellite systems which are integrated into the National Accounts. They often include information on investment, capital and labour. However, additional information on energy, emissions, natural resources, waste, sewage and water are also needed and can be added to the tables as well.

The extended input-output table of Germany as presented in the new UN Handbook (United Nations, 2018, p. 518) has the following seven extensions with information in values and quantities:

1. Gross fixed capital formation (Million Saudi Riyals)
2. Capital stock (Million Saudi Riyals)
3. Employment (1.000 persons)
4. Energy use (1.000 tons of oil equivalent)
5. Air emissions (1.000 tons)
6. Global warming, acid deposition, tropospheric ozone formation (1.000 tons)
7. Water use (Million cubic meters)

The traditional input-output indicators comprise direct input coefficients, cumulative input coefficients (Leontief inverse) and multipliers for output, income, employment and capital. Inter-industrial linkage analysis studies the interdependencies between industries by compiling forward and backward linkages of industries

Direct input coefficients reflect the direct input requirements of products for a specific industry, while cumulative input coefficients represent both direct and indirect input requirements of products at all stages of production. Cumulative input coefficients are often used to identify an industry's backward linkages. The column totals of the direct input coefficients and the Leontief Inverse input coefficients reflect the intensity of backward linkages. The row totals of the direct output coefficients and the Ghosh Inverse output coefficients show the intensity of forward linkages.

Economic diversity has often been promoted as a means to achieve economic stability and growth. Some empirical studies have related higher levels of diversity to both economic stability and overall levels of economic activity. Diversity measures used in these studies have tended to be narrow, usually emphasizing the distribution of employment across industries. Such measures are unsatisfactory because they do not capture inter-industrial linkages.

An alternative approach to measuring diversity, based on the technical coefficients matrix of an input-output model, was developed by Wagner and Deller (1998), who showed that higher levels of diversification within the theoretical construct of input-output are associated with higher levels of stability. Their Primary Diversity Measure (PDM) emphasizes inter-industry relations and provides the best way to evaluate economic diversification. It is derived by multiplying values assigned to three variables:

- Relative size of the economy—number of indigenous industries
- Density of the economy—number of non-zero elements in the Leontief matrix, indicating the diversity of transactions
- Condition number of the Leontief matrix—indicator of inter-industry linkages.

The Primary Diversity Measure (PDM) was applied by Al-Kawaz (2008) for Kuwait in 2000 and by Beutel (2019) for Kuwait and Saudi Arabia during 1995-2011.

4. Sustainable Development of ASEAN Countries

Sustainable development of nations involves economic, social and environmental changes. Within this process, diversification and structural change of production and demand are closely related to many areas of the economy and society. For income per capita to converge, countries must move towards more diversified, more complex production structures with more technology and knowledge. The long-term strategy for countries is to increase the gross national income per capita and to transform the non-renewable natural capital into other forms of capital like machinery, buildings and human capital.

Since a long time, the World Bank is engaged in measuring sustainable development of nations. The long-term strategy of countries should be to increase the gross national income per capita and transform the non-renewable natural capital into other forms of capital like machinery, buildings, and human capital (Beutel, 2013).

In the World Bank's World Development Indicators we find two prominent indicators for sustainable economic development: 'Adjusted net national income' and 'Adjusted net savings'.

'Adjusted net national income' is estimated by subtracting from gross national income the consumption of fixed capital and depletion of natural resources. The consumption of fixed capital reflects the decline in man-made physical capital through retirement of buildings, machinery, transport equipment, and the like; while the depletion of natural resources measures the decline in non-renewable natural resources through extraction.

Gross domestic product (GDP)
 + Net income from abroad
 = Gross national income (GNI)
 - Consumption of fixed capital
 = Net national income
 - Natural resources depletion
 = Adjusted net national income

The consumption of fixed capital is estimated as part of the national accounts. On the depletion of natural resources, the World Bank provides valuable information for 10 minerals, 4 energy sources, and net forest depletion.

'Adjusted net savings' is a national accounting aggregate designed to measure changes in assets including natural and human capital. The gross stock of natural capital (natural resources), physical capital (buildings,

machinery, transport equipment), and human capital (education, skills, knowledge) is growing if a nation's adjusted net savings are positive.

There is an intrinsic link between change in the wealth of a nation and the sustainability of its development path. If genuine (adjusted) savings are negative at a given point in time, then welfare in the future will be less than current welfare. Therefore, adjusted net savings can be regarded as a sustainability indicator.

The World Bank calculates adjusted net national savings as follows:

Gross national savings
 – Consumption of fixed capital
 = Net savings
 + Education expenditure
 – Energy depletion
 – Mineral depletion
 – Net forest depletion
 – Carbon dioxide emissions damage
 – Particulate emissions damage
 = Adjusted net savings (genuine savings)

The calculation of adjusted net national savings begins with gross national savings, calculated as gross national income minus total consumption plus net transfers from abroad. Deducting consumption of fixed capital from gross national savings, we arrive at net national savings. Finally, education expenditure (considered as investment into human capital) is added, and depletion of natural resources and damage from pollution are deducted.

The World Bank adds all current operating expenditures for education to net savings as a gross investment in human capital. I believe that it would also be appropriate to deduct the consumption of human capital, as is done for the consumption of physical capital. Consumption of fixed capital reflects the value of the retired physical capital. The pensions of persons who worked in the education system could be regarded as consumption of human capital. In this case, consumption of human capital corresponds to the costs for the retirements of personal in education.

An economy is sustainable if it saves more than the depreciation on its man-made and natural capital—in other words, if its net national savings measurement is positive.

Adjusted net national savings as a test for sustainable development: all ASEAN countries, 2016

	Indonesia	Thailand	Singapore	Philippines	Malaysia	Viet Nam	Myanmar	Cambodia	Laos	Brunei Darussalam	ASEAN
Gross national income (GNI)	902,417	392,423	293,765	367,012	288,414	196,687	61,780	18,788	15,126	12,236	2,548,648
	Million US \$ at current prices										
	% of gross national income (GNI)										
Gross national income (GNI)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
- Final consumption	-69.6	-70.2	-49.2	-70.5	-69.4	-78.3	-71.7	-86.6	-83.2	-44.1	-68.2
+ Net transfers	0.5	1.7	-2.1	6.7	-1.6	4.1	3.5	8.3	1.7	-3.3	1.5
= Gross national savings (GNS)	<u>30.9</u>	31.5	48.7	36.2	29.1	25.7	31.9	21.7	18.4	52.6	33.2
- Consumption of fixed capital	-17.0	-18.6	-14.8	-7.5	-17.3	-12.3	-4.4	-10.3	-15.7	-8.7	-14.9
= Net national savings	13.9	12.9	33.9	28.7	11.8	13.5	27.5	11.3	2.7	43.8	18.3
+ Education expenditure	3.3	4.1	2.7	1.8	4.8	4.6	0.7	2.0	3.0	4.0	3.3
- Energy depletion	-0.9	-0.6	0.0	-0.1	-1.7	-0.7	-0.4	0.0	0.0	-5.4	-0.7
- Mineral depletion	-0.4	0.0	0.0	-0.6	-0.1	-0.1	-0.4	0.0	-3.1	0.0	-0.3
- Net forest depletion	0.0	-0.5	0.0	-0.2	0.0	0.0	-0.7	-1.1	-2.9	0.0	-0.1
- Carbon dioxide damage	-1.8	-2.7	-0.7	-1.0	-2.9	-3.0	-1.1	-1.4	-0.4	-2.5	-1.9
- Particulate emission damage	-0.5	-0.2	-0.1	-0.5	-0.2	-0.3	-0.7	-0.8	-1.2	0.0	-0.3
= Adjusted net national savings	13.7	13.1	35.9	28.2	11.7	14.0	24.8	10.0	-1.9	39.9	18.3
	Population (Mio. persons)										
Population	261.115	68.864	5.607	103.320	31.187	94.569	52.885	15.762	6.758	0.423	640.492
	US \$ per person										
Gross national income (GNI)	3,456	5,699	52,390	3,552	9,248	2,080	1,168	1,192	2,238	28,913	3,979
Adjusted net national savings	472	745	18,788	1,001	1,082	291	290	120	-43	11,527	728

Source: The World Bank - World Development Indicators, February 2019

Long time series for 1995-2017 on adjusted net national income and adjusted net national saving can be extracted from the World Bank's World Development Indicators (World Bank, 2018).

An economy is sustainable if it saves more than the depreciation on its man-made physical and natural capital. In the Table above, a test for sustainable development has been made for all ASEAN countries for the year 2016. The results show large difference in sustainable development.

Adjusted net national saving is the key variable for sustainable development. The shares of adjusted net national saving in gross national income ranged from 39.9 % in Brunei Darussalam to - 1.9 % in Laos. The test for sustainable development failed in Laos. High values for adjusted net national saving per person are observed in Singapore (18,788 \$/person) and in Brunei Darussalam (11,527 \$/person). Low values for adjusted net national saving per person were realised in Laos (-43 \$/person), Cambodia (120 \$/person) and Myanmar (120 \$/person). All ASEAN countries combined had the following savings ratios of GNI: Gross national savings 33.2%, net national savings 18.3%, and adjusted net national savings 18.3%. The positive allocation

for education expenditures was fully absorbed by the allocation for the depletion of natural resources and other damages.

5. Discussion and Conclusion

Since 1970, all ASEAN countries except Singapore increased the diversification of industries. High and stable levels of diversification are reported for four countries: Indonesia, Malaysia, Philippines and Thailand. From 1995 onwards, Myanmar and Cambodia steadily reduced the concentration of industries. Large and volatile fluctuations of diversification are observed for Brunei Darussalam and Myanmar. In Vietnam and Laos the concentration of industries decreased since 1990, while in Singapore the concentration of industries steadily increased since 1970.

The test for sustainable development failed in Laos. A low positive savings ratio was turned into a negative one after the allocation for the mineral depletion and net forest depletion. All other ASEAN realised positive ratings. However, the differences of the adjusted net saving share in gross national income were very large, reaching from 39.9 percent in Brunei and 35.9 percent in Singapore to 10.0 percent in Cambodia and -1.9 percent in Laos.

A full implementation of the input-output approach will only be possible if comparable supply and use tables become available for all ASEAN countries. Ideally they should use the same classification of the System of National Accounts 2008 (United Nations, 2009). At the moment, OECD input-output tables for 1995-2015 are available for 8 ASEAN countries. The tables for Myanmar and Laos are still missing. The Asian Development Bank (2019) published input-output tables for 9 ASEAN countries excluding Myanmar covering 2010-2017. The national statistical offices of ASEAN countries should be encouraged to compile annual supply, use and input-output tables as an integral part of their national accounts.

References

1. Al-Kawaz, Ahmed (2008): Economic Diversification: The Case of Kuwait with Reference to Oil Producing Countries, in: *Journal of Economic Cooperation*, 29, pp.23-48.
2. Asian Development Bank (2019): *Data Library, Input-Output Economic Indicators*.
3. Beutel, Joerg (2012): Conceptual Problems of Measuring Economic Diversification as Applied to the GCC Countries, in: Giacomo Luciani (ed.): *Resources Blessed: Diversification and the Gulf Development Model*, Gulf Research Centre, Gerlach Press, pp. 29-70.
4. Beutel, Joerg, Isabelle Rémond-Tiedrez, José M. Rueda Cantuche (2013): The Importance of Input-Output Data for the Regional Integration and Sustainable Development of the European Union, in: Joy Murray and

- Manfred Lenzen (eds.): The Sustainability Practitioner's Guide to Multi-Regional Input-Output Analysis, Champaign, Illinois, USA, pp. 220-239.
5. Beutel, Joerg (2017): The supply and use framework of national accounts, in: Thijs ten Raa (ed.): Handbook of Input-Output Analysis, Cheltenham, UK, pp. 41-129.
 6. Beutel, Joerg (2019): Economic diversification and sustainable development of GCC Countries, 2018 Gulf Research Meeting, University of Cambridge, forthcoming in Palgrave Macmillan, p. 39.
 7. OECD (2018): Harmonised National Input-Output Tables.
 8. United Nations (2018): Handbook on Supply, Use and Input-Output Tables with Extensions and Applications.
 9. Wagner, John E.; Steven C. Deller (1998): Measuring the Effects of Economic Diversity on Growth and Stability, in: Land Economics, Vol. 74, No. 4, pp. 541-556.
 10. World Bank (2018): World Development Indicators, Washington, DC.



Value-added exports and trade-embodied carbon emission of China's industrial sectors with heterogeneity of firm size



Yang Cuihong^{1,2*}, Zhang Junrong^{1,2}

¹ Academy of Mathematics and Systems Science, CAS, Beijing, China

² University of Chinese Academy of Sciences, Beijing, China

Abstract

In recent years, with the rapid growth of the economy, China has become one of the largest CO₂ emitters as well as the important exporter and importer in the world. How to effectively reduce carbon emissions generated by China's trade has aroused wide attention of governmental agencies and the academia. However, the existing studies exploring the trade-embodied carbon emission abatement solutions are based on the national level or the industry level, which are lack of in-depth research from the firm level. In fact, all the policies such as foreign trade and carbon emission reduction rely on the specific implementation of firms following a policy process in a bottom-up way. Specifically, the rapid development of China's economy has bred a large number of small and medium-sized firms. Under this background, it is of great significance to study the respective roles of different sized firms in China's economy and environment. In this study, we construct a newly extended Chinese input-output model (*LMS* model) for the year 2012, considering the heterogeneity of firm size, to explore the contributions on value-added exports as well as trade-embodied carbon emission of different sized firms in China. The results show that small and medium-sized firms play a crucial role in promoting China's economic development, both generating more than 60% of China industrial sector's total output and value-added. Meanwhile, small and medium-sized firms also produce more than half of the China industrial sector's total emissions, acting as a dominant driving factor to the China's carbon emissions. This study also investigates that in China's foreign trade, the value-added created by per unit input of medium-sized firms in industrial sector is higher than the other firm types. While small-sized firms have the highest embodied carbon emissions generated by per unit export, which fully demonstrates that the export products structure of small-sized firms is facing with optimizing and upgrading in the future. This study may provide some differentiated policy implications at firm level which can help to promote trade mode and carbon emission abatement in China.

Keywords

Firm heterogeneity by size; input-output table; value-added; carbon emission; China

1. Background

The problem of global warming caused by the consumption of fossil energy has become a social focus all over the world. In recent years, with the globalization of trade and production, international trade carry a large amount of carbon dioxide embodied in the goods production chain. According to WTO data, the total value of global exports of goods in 1990 was only 3449 billion USD, and in 2017 it has reached 17729 billion USD, with an increase of more than 5 times. The accompanying problems of environmental pollution with the international trade have become increasingly prominent. Under this circumstance, there has been a rapid increase in the study of trade embodied carbon emissions of different countries (Wyckoff and Roop, 1994; Ahmad and Wyckoff, 2003; Peters and Hertwich, 2006; Weber and Matthews, 2007; Peters et al., 2011; Liu et al., 2016; Deng and Xu, 2017).

Specifically, with the rapid development of the economy, China has become one of the largest CO₂ emitters in the world, and also one of the countries with the greatest energy consumption. At the same time, since the implementation of the reform and opening up policy, China's foreign trade has developed rapidly. In 2017, China's foreign trade amounted to USD 4105 billion, ranking the first in the world. The gross export amounted to USD 2263 billion, accounting for 12.8% of the total world exports. China, as a big country in the world trade, produces a large amount of embodied carbon emissions from a large number of energy-intensive products. How to effectively reduce trade-embodied carbon emissions in China has aroused wide attention of government, experts and scholars all over the world. There has been a large number of domestic and foreign scholars carried out research on China's trade and carbon emissions (Weber et al., 2008; Yunfeng and Laike, 2010; Zhao and Liu, 2011; Su and Ang, 2014; Zhao et al., 2016; Huang and Zhao, 2018).

However, the existing studies exploring the China's trade-embodied carbon emissions are based on the national level or the industry level, which lack of in-depth research from the firm level. In fact, although all the policies such as foreign trade and carbon emission reduction are carried out by government, they have to rely on the specific implementation of firms, so it is essential to start from the firm level to solve these problems from bottom to top. However, there are obvious differences in production structure, technology level, import and export trade, energy consumption, carbon emissions and economic impact among different types of firms, so to a large extent, the implementation of a trade or emission reduction policy will produce different feedbacks for different types of firms. Especially with the rapid development of the China's economy, many kinds of diversified firms have been cultivated. Therefore, it is extremely necessary to study trade and carbon emission from the perspective of firm heterogeneity.

At present, a small number of scholars have studied China's foreign trade and carbon emissions from the perspective of heterogeneity of firm ownership and foreign trade mode (Dietzenbacher et al., 2012; Ma et al., 2015; Jiang et al., 2016). However, it is noteworthy that the vigorous development of China's economy in recent years has bred a large number of small and medium sized firms. According to the data of the National Bureau of Statistics (NBS), the total number of small and medium-sized firms accounting for 97% of the industrial total firms in 2017, which dominates in terms of quantity¹. The data shows that China's small and medium-sized firms have contributed 60% of China's GDP, 50% tax and 80% of urban employment in 2018². Therefore, small and medium-sized firms have become an important force in China's economy and played a significant role in promoting the China's economy progress.

Accordingly, there are significant differences in the production structure, technical level, import and export trade, energy consumption, carbon emissions, and economic impact among firms of different size types. Under this background, it is of great significance to study the respective roles in China's economy and the contributions on the value-added and trade-embodied carbon emission patterns of different sized firms. Therefore, based on the non-competitive table, we construct an extended Chinese input-output model (*LMS* model) for the year 2012, considering the heterogeneity of firm size, to explore the value-added as well as trade-embodied carbon emission of different sized firms in China. This study may provide some differentiated policy implications at firm level on the following issues: What are the respective roles of different sized firms in China's economy and environment? Are there any significant differences in the value-added among large, medium and small firms? Who is the main driving force of China's trade-embodied carbon emissions?

2. Methodology

2.1 LMS model

The traditional Chinese input-output model does not concern the type of firm size, but there are significant differences in production technology, energy efficiency and import and export among the big, medium and small-sized firms. Therefore, an extended input-output model capturing the heterogeneity of firm size is constructed in our study, based on the 2012 non-competitive input-output table. The model distinguishes the large, medium and small firms of industrial sectors, abbreviated as *LMS* model (Table 1).

¹ Available at: <http://www.stats.gov.cn/tjsj/ndsj/2018/indexch.htm>

² Available at: http://www.gov.cn/guowuyuan/2018-08/20/content_5315204.htm

In *LMS* model, China's production activity is partitioned into three parts: namely, large firms production (*L*), medium firms production (*M*), and small firms production (*S*).

$$\text{Denote, } A^D = \begin{bmatrix} A^{LL} & A^{LM} & A^{LS} \\ A^{ML} & A^{MM} & A^{MS} \\ A^{SL} & A^{SM} & A^{SS} \end{bmatrix}, F = \begin{bmatrix} F^{LD} \\ F^{MD} \\ F^{SD} \end{bmatrix} + \begin{bmatrix} F^{LE} \\ F^{ME} \\ F^{SE} \end{bmatrix}, X = \begin{bmatrix} X^L \\ X^M \\ X^S \end{bmatrix}$$

Define direct input coefficient $A_D = Z^D(\hat{X})^{-1}$, \hat{X} is diagonal matrix of output value X , row vector of value-added ratio is $A_V = V(\hat{X})^{-1}$, we then have:

$$X = (1 - A_D)^{-1}F \tag{1}$$

Table 1 China's Non-Competitive Input-Output Table Capturing firm size heterogeneity (LMS model)

Output		Intermediate use			Final use		Total Gross Output
		Large Firm	Medium Firm	Small Firm	Domestic Final Demands	Exports	
Input		(L)	(M)	(S)	(FD)	(FE)	(X)
	Domestic Intermediate Inputs (ZD)	Large Firm (L)	Z^{LL}	Z^{LM}	Z^{LS}	F^{LD}	F^{LE}
Medium Firm (M)		Z^{ML}	Z^{MM}	Z^{MS}	F^{MD}	F^{ME}	X^M
Small Firm (S)		Z^{SL}	Z^{SM}	Z^{SS}	F^{SD}	F^{SE}	X^S
Imported Intermediate Inputs (ZN)		Z^{NL}	Z^{NN}	Z^{NS}	F^{ND}	F^{NE}	X^N
Value-added (V)		$V^{L'}$	$V^{M'}$	$V^{S'}$			
Total Gross Inputs (X)		$X^{L'}$	$X^{M'}$	$X^{S'}$			

Note: Table 1 gives a typical non-competitive I/O table which distinguishes domestic inputs and imported inputs. The superscript *L, M, S* denotes products from large-sized firms, medium-sized firms, and small-sized firms; *ZD* and *ZN* denote domestic and imported intermediate inputs, respectively; *FD* and *FE* denotes final demands vectors for domestic products and exported ones, respectively; *X* and *V* denotes total output (input) and primary inputs vector, respectively.

Then, value-added generated by export and final demands can be calculated as:

$$V_e = B_v e = A_v(I - A)^{-1}e \tag{2}$$

$$V_f = B_v e = A_v(I - A)^{-1}f \tag{3}$$

Correspondingly, we can estimate the embodied CO₂ emissions in exports or final demands as follows:

$$C_e = c(I - A)^{-1}e \tag{4}$$

$$C_f = c(I - A)^{-1}f \quad (5)$$

Where, c is a row vector of row vector of CO₂ emissions coefficients representing CO₂ emissions per unit of economic output by sector of different sized firms.

2.2 Data source

For the convenience of the study and data availability, the original China's 2012 non-competitive input-output table was aggregated into 29 sectors, including 24 industrial sectors. The firm size division criteria of different sectors is based on the '*Statistical Dividing Method for Large, Small and Medium-sized Micro-firms (2011)*' published by the National Bureau of Statistics of China(NBS). The division criteria of industrial sector are shown in Table 2, and the firms of industrial sector are divided into three firm size types: large, medium, and small according to the indexes of employment and revenue. Based on the firm-level data and the economic census data, we can obtain the initial estimated value of LMS model including intermediate flow matrix, the import matrix, the total matrix, and the final demand matrix, respectively. After obtained all the initial data of LMS model, then RAS method was applied to get the new extended input-output table. Specifically, because of the firm-level data limitation of the other sectors, some assumptions of different sized firms are applied. So in this study, we only analyse the results of the industrial sectors considering the data accuracy.

Table 2 The division criteria of firm size

Sector	Index	Unit	Large	Medium	Small
Industrial	Employment(X)	Person	X≥1000	300≤X<1000	X<300
	Revenue (Y)	10 ⁴ RMB	Y≥40000	2000≤Y<40000	Y<2000

3. Results

This section illustrates the different roles of large, medium, and small-sized firms in China's economy and environment. First, it shows the distribution of the value added, output, export, and CO₂ emissions generated by different sized firms in the industrial sector. Then, we present results of several input-output analysis obtained by using the newly constructed extended input output table.

3.1 Firm-level contribution to China's economy and CO₂ emisison

Using the estimation method described in the previous section, we estimated an extended noncompetitive I/O table distinguishing the firm size types. Table 3 shows the output, value added, export and CO₂ emissions share of different sized firms in the industrial sector.

Table 3 Contribution to China's economy and CO₂ emissions of industrial sector

	Value-added ratio	Output share	Value added share	Export share	CO ₂ emissions share	The share of export to final demand
Large firm	20.02%	37.11%	33.35%	48.38%	41.17%	50.20%
Medium firm	24.57%	25.98%	28.66%	28.70%	27.08%	43.98%
Small firm	22.93%	36.91%	37.99%	22.92%	31.75%	31.60%

Overall, the total output and value-added of small and medium-sized firms accounted for more than 60% in the industrial sector, sufficiently demonstrating that small and medium-sized firms play a crucial role in promoting China's economic development. It is noteworthy that the proportion of output and value-added of small-sized firms reached 36.91% and 28.66% respectively, which was similar to the proportion of large firms and playing an important role in the national economy. As for the export, large-sized firms accounted for half of China industrial sector's total exports (48.38%), which is enough to show that large-sized firms have certain advantages over small and medium-sized firms in export that is closely related to the characteristics of large-sized firms. In 2012, the top two sectors of export scale were telecommunication equipment, computer and other electronic equipment manufacturing and electrical machinery and equipment manufacturing, which accounted for 26.99% and 9.75% of the total export of industrial sector, respectively. These two sectors are typical technology-intensive industries, but small and medium-sized firms are relatively lack of technological advantages in the two sector, which leading to the export of large-sized firms in these two sectors occupied a very obvious position. In addition, in terms of the CO₂ emissions, large-sized firms contributed about 41.17% CO₂ emissions of the industrial sector's total emissions, while the share of the medium firms was the smallest, only occupying 27.08%. The proportion of export in the final demand of large-sized firms and medium-sized firms was 50.20% and 43.98% respectively, while the share of small firms was only 31.60%. This shows that the final products of small firms are mainly used for domestic final consumption due to insufficient comparative advantage and other factors, and the proportion of export is significantly lower than that of large and medium firms.

It is worth noting that in the year of 2012 that the average value-added ratio of China's industrial industry was 22.28%. Specifically, the value-added ratio of medium-sized firms was 24.57%, while the value-added ratio of large-sized firms was at the lowest level of 20.02%. This illustrates

that the value-added created by per unit input of medium-sized firms is higher than that of large-sized firms.

3.2 Roles of different sized firms in the China's export

In this section, we use the extended Chinese input-output table, to explore the value-added exports as well as export-embodied carbon emission of different sized firms in China. Table 4 shows the main indicators of export calculated by this *LMS* model.

Table 4 Major export indicators for the industrial sector

	Value-added export share	Export embodied carbon emissions share	Value-added per export	Per export embodied carbon emissions (t CO ₂ /RMB)	Carbon emissions of per value-added export (t CO ₂ /10 ⁴ RMB)
Large firm	44.05%	45.62%	0.65	1.82	2.82
Medium firm	31.37%	29.57%	0.78	1.99	2.56
Small firm	24.58%	24.80%	0.76	2.09	2.74

We estimate that China's total value-added that generated by export in 2012 were 10338 billion RMB, accounting for 19.26% of China's total value-added. Meanwhile, the total export-embodied carbon emissions were 2487 MtCO₂, which occupied 21.17% of the China's total CO₂ emissions. Specific to industrial sectors, the corresponding values are 40.01% and 40.83%, respectively. Of these value-added and embodied carbon emissions that generated by export in industrial sectors, large firms contributed 44.05% and 45.62% respectively, playing a dominant position.

It can be concluded from Table 4 that medium-sized firms had the highest value-added of 7,800 RMB that generated by 10,000 RMB export, which indicates that in China's foreign trade, the value-added generated by per export of medium-sized firms is higher than that of large-sized firms and small-sized firms in the industrial sector. It is worth noting that the value-added per export of large-sized firms is the lowest, which deserves close attention in China's foreign trade in the future. For the embodied carbon emissions generated by per export, small-sized firms had the highest per export embodied carbon emissions, while large firms had the lowest value. This illustrates that the carbon emissions generated by the export production of small-sized firms are the highest, which is closely related to the technological level and production mode of small-sized firms. It is worth noting that for carbon emissions of per value-added export, medium-sized firms produced the lowest carbon emissions per 10,000 RMB value-added export, which is 2.56 tons. Carbon emissions of per value-added export can effectively measure the environmental costs

required to create unit economic benefits in export trade. This result shows that the environmental cost of medium-sized firms in creating the same value-added is the lowest.

4. Conclusion

This research constructed an extended China's input-output table distinguishing the firm heterogeneity by size to measure the specific role of the different firms by size in China's economy and environment. The results show that small and medium-sized firms play a crucial role in promoting China's economic development, of which the output and value-added accounted for more than 60% in the industrial sector. As for the exports, large firms composed about half of China's total exports, which is obviously to show that large firms have certain advantages over small and medium firms of exports. In terms of CO₂ emissions, large firms contributed about 41.17% of the industrial total emissions, acting as a dominant factor to the China's carbon emission in 2012.

The value-added ratio of medium-sized firms reached 24.57%, which was higher than large-sized firms and small-sized firms; and the value-added per export of medium-sized firms was also higher than the other firm size types, reaching 0.78. This investigates that in China's foreign trade, the value-added created by per unit input of medium-sized firms in industrial sector is higher than large-sized firms and small-sized firms. In particular, the value-added ratio of large enterprises was the lowest among the three firm types, which is closely related to the excessive intervention of local governments aiming to stimulate the economic development. Therefore, it is necessary for the government to carry out more reasonable taxation and investment policies, so as to propel the financial market be more inclined to small and medium-sized firms to help them solve the difficulty of financing. For the per export embodied carbon emissions, small-sized firms had the highest per export embodied carbon emissions, while large-sized firms had the lowest value-added, which fully indicates that the export products structure of small-sized firms is facing with optimizing and upgrading in the future. However, small firms generally have difficulty in financing which is so hard for them to carry out technological innovation or introduce environmental protection technologies by themselves. Therefore, under this circumstance, it is significant for the government to increase its support and issue some policies in favor of the development of small-sized firms.

References

1. Ahmad, N., & Wyckoff, A. (2003). Carbon dioxide emissions embodied in international trade of goods.

2. Deng, G., & Xu, Y. (2017). Accounting and structure decomposition analysis of embodied carbon trade: A global perspective. *Energy*, *137*, 140-151.
3. Dietzenbacher, E., Pei, J., & Yang, C. (2012). Trade, production fragmentation, and China's carbon dioxide emissions. *Journal of Environmental Economics and Management*, *64*(1), 88-101.
4. Huang, L., & Zhao, X. (2018). Impact of financial development on trade-embodied carbon dioxide emissions: Evidence from 30 provinces in China. *Journal of Cleaner Production*, *198*, 721-736.
5. Jiang, X., Chen, Q., Guan, D., Zhu, K., & Yang, C. (2016). Revisiting the global net carbon dioxide emission transfers by international trade: the impact of trade heterogeneity of China. *Journal of Industrial Ecology*, *20*(3), 506-514.
6. Liu Z, Davis S J, Feng K, et al. Targeted opportunities to address the climate–trade dilemma in China[J]. *Nature Climate Change*, 2016, 6(2): 201.
7. Ma, H., Wang, Z., & Zhu, K. (2015). Domestic content in China's exports and its distribution by firm ownership. *Journal of Comparative Economics*, *43*(1), 3-18.
8. Peters, G. P., & Hertwich, E. G. (2006). Pollution embodied in trade: The Norwegian case. *Global Environmental Change*, *16*(4), 379-387.
9. Peters, G. P., Minx, J. C., Weber, C. L., & Edenhofer, O. (2011). Growth in emission transfers via international trade from 1990 to 2008. *Proceedings of the national academy of sciences*, *108* (21): 8903-8908.
10. Su, B., & Ang, B. W. (2014). Input–output analysis of CO₂ emissions embodied in trade: a multi-region model for China. *Applied Energy*, *114*, 377-384.
11. Weber, C. L., Peters, G. P., Guan, D., & Hubacek, K. (2008). The contribution of Chinese exports to climate change. *Energy Policy*, *36*(9), 3572-3577.
12. Wyckoff, A. W., & Roop, J. M. (1994). The embodiment of carbon in imports of manufactured products: implications for international agreements on greenhouse gas emissions. *Energy policy*, *22*(3), 187-194.
13. Yang, C., Dietzenbacher, E., Pei, J., Chen, X., Zhu, K., & Tang, Z. (2015). Processing trade biases the measurement of vertical specialization in China. *Economic Systems Research*, *27*(1), 60-76. Yunfeng, Y., & Laike, Y. (2010). China's foreign trade and climate change: A case study of CO₂ emissions. *Energy policy*, *38*(1), 350-356.
14. Zhao, Y., Wang, S., Zhang, Z., Liu, Y., & Ahmad, A. (2016). Driving factors of carbon emissions embodied in China–US trade: a structural decomposition analysis. *Journal of Cleaner Production*, *131*, 678-689.



Extended input-output model for demographic change – Preliminary application to the Chinese urbanisation



Nobuhiro Okamoto

Daito Bunka University, 560 Iwadono, Higashimatsuyama, Saitama, 355-8501, Japan

Abstract

China carries out “Urbanisation” as an economic policy which intends to concentrate people in the urban area and boost the whole economic growth on the basis of “Economy of Agglomeration” struggling with the pressure of “middle-income trap” or “new normal.” The research question here is about how the labour migration from rural areas to urban areas has an economic and industrial impact on the Chinese economy, and whether or not the geographical change between space is truly beneficial for Chinese economic growth in the near future. To answer this question, this research develops the extend input-output model based on the previous research such as Batey (2018) and their other research, which focus on incorporating labour account with Input-Output model. In Batey and others’ original model, the Input-Output model has been developed into the economic model with a household which takes account of immigrants from other regions, people who are out of work, and ordinary labour force. This study develops this extended Input-Output model for demographic change, in particular, change of population movements from villages to cities in China since the urbanisation process is seen as the continuous concentration of people in the certain areas, especially, cities. The study will illustrate the preliminary results in the case of China by using this model. Furthermore, the paper will discuss the possibilities of a wide range of application of the Input-Output table in terms of demography.

Keywords

Demography; Urbanisation; Migration; Input-output analysis

1. Introduction

For the sustainable economic development of any region, its demography is important. China, which is large populous country, has recently proceeded the urbanisation and townisation as an economic policy since in order to overcome the so-called ‘middle-income gap’, increasing the productivity of cities, which is considered the decisive engine of the economic growth, by concentrating the people in cities. The swelling population in urban areas might have an important effect on the regional economy. However, the previous researches have not clarified the interrelational process between migration and economy such as how the migrants in cities affect the economy.

This study aims to make clear the above-mentioned interrelation by using the input-output model, which focus on the interdependency of economic factors. The conventional input-output analysis has focused on the changes in output brought about by industrial activities; households have been incorporated into the model wherein they are treated as an industry within the framework. Batey and his co-researchers, mainly Madden, have positively contributed to the field of the extended input-output model involving population, especially labour accounts (Batey and Madden 1981; Batey 1985; Batey and Weeks 1987; Batey, Madden, and Weeks 1987; Batey and Madden 1988; Batey and Weeks 1989; Batey and Madden 1999a, 1999b). The development and potential of this model were discussed by Batey and Rose (1990); more recently, the model has been reviewed in the context of declining regional economies (Batey 2018).

The paper consists of the following parts. First, it starts with a description of extended models of households to its derivative model of demography—which Batey and others focused on in their literature. Second, we developed the input-output model for urbanisation, applying the ‘Batey–Madden model’, and further discussion on its multipliers. Then, the empirical analysis of the Chinese urbanisation process is conducted. Finally, it reflects on the results and conclusions.

2. Methodology

The urbanisation is defined as the process of the movement from people in rural areas to urban areas. Farmers are mainly devoting themselves to agriculture production, migrate to cities to find better jobs and seek for their better life. City dwellers transformed from farmers become an important labour force in factories and offices, at the same time, they enjoy the modern consumption life.

Dividing households into one in cities and the other in villages, the input-output model for urbanisation is constructed as the application of rudimentary Batey-Madden Model.

$$\begin{bmatrix} I - A & -\dot{h}_c^u & -\dot{h}_c^r \\ -l & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_I \\ u \\ r \end{bmatrix} = \begin{bmatrix} d_I \\ u \\ p \end{bmatrix}$$

where

$I - A$: Leontief matrix

\dot{h}_c^u : a column vector of consumption coefficients, expressed as consumption per household, for an urban residence

\dot{h}_c^r : a column vector of consumption coefficients, expressed as consumption per household, for rural residence

l : a row vector of urban employment-production (urban employment/gross output ratios) functions by industrial sector

u : a scalar, the number of urban workers
 r : a scalar, the number of rural workers
 p : a scalar, the level of labour supply

If the matrix is partitioned with the economic and demographic activity, then it can be converted to a simple form of the equation as follows:

$$\begin{bmatrix} I - A & -H_c \\ -H_l & D \end{bmatrix} \begin{bmatrix} x_l \\ x_d \end{bmatrix} = \begin{bmatrix} d_l \\ d_d \end{bmatrix}$$

where $H_c = [\dot{h}_c^u \ \dot{h}_c^r]$, $H_l = \begin{bmatrix} l \\ 0 \end{bmatrix}$, $D = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. In addition, $x_d = \begin{bmatrix} u \\ r \end{bmatrix}$ is defined as the number of urban and rural workers, $d_d = \begin{bmatrix} 0 \\ p \end{bmatrix}$ is the number of commuting workers from rural areas¹, which is assumed to be an imbalance and therefore set at zero, and the number of active economic population or labour supply of the country. The equation can then be rewritten as:

$$\begin{bmatrix} x_l \\ x_d \end{bmatrix} = \begin{bmatrix} I - A & -H_c \\ -H_l & D \end{bmatrix}^{-1} \begin{bmatrix} d_l \\ d_d \end{bmatrix} = \begin{bmatrix} L^{11} & L^{12} \\ L^{21} & L^{22} \end{bmatrix} \begin{bmatrix} d_l \\ d_d \end{bmatrix} = \begin{bmatrix} B(I + H_c L^{22} H_l B) & B H_c L^{22} \\ L^{22} H_l B & L^{22} \end{bmatrix} \begin{bmatrix} d_l \\ d_d \end{bmatrix}$$

Through the detailed analysis of each element of this inverse matrix, we define them as in Table 1 and 2..

Table 1 Image of inverse matrix of model for urbanisation

	Industry	Urban and rural household
Industry	$B(I + H_c L^{22} H_l B)$ Leontief multiplier	$B H_c L^{22}$ Induced per capita consumption
Urban and rural employment	$L^{22} H_l B$ Induced urban employment and reduced rural employment	L^{22} urban and rural labour allocation multiplier, simply urbanization multiplier

Table 2 Interpretation of urban and rural labour allocation multiplier (L^{22})

	Urban household	Rural household
Urban employment	Urban employment multiplier	The probability of urban employment
Rural employment	Rural employment multiplier	The probability of rural employment

The rather complex point to this inverse matrix is how to interpret the submatrix of demographic change between rural and urban areas, defined as ‘urban and rural labour allocation multiplier’, more simply, ‘urbanisation

¹ The model assumes that there are no commuting workers from rural areas to urban areas for their job. It means that rural workers work only at villages as long as they live in the countryside.

multiplier', indicating that the increased total labour force in the country would be reallocated into cities and villages by a certain probability, and have an impact on urban employment. That involves important information about urbanisation in terms of employment and movement. Table 2 shows this definition of L^{22} .

3. Result

We are going to test the model of how it works using the relevant data. First, the data used in our analysis is the 2015 input-output table for China, which is the latest but the updated table from the 2012 benchmark table, and data related to the labour account for the latest ten years, namely from 2008 through 2017. First, the overall labour accounts are shown in figure 1 and 2

Figure 1 Economically active population

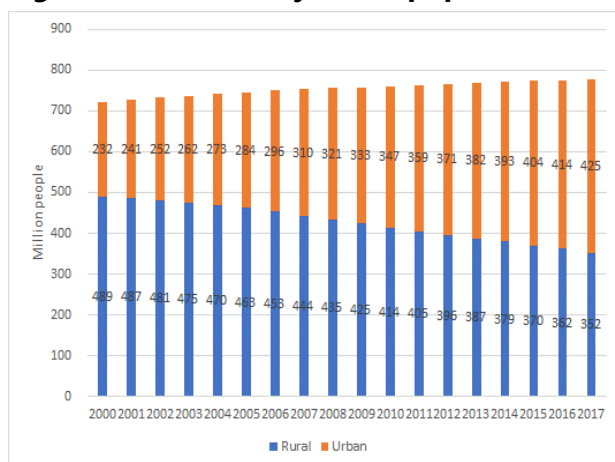
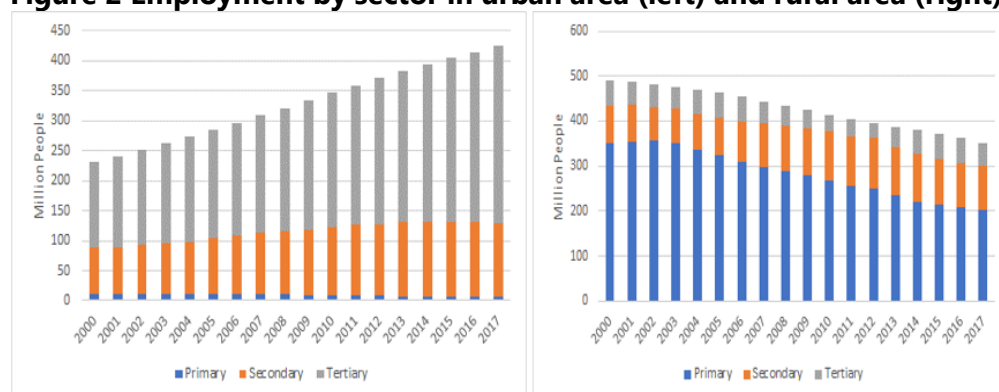


Figure 1 shows the labour force allocation of where they work (without the unemployment population). Total population is slightly on the increase, and the number of urban workers has been increasing significantly. The rural workers were smaller than the urban workers in 2015, meaning that the

majority of the labour force exists in cities after that.

Figure 2 indicates the number of the employee by sector in cities; it also describes which sector has absorbed the population from rural areas. It clearly reveals that the tertiary sector plays an important role in employment in urban areas.

Figure 2 Employment by sector in urban area (left) and rural area (right)



On the contrary, the primary sector in the rural area is the main source of labour supply to urban industries.

Next, Table 3 shows the outcome implemented by the model for urbanisation. Each part of the table corresponds with the inverse matrix shown in Table 1 and 2. The latest input-output data for 2015² was used in combination with the labour account of the same year.

Table 3 Result of the model execution for 2015

	Primary	Secondary	Tertiary	Urban Household	Rural Household
Primary	1.215	0.208	0.106	6429	4275
Secondary	0.939	3.088	1.200	67089	31359
Tertiary	0.406	0.959	2.000	54662	22237
Total	2.560	4.255	3.307	128180	57871
Urban employment	3.18	6.87	8.94	1.280	0.119
Rural employment	-3.18	-6.87	-8.94	-1.280	0.881

Note: A unit of household consumption is yuan and employment in person.

First, since the households are incorporated into the model, each cell of the multiplier is larger than the conventional Leontief inverse, and the total of the column which is seen as the total backward linkage effect indicates 2.560 for primary, 4.255 for secondary, and 3.307 for tertiary industry respectively.

Next, the upper-right part of the table shows the output generated by a unit increase in consumption induced by demographic change. The increase in economically active population will induce the total output in the industry to meet their increased consumption. We can see the total induced effects which are the sum of the column, 128,180 yuan for urban areas and 57,871 yuan for rural areas. This shows that urban areas consumption raised by the

² The data is provided as the format of so called import-competitive type. Each transaction includes imported goods and services. This elimination work will be done in future research.

increase in the number of the population play a crucial factor in the growth of industrial production.

Third, we find the employment induced by final demand in the lower-left of the table. This model considers only employment in urban areas assumed to be induced by the final demand of sectors. The number of employment in urban areas is increased, 3.18 people in primary, 6.87 people for secondary, and 8.94 people for the tertiary sector respectively. And also, the same number of people are decreased in rural areas. Thus, approximately 19 people are migrated from villages to cities by a unit increase in final demand.

Finally, the lower-right of the table illustrates the information about the demographic change, in particular, the process of urbanisation in China. That is defined here as 'urban and rural labour allocation multiplier', simply, 'urbanization multiplier.' Urban employment multiplier is 1.280 whereas the same amount of negative figure is the rural employment multiplier. From the viewpoint of rural households, their probability of taking a job in urban areas is 0.119 (11.9%), and the probability of remaining in rural areas is 0.881 (88.1%).

In order to understand how the model works deeply, the changes in each element of the multiplier are investigated by the changes of labour accounts from 2008 to 2017 with remaining constant of input-output data for 2015. That is, we can see the changes in impacts on the whole economy and demography induced by the changes in labour allocation between urban and rural areas, assuming that the economic structure is unchanged.

The results are shown in Figure 3 and 4. As seen in Figure 1 and 2, there was a tendency of labour migration from rural areas to urban areas with a slight increase in total labour supply in China from 2008 to 2017. Figure 3 indicates that with the constant of input-output structure, the Leontief multiplier or total sum of it called backward linkages is marginally decreasing even though the secondary sector has the strongest backward linkage among sectors. The movement of labour from the urban sector to rural sector might have a force to reduce the backward linkages in the whole country.

Reflecting this change in backward linkages, the output induced by total household consumption is also decreasing, in particular, urban household, but the consumption of rural areas remain relatively stable (The right of Figure 3). We still need further analysis of the reason why they are decreasing as the urban population increases. The result seems to be opposite to our intuition.

Figure 3 Leontief multiplier (Backward linkage) and Output induced by consumption

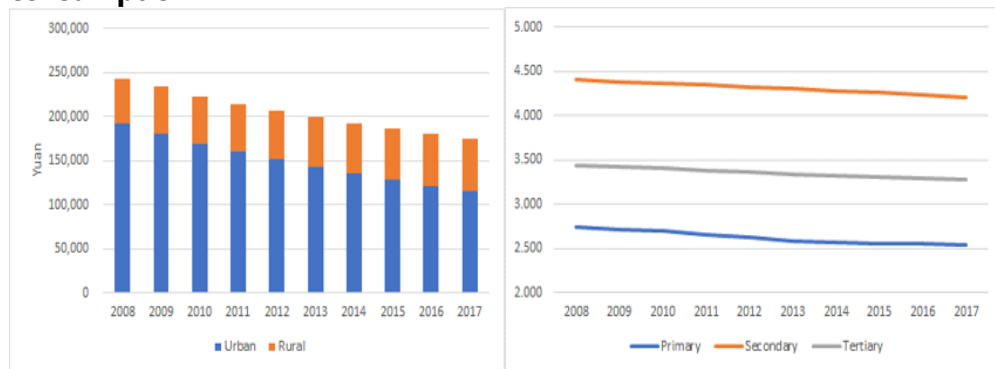


Figure 4 Induced employment and related multiplier and probability of labour in urban areas

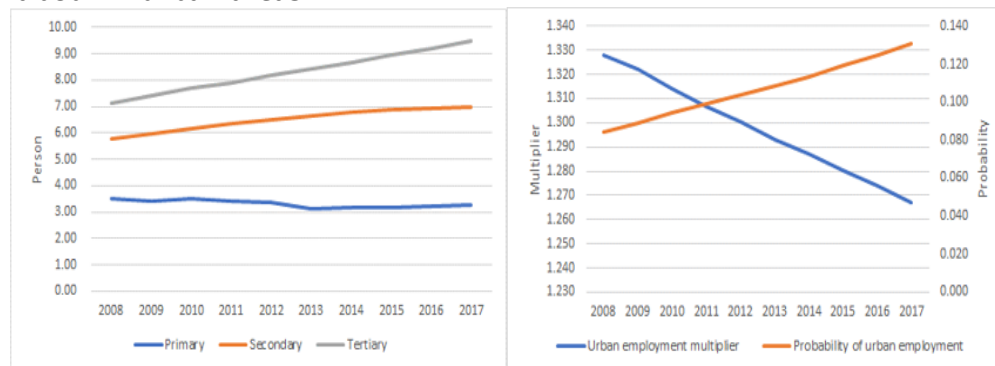


Figure 4 shows the changes in the lower side of the model inverse matrix, specifically, urban employment (or movement to urban sectors) in the left, and urbanisation multiplier in the right. There is a constant increase in the number of people absorbed in the urban tertiary sector, whilst employment in primary sectors remain constant. Nevertheless, urban employment multiplier is declining whereas we can see an upward movement in the probability of urban employment.

4. Discussion and Conclusion

This paper has proposed the extended input-output model with regard to the urbanisation, which is considered as the demographic change, i.e., the population movement from the countryside to the cities, by applying the Batey-Madden model. It has also analysed the model structure and the new multiplier called 'urbanisation multiplier' has been provided after its multiplier have been thoroughly studied.

The framework of our extend input-output model for urbanisation provides a useful basis for studying the relationship between urbanisation and

economic change. An important aspect of this change is the increase or decrease in the number associated with urban workers and rural workers, together with national labour supply. In fact, The model has been used to analyse the urbanisation process, labour migration from villages to cities, in China. The findings provided us with an insightful and new aspect of understanding the urbanisation, which has been carried out as an economic policy in China.

Although figures calculated by the model still need to be thoroughly investigated, this model has a strong potential for further revelations and hence is apt for more in-depth analysis. We could elaborate the valuables in the model by reflecting the current situations in China, for instance, the existence of rural migrant workers in cities who are not treated as inhabitants in cities, as well as accumulating the empirical studies in wider aspects.

References

1. Batey P. W. J. (1985) "Input-Output Models for Regional Demographic-Economic Analysis: Some Structural Comparisons," *Environment & Planning A*, Vol. 17 (1), pp. 73-99.
2. Batey, P. W. J. (2018) "What Can Demographic-Economic Modelling Tell Us About the Consequences of Regional Decline?" *International Regional Science Review*, Vol. 41 (2), pp. 256-281.
3. Batey, P. W. J. and A. Z. Rose (1990) "Extended Input-Output Models: Progress and Potential," *International Regional Science Review*, Vol. 13 (1-2), pp. 27-49.
4. Batey, P. W. J. and M. Madden (1981) "Demographic-Economic Forecasting Within an Activity-Commodity Framework: Some Theoretical Considerations and Empirical Results," *Environment and Planning A*, Vol. 13 (9), pp. 1067-1083.
5. Batey, P. W. J. and M. Madden (1988) "The Treatment of Migration in an Extended Input-Output Modelling Framework," *Ricerche Economiche*, Vol. 42 (2), pp. 344-66.
6. Batey, P. W. J. and M. Madden (1999a) "Interrelational Employment Multipliers in an Extended Input-Output Modeling Framework," in *Understanding and Interpreting Economic Structure*, edited by Hewings, G. J. D., Sonis, M., M. Madden, and Y. Kimura, Heidelberg and New York: Springer, pp. 73-89.
7. Batey, P. W. J. and M. Madden (1999b) "The Employment Impact of Demographic Change: A Regional Analysis," *Papers in Regional Science*, Vol. 78 (1), pp. 69-87.
8. Batey, P. W. J., M. Madden, and M. J. Weeks (1987) "Household Income and Expenditure in Extended Input-Output Models: A Comparative

- Theoretical and Empirical Analysis," *Journal of Regional Science*, Vol. 27 (3), pp. 341-356.
9. Batey, P. W. J. and M. J. Weeks (1987) "An Extended Input-Output Model Incorporating Employed, Unemployed, and In-Migrant Households," *Papers in Regional Science*, Vol. 62 (1), pp. 93-115.
 10. Batey, P. W. J. and M. J. Weeks (1989) "The Effects of Household Disaggregation in Extended Input-Output Models," in *Frontiers of Input-Output Analysis*, edited by Miller, R. E., K. R. Polenske, and A. Z. Rose, New York and Oxford: Oxford University Press, pp. 119-33.



Using input-output tables to study trade and international production sharing arrangements



Joseph Mariasingham
Asian Development Bank

Abstract

Economic globalization is increasingly being characterized by fragmented production processes that are distributed internationally. As enterprises seek to capitalize on factor cost differentials and the lowering of barriers to trade and investment, cross-border transactions in intermediate products have come to dominate international trade. Such internationalization of the production process, however, poses a number of critical definitional and measurement challenges and issues, as conventional approaches to characterizing trade flows and presenting trade statistics have shown to be inadequate in capturing the essential characteristics of international production sharing. To fill this important analytical gap, a number of multilateral institutions such as the Asian Development Bank (ADB) have taken the initiative to produce input-output table based statistics and quantitative analyses to complement basic trade statistics. This paper discusses such a framework for producing statistics on international production sharing.

Keywords

Economic globalization; input-output analysis; international production sharing.

1. Introduction

The principal sources of trade data are customs records. Goods that cross territorial boundaries are recorded primarily as exports, re-exports, imports, or re-imports at full value. Valuation methods are generally based on the purchase price or cost of production. Trades in services are discerned through an economy's balance of payment accounts maintained by its central bank. Deeper analysis of relevant data gathered through enterprise and trade surveys could provide additional insights on origin and destination as well as components of the traded commodities. Trade data in themselves do not provide information on the effects of cross-border transactions on the economy. The underlying issue is that trade data are recorded and presented in gross value terms without any attempt to delineate the local and foreign contents in the traded commodity or the contributions of different industrial sectors to its production. Commodities are produced either completely locally or by incorporating at least one non-local (imported) component (good or

service). The territorial apportionment of the benefits of productive activity is primarily determined by the origin of the components. In this regard, statistics on a territory's actual contribution to the production of a commodity is essential for economic analysis and policy-making. Such granularity in data becomes even more significant as countries seek economic growth by expanding the market for their products beyond their territorial boundaries. Extending the decomposition by origin to product and sector levels will further enhance the analytical utility of the information. Further, standard trade statistics do not readily facilitate the measurement of an economy's involvement in globalized production processes. For example, an economy whose sole export is a basic low-valued, yet key, component of a commodity assembled primarily in and shipped globally from another economy would not be discerned as highly integrated into the global market through traditional measures. Nonetheless, the value of the work done in the economy producing the key component (its "value added") is intrinsic in the commodity. The criticality of the economy, and its sector producing the key component, in the production process of the commodity is concealed in standard measures. Likewise, the contributions of other local sectors that support the exporting sector are also not apparent. Traditional approaches do not support a mechanism for tracing the path of a value added from its initial creation to final consumption. Only the economic input–output analysis framework provides such a facility.

2. Studying production and trade through an IO analysis framework

Figure 3.1 depicts an elementary open economy in IOT form at a given point in time. There are three principal matrices: intermediate use, final use, and value added. The total output, or supply, by industrial sector is provided in the row vector and the total demand by industrial sector is given in the column vector, which are also the row and column sums, respectively, of the system of matrices. The economy has three industrial sectors ($i, j = 1, 2, 3$), two final use domestic sectors (e.g., households), and the rest of the world (ROW). The intermediate use matrix records bilateral and bisectoral transactions in intermediates, which are commodities used in the production of other commodities. The value added matrix details the shares of labor (compensation), capital (interest and depreciation), entrepreneurial effort (operating surplus or profit), and government (production and commodity taxes and subsidies) in a given sector's output. The sectors produce differentiable commodities valued X_j . Assume that sector 1 of the domestic economy imports an intermediate commodity valued M_1 , transforms or enhances it using domestic labor valued V_1 , and produces output valued X_1 . Sector 2 uses sector 1's output as input in its production process, employing labor valued V_2 to produce output valued X_2 , which, in turn, becomes the input

in the production process of sector 3. The chain of production and bisectoral trade in intermediates continues until the product of sector 3 valued X_3 is either exported (E_3) to the ROW or consumed by the domestic final use sector (F_3) and is thereby no longer used in the economy's domestic production processes.

Figure 3.1: Input Output Transactions Table

		Industrial Sectors as Consumers (j)			Rest of the World (ROW)			Total demand
		1	2	3	Domestic	Imports	Exports	
Industrial Sectors as Producers (i)	Intermediate use							
	1	M_1	X_1					X_1
	2			X_2				X_2
	3							X_3
	Labor Capital and entrepreneurship	Value Added						
	Government							
	Total value added	V_1	V_2	V_3				
	Total output	X_1	X_2	X_3				
	Final use				F_3	$-M_1$	E_3	

A salient feature of the IOT is that it provides the mechanism for detailing the direct and indirect linkages between production and trade in a systematic and mathematical manner. Since every sector-specific production process (resulting in the production of $X_j >= 0$) can be represented as the linear combination of the contributions of all industrial sectors ($z_{ij} >= 0$) in the sector i -sector j space ($i, j = 1, \dots, n$), the intermediate use matrix (Z) and the associated matrix of technical coefficients (A) are square. Further, in the matrix representation of a realistic economy, no column sum in A is greater than 1, and at least one column sum is less than 1 (implying non-negative value added in every sector). Given these characteristics of the technical coefficient matrix A , a powerful economic analytical tool known as Leontief inverse can be derived from it. Formulaically, it is expressed as

$$L = (I - A)^{-1}$$

where I is the identity matrix whose dimensions are same as that of A . L is also known as the total requirements matrix, whereas the matrix of technical coefficients, A , is also referred to as the direct requirements matrix. The matrix of total output X (accounting for all direct and indirect effects) required to support final demand F is given by

$$X^r = (I - A^r)^{-1} F^r$$

where r refers to the economy being analyzed. A^r is the technical coefficient matrix of transactions within r .

A defining contribution of the input–output system—from the tables to the Leontief inverse—to economic analysis is the quantified mapping of the continuum of linkages and relationships between production and trade, making it the ideal framework for studying the globalized production environment. Figure 3.2 situates an economy in an international context by incorporating the input–output details of the trading partners in the system of matrices, resulting in a simple international or interregional IOT with two economic territories. In this articulation, the intermediate and final use matrices are decomposed as use of domestically produced commodities and use of imports. Given that the imports of an economy are the exports of its trading partners and all commodities have to be produced, and consumed, in the world characterized by the two economies, Figure 3.2 describes a complete global system of production, trade, and consumption.

Figure 3.2: Numerical Example of an International Input Output Transactions Table

		The Economy			Rest of the World (ROW)			Final use			
		Industrial Sectors as Consumers (I)			Industrial Sectors as Consumers (I)						
The Economy	Industrial Sectors as Producers (I)	Intermediate use						Domestic		Exports	Total demand
		1	2	3	1	2	3				
	1		50								50
	2			80							80
	3					5		75	20		100
ROW	1	40									40
	2				20			5	5		30
	3										0
Value added											
	Labor	5	15	10	10	15					
	Capital and entrepreneurship	5	15	10	10	10					
	Government										
	Total value added	10	30	20	20	25	0				
Total output											
		50	80	100	40	30	0				

The interpretation of the matrices is the same as discussed earlier, but now the input requirements of a production process are also presented in another dimension: territorial origin of inputs. The resulting total requirements matrix details, maps, and quantifies the global (direct and indirect) effects of a final consumption decision, regardless of its origin, in the three-dimensional, geography–sector–sector space. By relocating, for example, the production of the economy’s sector 1 intermediate input from itself to sector 1 of the ROW and by enabling the ROW’s sector 2 to use the economy’s sector 3 output, Figure 3.2 creates a new set of direct and indirect interregional and intraregional productive dependencies; we now have a simple globally shared production process or global production chain.

3. Value added approach to analysing trade data

The cost of production or purchase of commodities forms the basis for the conventional presentation and analysis of the statistics on trade. However, as discussed earlier, such an approach has limited analytical utility in an economic environment characterized by highly fragmented production processes distributed globally. In particular, the actual contribution of an economy, or a given sector of an economy, to the production of a commodity is not readily discernible. Hence, a more insightful and analytical illustration of the data is needed to fully understand the state and dynamics of modern day trade, and its correlation to international production-sharing arrangements. The last section showed that the input–output framework provides the right setting for a quantitative exploration of trade and trade patterns and linkages. This section delves deeper into the framework to extract a mechanism for decomposing the total or gross value of a commodity according to where its componential values are created (added). Figure 4.1 elaborates on the IOT example provided in Figure 3.1 by denoting the complete set of transactions possible within the economic framework provided therein. The sector specific production technologies depicted in industrial sector columns can be represented by a system of equations as follows:

$$p_{11}X_1 + p_{21}X_2 + p_{31}X_3 + V_1 = X_1$$

$$p_{21}X_1 + p_{22}X_2 + p_{23}X_3 + V_2 = X_2$$

$$p_{31}X_1 + p_{32}X_2 + p_{33}X_3 + V_3 = X_3$$

Where $0 \leq p_{ij} < 1$ and $V_j > 0$.

The solution to the system of equations for X_j shows that the output of sector j , and hence its imports and exports, can be completely decomposed as the value added terms V_j of all the industrial sectors.

Figure 4.1: Input Output Transaction Table

		Industrial Sectors as Consumers (j)			Rest of the World (ROW)			Total demand
Intermediate use		1	2	3	Domestic	Imports	Exports	
Industrial Sectors as Producers (i)	1	$p_{12}X_2$	$p_{13}X_3$	$p_{11}X_1$	F_1	$-M_1$	E_1	X_1
	2	$p_{22}X_2$	$p_{23}X_3$	$p_{21}X_1$	F_2	$-M_2$	E_2	X_2
	3	$p_{32}X_2$	$p_{33}X_3$	$p_{31}X_1$	F_3	$-M_3$	E_3	X_3
Value added								
Labor Capital and entrepreneurship								
Government								
Total value added		V_1	V_2	V_3				
Total output		X_1	X_2	X_3				

where $0 \leq p_{ij} < 1$

By decomposing output in value added terms, one can quantify the contribution of each sector, and that of the territory where it is located, in the

output of any given sector. The decomposition within the input–output analysis framework provides a facility to discern the length of the production chain, degree of the distribution of the production process globally (production sharing), and position of an economy or sector in the production sequence of a commodity. Crucially, by identifying and quantifying the contribution (value added) of each economy or sector in the production of a commodity, the value added decomposition permits the measurement of the benefits accruing to the sector or economy as a result of participating in the production, and trade, of the commodity. In terms of standard System of National Accounts concepts, the net benefit (or income or value) accruing to an economy or sector in order for it to be counted in as part of GDP is the value added. The value added approach can succinctly be encapsulated in the input–output framework tracing both the sector or economy contribution to the full set of production processes (forward linkages) and contributions of all the sectors or economies to the production process of a given sector (backward linkages). An abstract technical coefficient matrix and its Leontief inverse are presented in Figure 4.2. To recapitulate, each term in the Leontief inverse, or total requirements matrix, of the technical coefficient matrix of an economy shows how much of sector i 's output is needed to meet the economy's productive, direct and indirect, requirements to supply one unit value of the final demand, including exports, for the output of sector j . The column j thus gives the total requirements by the producing sector for all the intermediates needed to produce output X_j of sector j in order to meet an additional unit value of the final demand for the product of sector j . Row i shows the total amount of the output of sector i needed, directly and indirectly, by the economy to meet the final demand for the product of each sector j . Since each element in the matrix is given in terms of output of sector i , it can be converted into value added terms by multiplying it by the proportion of value added, V_i , embedded in the products as shown in Figure 4.3. However, as discussed above, the total value added, even at the sector level, translates into final use or final demand. Thus, the column sum of the value added embedded in each term of the total requirement matrix is equal to the final demand for a sector's output, which is unity by definition. Extending this mathematical formulation and multiplying the total requirement matrix of value added (VB) by the matrix of actual level of final demand for each sector's product Y results in a matrix (VBY) that provides a framework for decomposing the final demand of a sector's product into various, economy-sector-specific, value added components (Figure 4.4).

Figure 4.2: Direct and Total Requirements Matrices

		The Economy			ROW			
		Industrial Sectors as Consumers (I)			Industrial Sectors as Consumers (I)			
		1	2	3	1	2	3	
The Economy ROW	Direct Requirements Matrix, A	Intermediate use						
		1	a^{11}	a^{12}	a^{13}	a'^{11}	a'^{12}	a'^{13}
		2	a^{21}	a^{22}	a^{23}	a'^{21}	a'^{22}	a'^{23}
	3	a^{31}	a^{32}	a^{33}	a'^{31}	a'^{32}	a'^{33}	
	1	a^{11}	a^{12}	a^{13}	a'^{11}	a'^{12}	a'^{13}	
	2	a^{21}	a^{22}	a^{23}	a'^{21}	a'^{22}	a'^{23}	
3	a^{31}	a^{32}	a^{33}	a'^{31}	a'^{32}	a'^{33}		
		The Economy			ROW			
		Industrial Sectors as Consumers (I)			Industrial Sectors as Consumers (I)			
		1	2	3	1	2	3	
The Economy ROW	Total Requirements Matrix, B	Intermediate use						
		1	b^{11}	b^{12}	b^{13}	b'^{11}	b'^{12}	b'^{13}
		2	b^{21}	b^{22}	b^{23}	b'^{21}	b'^{22}	b'^{23}
	3	b^{31}	b^{32}	b^{33}	b'^{31}	b'^{32}	b'^{33}	
	1	b^{11}	b^{12}	b^{13}	b'^{11}	b'^{12}	b'^{13}	
	2	b^{21}	b^{22}	b^{23}	b'^{21}	b'^{22}	b'^{23}	
3	b^{31}	b^{32}	b^{33}	b'^{31}	b'^{32}	b'^{33}		

Figure 4.3: Total Value Added Coefficient Matrix

		\hat{V}	B			$\hat{V}B$			
The Economy ROW	X	1	v^1_1	b^{11}	b^{12}	b^{13}	$v^1_1 b^{11}$	$v^1_1 b^{12}$	$v^1_1 b^{13}$
		2	v^2_1	b^{21}	b^{22}	b^{23}	$v^2_1 b^{21}$	$v^2_1 b^{22}$	$v^2_1 b^{23}$
		3	v^3_1	b^{31}	b^{32}	b^{33}	$v^3_1 b^{31}$	$v^3_1 b^{32}$	$v^3_1 b^{33}$
		4	v^4_1	b^{41}	b^{42}	b^{43}	$v^4_1 b^{41}$	$v^4_1 b^{42}$	$v^4_1 b^{43}$
		5	v^5_1	b^{51}	b^{52}	b^{53}	$v^5_1 b^{51}$	$v^5_1 b^{52}$	$v^5_1 b^{53}$
		6	v^6_1	b^{61}	b^{62}	b^{63}	$v^6_1 b^{61}$	$v^6_1 b^{62}$	$v^6_1 b^{63}$

Figure 4.4: Value Added Decomposition of Final Demand

		$\hat{V}B$	\hat{Y}	$\hat{V}B\hat{Y}$		
The Economy ROW	X	1	y^1_1	$v^1_1 b^{11} y^1_1$	$v^1_1 b^{12} y^1_1$	$v^1_1 b^{13} y^1_1$
		2	y^2_1	$v^2_1 b^{21} y^2_1$	$v^2_1 b^{22} y^2_1$	$v^2_1 b^{23} y^2_1$
		3	y^3_1	$v^3_1 b^{31} y^3_1$	$v^3_1 b^{32} y^3_1$	$v^3_1 b^{33} y^3_1$
		4	y^4_1	$v^4_1 b^{41} y^4_1$	$v^4_1 b^{42} y^4_1$	$v^4_1 b^{43} y^4_1$
		5	y^5_1	$v^5_1 b^{51} y^5_1$	$v^5_1 b^{52} y^5_1$	$v^5_1 b^{53} y^5_1$
		6	y^6_1	$v^6_1 b^{61} y^6_1$	$v^6_1 b^{62} y^6_1$	$v^6_1 b^{63} y^6_1$

Extending the earlier analysis on the total requirements matrix to $\hat{V}B\hat{Y}$, it can be seen that the columns of the matrix show the value added of each economy-sector embedded in the final demand for a given sector's output, essentially showing how the portion of a sector's output designated as final demand y_i was produced with the value added from different economy-sectors. Thus, the columns detail, in terms of value added, the productive linkages between the final demand for a sector's output and the economy-sector specific contributions required to produce it. These productive dependencies of a given sector on all sectors upstream, including on itself, are termed backward linkages. It shows how a change in the demand for a sector's

product affects the output of sectors supplying intermediates to it—that is, a sector’s supply dependency.

The elements across the rows of the *VBY* matrix provide the amount of a sector’s value added in the final demand for any given economy-sector’s output, thus specifying the sector’s contributions to the productive processes of all economy-sectors, including to its own processes. These downstream productive linkages that detail how and by whom a sector’s products are being used are called forward linkages. The degree of criticality of an economy-sector’s products to the production processes of various sectors and economies can be discerned from the information provided across the rows. It shows how a sector’s output would be affected by changes in the final demand for other sectors’ output—that is, an economy-sector’s demand dependence. From an economy’s perspective, the non-diagonal blocks across the rows indicate its level of export dependence or export concentration. It also shows how regionally diversified an economy-sector’s export markets are. The impacts of all exports on any given sector can also be discerned through the information across the relevant row.

4. Discussion and Conclusion

Basically, in a *VBY* matrix, the value added terms across and along the rows and columns referring to a given sector show, respectively, how its output was produced and how it was used. Information discerned across the rows and along the columns show the length, distribution, and concentration of a given commodity’s production chain. The entire economic input–output system expressed in terms of transactions in value added, as detailed by the *VBY* matrix, facilitates the measurement, analysis, and evaluation of the sectoral and economy-wide impacts of economic decisions on production and consumption. Since a multi-sectoral and multi-economy global economic system and the intersectoral and inter-economy interconnectedness and dependencies are depicted comprehensively by the input–output framework, the economy-wide and sectoral transmission and diffusion of the economic effects of the decisions can be traced, mapped, and quantified by the *VBY* matrix. This feature of the input–output system makes it a very powerful economic analysis tool especially in studying the level of economic integration of regions and in tracking temporally and spatially the impact of major investment activities such as economic corridor development.

References

1. Miller, R.E. and P.D. Blair. 2009. Input-Output Analysis: Foundations and Extensions. New York: Cambridge University Press.
2. United Nations. 2009. System of National Accounts 2008. New York: United Nations.
3. Eurostat. 2008. Eurostat Manual of Supply, Use and Input-Output Tables. Eurostat Methodologies and Working Papers. Luxembourg.
4. Koopman, Robert, Zhi Wang and Shang-Jin Wei. 2012. Estimating Domestic Content in Exports When Processing Trade is Pervasive. *Journal of Development Economics*. 99. pp. 178–189.



The effects of macroeconomic variables on future credit ratings



Gan Chew Peng¹; Pooi Ah Hin¹; Ng Kok Haur²

¹ School of Mathematical Sciences, Sunway University, Malaysia.

² Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia.

Abstract

The method based on the multivariate power-normal distribution is used to analyse the data on credit ratings and macroeconomic variables. Initially the macroeconomic variable is incorporated, one at a time, into the non-Markovian model which attempts to predict the credit rating in the next quarter when the ratings in the present and previous quarters are given. The effects of macroeconomic variables on future credit rating are found to agree basically with the corresponding results reported so far in the literature. We next summarize the effects of the set of macroeconomic variables by a small number of latent factors, and include the latent factors into the non-Markovian model. It is found that a small number of latent factors is sufficient to release the explanatory power of the macroeconomic variables in improving the estimation of the transition probability of the future credit rating.

Keywords

Credit ratings; Macroeconomic variables; Non-Markovian model

1. Introduction

A credit rating is an evaluation of the credit risk of a prospective debtor, predicting their ability to pay back the debt, and an implicit forecast of the likelihood of the debtor defaulting (Kronwald, 2009). Credit evaluation for companies and governments is generally done by a credit rating agency such as Standard & Poor's (S&P), Moody's, or Fitch. These rating agencies are paid by the entity that is seeking a credit rating for itself or for one of its debt issues.

Macroeconomic variables play a vital role in determining the rating for a company's credit rating. Banks' credit risks are assumed to be affected by macroeconomic variables, that is, when the macroeconomic conditions have been improved, the credit risk will be reduced. Several models have been developed and the results support the assumption.

Alves (2005), and Shahnazarian and Asberg-Sommer (2008) are the pioneers in the analysis of the relation between default probabilities and macroeconomic factors using Vector Autoregressive approach. They found short-term interest rates, economic growth and inflation to be the variables with significant effects on default frequencies. Distinguin, Raus and Tarazi (2006) found that market data helped to obtain more reliable default

probability. Simons and Rolwes (2012) however found that GDP growth is negatively related to probability of default while interest rate and exchange rate are positively related to the probability of default in certain sectors.

Castrén, Déés, and Zaher, (2008), Wong, Choi and Fong (2006), Otani, Shiratsuka, Tsurui and Yamada (2009), Avouyi-Dovi, Mireille, JarDET, Kendaoui and Moquet (2009), Jakubik and Schmieder (2008), Küçüközmen and Yüksel (2006), Atlintas (2012), and Figlewki, Frydman and Liang (2012) also investigated the dependence of credit risk on some selected macroeconomic variables. Rinaldi and Sanchis-Arellano (2006) reported that inflation rate, the ratio of financial assets to disposable income, disposable income itself, the ratio of household debt to household disposable income, and real lending interest rates were important variables affecting non-performing loans. Yurdakul (2014) reported that growth rate and ISE index reduce banks' credit risk, while money supply, foreign exchange rate, unemployment rate, inflation rate, and interest rate increase banks' credit risk.

The layout of this paper is as follows. In Section 2, the method based on multivariate power-normal (MPN) distribution (Pooi, 2012) is used to fit the data on credit ratings and macroeconomic variables. Section 3 presents the results on the effects of macroeconomic variables on the credit rating transition probabilities. Section 4 concludes this paper.

2. Method based on Multivariate Power Normal Distribution

The index of the i -th company in a group of N companies may be represented by the $N-1$ binary codes $0 \cdots 010 \cdots 0$ in which the value "1" takes the i -th position for $1 \leq i \leq N-1$, while the N -th company may be represented by $N-1$ zeros. In the case when the number of possible credit ratings is M , the rating j of a company may be represented by the $M-1$ binary codes $0 \cdots 010 \cdots 0$ in which the value "1" takes the j -th position for $1 \leq i \leq M-1$, while the rating M may be represented by $M-1$ zeros.

Consider a vector \mathbf{y} of $[N+3(M-1)]$ components consisting of the value of a particular selected macroeconomic variable in the present quarter and the codes for the index of a company together with those for the company's ratings in the previous, present and next quarters. We note that the use of ratings in the previous quarter will enable us to form a non-Markovian model (Gan et. al, 2017). From the credit rating data of N companies over N_q quarters, a table of $N \times (N_q - 2)$ rows is formed with each row representing a value for \mathbf{y} . The table can be partitioned into N sub-tables with the i_c -th ($1 \leq i_c \leq N$) sub-table representing the values of \mathbf{y} derived from the i_c -th company. Let n_w be a positive integer which is less than $(N_q - 2)$. From the i_c -th sub-table, we form the i_w -th sub-window using the first until the $(n_w - 1 + i_w)$ -th row

of the sub-table. Combining the i_w -th sub-windows for the N companies, we get the i_w -th window of $N \times (n_w - 1 + i_w)$ rows.

The data for \mathbf{y} in the i_w -th window is fitted with an $[N + 3(M - 1)]$ -dimensional MPN distribution. From the MPN distribution for the vector \mathbf{y} of $[N + 3(M - 1)]$ values, a large number N_s of the values of \mathbf{y} are generated. The components of \mathbf{y} may be divided into five groups of which group 0 consists of the value of the selected macroeconomic variable, group 1 consists of the second till the N -th components, group j ($3 \leq j \leq 5$) consists of the next $M - 1$ components. By using the criterion based on the distance defined in Gan and Pooi (2015), the codes for company and credit ratings are converted to integer values. Thus the components of \mathbf{y} are transformed to the vector $\mathbf{y}^{(1)}$ of which the first component gives the values of the selected macroeconomic variable, the second component represents the index of the company, while the last 3 components are the ratings in the previous, present and future quarters.

From the large number of the $\mathbf{y}^{(1)}$ generated, we form a table consisting of the values of $\mathbf{y}^{(1)}$ which correspond to a chosen company and the chosen ratings ($r^{(v)}$ and $r^{(p)}$, say) in the previous and present quarters. We next form a sub-table by deleting the second to fourth columns of the original table. A row in the sub-table then gives the value of a vector $\mathbf{y}^{(2)}$ of which the first component is the value of the selected macroeconomic variable and the second component is the rating in the next quarter for the selected company with the specified rating $r^{(v)}$ in the previous quarter and the rating $r^{(p)}$ in the present quarter.

When the first value of $\mathbf{y}^{(2)}$ is given by the first value of the i -th row of the sub-table, a conditional distribution is obtained for the second value of $\mathbf{y}^{(2)}$. From the conditional distribution, we obtain the probability P_j that the second component of $\mathbf{y}^{(2)}$ lies in the interval I_j :

$$I_j = (r^{(p)} + j - 0.5, r^{(p)} + j + 0.5], j = 0, -1, +1 \text{ if } r^{(p)} < 10,$$

or the interval

$$I_j = (r^{(p)} + j - 0.5, r^{(p)} + j + 0.5], j = 0, -1, -2 \text{ if } r^{(p)} = 10.$$

We may investigate the dependence of the probability P_j on the value of the selected macroeconomic variable given by the first component of $\mathbf{y}^{(2)}$.

Instead of investigating the effects of the macroeconomic variables, one at a time, we may summarize the effects of the 8 macroeconomic variables by a small number of latent factors, and include the latent factors into the non-Markovian model.

Ley \mathbf{y}^* be an $n_m \times 1$ vector consisting of the values of n_m macroeconomic variables. A table consisting of N_q rows may be formed such that in the table,

the j -th row represents the value of n_m macroeconomic variables in the j -th quarter. We form the i_w -th sub-table using the first to $n_w - 1 + i_w$ rows of the table.

A factor model with its static representation given by

$$\mathbf{y}^* = \mathbf{\Lambda}\mathbf{F} + \mathbf{e}, \tag{2.1}$$

may be used to describe the vector \mathbf{y}^* . In Equation (2.1), $\mathbf{\Lambda}$ is an $n_m \times r$ matrix of factor loadings, \mathbf{F} is an $r \times 1$ vector of common latent factors underlying \mathbf{y}^* and \mathbf{e} is a $n_m \times 1$ vector of random errors.

We perform a principal component analysis of the n_m columns of the observations in the i_w -th sub-table. Suppose the principal component with the i -th largest variance is f_i . We obtain the first r^* principal components ($r^* < (n_m + 1)$) $f_{1j}, f_{2j}, \dots, f_{r^*j}$. Suppose f_{ij} is the value of f_i extracted from the j -th row of the sub-table. The row vector $\mathbf{f}_j = (f_{1j}, f_{2j}, \dots, f_{r^*j})$ then represents the values of r^* important latent factors in the j -th quarter.

Whenever the first value of the vector \mathbf{y} represents the value of the macroeconomic variable in the j -th quarter, we replace this first value by the value of \mathbf{f}_j which represents the values of r^* important latent factors in the j -th quarter. In this way we can obtain the i_w -th window of $N \times (n_w - 1 + i_w)$ rows, each of which represents of an updated value of \mathbf{y} .

The data for \mathbf{y} in the i_w -th window is fitted with an $[r^* + (N - 1) + 3(M - 1)]$ -dimensional MPN distribution. From the fitted MPN distribution, a large number N_s of the values of \mathbf{y} are generated. The components of \mathbf{y} are transformed to the vector $\mathbf{y}^{(1)}$ of which the first r^* components gives the values of the r^* latent factors, the $(r^* + 1)$ -th component represents the index of the company, while the last 3 components are the ratings in the previous, present and future quarters.

From the large number of the $\mathbf{y}^{(1)}$ generated, we form a table consisting of the values of $\mathbf{y}^{(1)}$ which correspond to a chosen company and the chosen ratings in the previous and present quarters. We next form a sub-table by deleting the $r^* + 1$ to $r^* + 3$ columns of the original table. A row in the sub-table then gives the value of a vector $\mathbf{y}^{(2)}$ of which the first r^* component are the value of the r^* latent factors and the $(r^* + 1)$ -th component is the rating in the next quarter for the selected company with the specified rating $r^{(v)}$ in the previous quarter and the rating $r^{(p)}$ in the present quarter.

When the first r^* values of $\mathbf{y}^{(2)}$ are given by the first r^* values in the i -th row of the sub-table, a conditional distribution is obtained for the $r^* + 1$ value of $\mathbf{y}^{(2)}$. From the conditional distribution, we obtain the probability P_j that the $(r^* + 1)$ -th component of $\mathbf{y}^{(2)}$ lies in the interval I_j . We may investigate the dependence of the probability P_j on the values of the r^* latent variables given by the first r^* components of $\mathbf{y}^{(2)}$.

3. Numerical Results

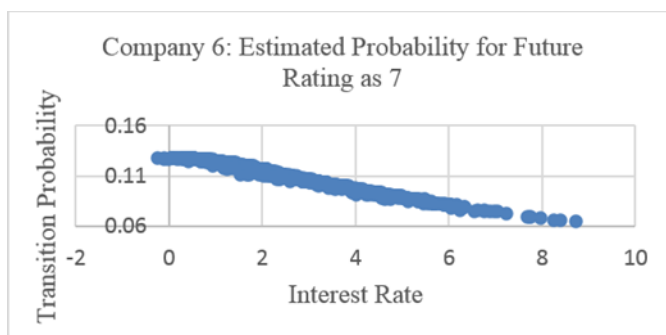
Consider the credit ratings of 10 companies in the electronics sector in 15 years (from 1999 to 2013) taken from the database of Taiwan Economic Journal. The values of eight selected Taiwan macroeconomic variables from year 1999 to year 2013 are also obtained.

The results for Companies 2 and 6 are chosen to illustrate the impact of macroeconomic variables. From Figure 3.1, we note that for Company 6, an increase in the interest rate tends to (a) decrease the probability of improving the rating to 7 from present rating of 8 and (b) increase the probability of staying in rating 8 or deteriorating to rating 9 in the next quarter when the rating in the present quarter is 8. Thus an increase in interest rate would increase the credit risk. From the plots of transition probabilities for other macroeconomic variables, we obtain Table 3.1.

Table 3.4 Effects of increased values of Taiwan macroeconomic variables on credit risk (“+” (or “-”) sign indicates that the macroeconomic variable tends to increase (or decrease) the credit risk)

Macroeconomic Variable	Company 2	Company 6
Interest Rate	+	+
Exchange Rate	+	+
TSEC weighted Index	-	-
Gold Prices	-	-
Money Supply M2	-	-
Inflation Rate	No clear indication	No clear indication
Unemployment Rate	-	+
GDP	-	-

The results in Table 3.1 basically agree with the corresponding results reported so far in the literature.



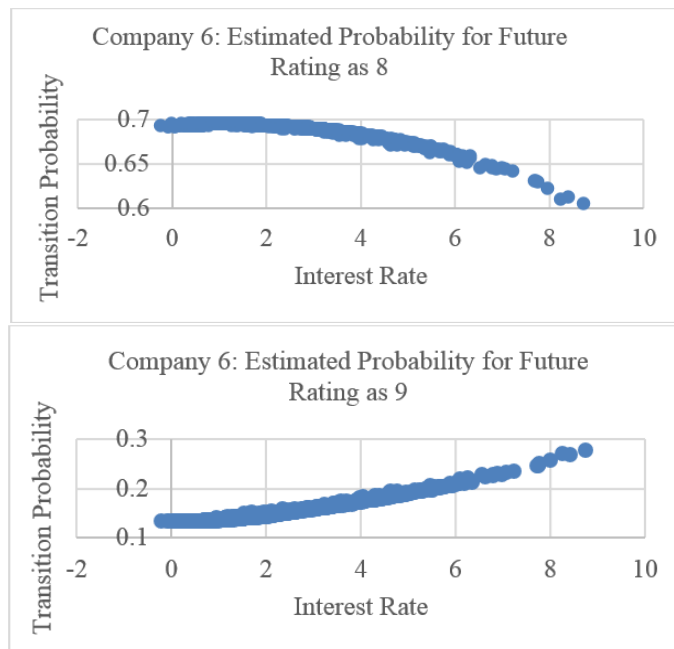


Figure 3.1 The estimated probability that the future rating is 7, 8, and 9, respectively, for Company 6, given the interest rate for present quarter, and the ratings for previous and present quarters are both equal to 8.

We have obtained the plot (not presented here) of the value of the i -th ($1 \leq i \leq r^*$, $r^* = 4$) latent factor with the probability of Company 2 staying in rating 10 given that its ratings in the previous and present quarters are both 10. The plots when $i = 1$ and 3 show obvious variation of the probability of staying in rating 10 when the first and third latent factors vary.

The similar plots for other companies with given ratings in the previous and present quarters are expected to exhibit variation of transition probabilities when the values of some of the latent factors vary. The explanatory power shown by the latent factors is consistent with the previous finding that some of the individual macroeconomic variables have effects on the credit risk.

4. Discussion and Conclusion

The latent factors extracted from the Taiwan macroeconomic variables exhibit an explanatory power over the credit risk. It might be possible to use the predicted probabilities in a number of consecutive quarters to get an indication of the possible transition of rating in the near future. It is also possible that the latent factors extracted from the macroeconomic variables in other countries will also help to give a better prediction of the credit risk.

References

1. Altıntaş, A. (2012). Kredi kayıplarının makro ekonomik değişkenlere dayalı olarak tahmini ve stres testleri. Türk bankacılık sektörü için ekonometrik bir yaklaşım (Estimation of credit losses with respect to macroeconomic indicators and stress tests – An econometric approach for Turkish banking). *Türkiye Bankalar Birliği*, 281, İstanbul.
2. Alves, I. (2005). Sectoral fragility: Factors and dynamics. *Investigating the Relationship between the Financial and Real Economy*, Bank for International Settlements, 22, 450-80.
3. Avouyi-Dovi, S., Mireille, B., Jarret, C., Kendaoui, L., & Moquet, J. (2009). Macro stress testing with a macroeconomic credit risk model: Application to the French manufacturing sector. *Banque de France Working Paper No. 238*.
4. Castrén, O., Déés, S., & Zaher, F. (2008). Global macro-financial shocks and expected default frequencies in the euro area. Working Paper Series 875, European Central Bank.
5. Distinguin, I., Rous, P., & Tarazi, A. (2006). Market discipline and the use of stock market data to predict bank financial distress. *Journal of Financial Services Research*, 30(2), 151-176.
6. Figlewski, S., Frydman, H., & Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, 21(1), 87-105.
7. Gan, C. P., & Pooi, A. H. (2015). Estimation of transition probabilities of credit ratings. *AIP Conference Proceedings* (Vol. 1691, No. 1, p. 050005). AIP Publishing.
8. Gan, C.P., Pooi, A.H. and Ng, K. H. (2017). Prediction of Future Credit Rating Using a Non-Markovian Model. *AIP Conference Proceedings* (Vol. 1830, p. 080002). AIP Publishing.
9. Jakubik, P., & Schmieder, C. (2008). *Stress Testing Credit Risk: Comparison of the Czech Republic and Germany, FSI Award 2008 Winning Paper*. Financial Stability Institute, Bank for International Settlements, Switzerland.
10. Kronwald, C. (2009). *Credit Rating and the Impact on Capital Structure*. Munich, GRIN Verlag.
11. Küçüközmen, C. C., & Yüksel, A. (2006). A macroeconometric model for stress testing credit portfolio. In *13th Annual Conference of the Multinational Finance Society*.
12. Otani, A., Shiratsuka, S., Tsurui, R., & Yamada, T. (2009). Macro stress-testing on the loan portfolio of Japanese banks. Bank of Japan Working Paper Series No. 09-E-1.
13. Pooi, A. H. (2012). A model for time series analysis. *Applied Mathematical Sciences*, 6(115), 5735-5748.

14. Rinaldi, L., & Sanchis-Arellano, A. (2006). Household debt sustainability: What explains household non-performing loans? An empirical analysis. European Central Bank Working Paper, No. 570.
15. Shahnazarian, H., & Asberg-Sommer, P. (2008). *Macroeconomic impact on expected default frequency*. Sveriges Riksbank Working Paper Series 219.
16. Simons, D., & Rolwes, F. (2012). Macroeconomic default modeling and stress testing. BiblioGov.
17. Wong, J., Choi, K. F., & Fong, T. (2006). A framework for macro stress testing the credit risk of banks in Hong Kong. *Hong Kong Monetary Authority Quarterly Bulletin*, 10, 25-38.
18. Yurdakul, F. (2014). Macroeconomic modelling of credit risk for banks. *Procedia-Social and Behavioral Sciences*, 109(8), 784-793.



Probability distribution model for predicting ozone (O₃) exceedances at two air quality monitoring sites in Malaysia during dry season



Norshahida Shaadan^{1,2,3}; Nabihah Jasri¹

¹ Centre for Statistical and Decision Science Studies, Faculty of Computer & Mathematical Sciences, UiTM

² Advanced Analytic Engineering Center, Faculty of Computer & Mathematical Sciences, UiTM

³ Business Datalytic Research Group, UiTM

Abstract

Tropospheric Ozone (O₃) is one of the strongest atmospheric oxidants and has become an important criteria pollutant in the Malaysia environment other than PM₁₀. Many studies worldwide have proven that, high concentration of O₃ contributes to a certain environmental problem including health problems, vegetation and materials, as well as climate changing. Thus due to the facts, it is necessary to gain a good understanding of the characteristics of O₃ pollution so that the information can be the input for managing the problem. In this study, several probability distributions including Gamma, Lognormal, Normal, and Weibull were compared with the aim to find the best distribution that can fit the O₃ at two selected air quality monitoring stations in Malaysia located in Shah Alam and Putrajaya. Based on the two years (2013 and 2014) period of hourly recorded data, the model parameters were estimated using the method of maximum likelihood estimator (MLE). The best distribution was determined using the plot of cumulative distribution function (CDF) and the goodness of fit statistic including the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling. The study results have shown that Weibull is found to be the best model for Shah Alam while Gamma model is for Putrajaya with expected exceedance return period of seven days.

Keywords

Ozone; Prediction Model; Probability Distribution; Pollutant Exceedances; Air Quality

1. Introduction

Nowadays, air pollution has become a global problem. Over exposed to polluted air that contained mixed hazardous gases, dust or hazes has been proven to be negatively affect human's health, animals, plants and the surroundings (Verma et al., 2015; Felzer et al., 2007). The substances that caused air pollution are called pollutants. In the Malaysia environment, Ozone (O₃) has been identified as the second most dominant pollutant other than particulate matter (PM₁₀). Ozone (O₃) is defined as a secondary type of gas pollutant and also known as photochemical oxidant that is formed via

photochemical reactions of precursors such as volatile organic compound (VOC_s) and Nitrogen Oxides (NO_x) (Lee et al., 2010). During hot and sunny day, the formation of O₃ is greatly enhanced. The variation of O₃ is largely depends on meteorology variables such as sunlight, temperature and wind (Ling et al., 2014).

Surface or Troposphere O₃ contributes to a certain environmental problem including health problems, vegetation and materials, as well as climate changing. O₃ is defined as a secondary pollutant whereby volatile organic compounds and NO_x from point sources are the precursors of Ozone in the country. Ozone shows strong day-to-day variation and sometimes virtually undetectable (Ramli et al., 2010). In the Malaysia environment, the main sources of VOC_s and NO_x emissions are from industries and motor vehicles particularly in urban areas (DOE 2014). In Klang Valley for example, it is found that O₃ exceedances pattern is strongly influenced by local pollutant emission and its dispersion characteristics. It is also observed that the variation of O₃ concentration is continuous in nature and its severity behaviour is often uncertain (Shaadan et al., 2018). O₃ pollution is odourless, the physical form is invisible and the pattern of occurrence is less understood. Thus, the existence of ozone pollution is often unaware.

According to Ghazali et al. (2014), Malaysia has not yet reached a critical level of the air pollution as in other metropolitan areas in Asia but its potential increased need to be assessed and predicted. Data monitoring and studies on air quality shows that some of the air pollutants in several large cities such as Johor, Selangor and Sabah are increasing with time and are not always at a satisfied level in comparison to the national standard (Rafia & Ibrahim, 2002). Since Ozone is one of the important pollutants, for the purpose of managing and mitigating air pollution, it is important to study and predict the return period of ozone pollution in the future period. Malaysia weather is characterized by two main seasons, the Northwest and Southwest monsoons and other two transition periods. Drier weather condition often recorded during Southwest monsoon and during the second and the first transition period at several regions. On the other hand, more frequent and larger amount of rain were recorded during the Northwest monsoon particularly at the east coast region in Peninsular Malaysia. As reported by Siti et al., (2015)), air pollution occurrence and high level of pollutants concentration often recorded during the dry season.

There are many statistical methods and models can be used for predicting air pollutant level. However, the application of probability models for predicting high pollutant levels and the exceedance is rarely discussed on O₃ data in the Malaysia environment. Majority of the studies had discussed on PM₁₀ pollutant since PM₁₀ is reported as a major component of haze. To fill the gap, this study will focus on the comparison of several chosen probability

models that can appropriately represent O₃ distribution at Shah Alam and Putrajaya air quality monitoring sites, during dry season in Malaysia (i.e. Southwest monsoon) and further, the model will be used to predict O₃ exceedances.

2. Methodology

This study used secondary data which was obtained from the Air Quality Division, Department of Environment (DOE) in Putrajaya, Malaysia. The data consist of daily by hourly recorded O₃ concentration for two years period of time from 2013 to 2014 at two air quality monitoring stations; the Shah Alam and Putrajaya stations. In this study, four probability distributions were used to fit the hourly O₃ concentration (ppm) that were recorded during dry season which include; Gamma, Lognormal, Normal and Weibull. Due to the small values of the observed O₃, the analysis was then conducted using modified scale values of 10 times of the observed.

The estimation of Gamma probability density function parameter (α =shape parameter and β =scale parameter), Lognormal probability density function parameters (α =scale parameter and β =shape parameter), Normal probability density function parameter (α =scale parameter and β =location parameter) and Weibull probability density function parameter (α =scale parameter and β =shape parameter) were done using the maximum likelihood estimators (MLE) methods. Several goodness of fit statistic was used to determine the distribution that can give the best fit to the data. The goodness of fit criteria or statistics includes the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling. The three goodness of fit were compared and the one with the lowest value indicate the best distribution.

Table 1.1
Goodness of Fit Statistics as define by D'Agostino and Stephens (1986)

Statistic	General formula	Computational formula
Kolmogorov-Smirnov (KS)	$\sup F_n(x) - F(x) $	$\max(D^+, D^-)$ with $D^+ = \max_{i=1, \dots, n} (\frac{i}{n} - F_i)$ $D^- = \max_{i=1, \dots, n} (F_i - \frac{i-1}{n})$
Cramer-von Mises (CvM)	$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^n (F_i - \frac{2i-1}{2n})^2$
Anderson-Darling (AD)	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} \sum_{i=1}^n (2i-1) \log(F_i(1-F_{n+1-i}))$

** Where $F_i \triangleq F(x_i)$, F =fitted cumulative distribution function, F_n =empirical distribution function, x_i =observation

To evaluate the performance of a fitted distribution model in comparison to the empirical distribution, the visualization approach is used. The CDF plots for the theoretical distribution curve and the empirical distribution curve were plot together in the same graph.

The performance of the best distribution model obtained was validated using the coefficient of determination R^2 . High value of R^2 (> 0.8) and a linear pattern of the plotted points between the predicted and the observe values indicate that the fitted model is a good model to represent the random distribution of Ozone hourly concentrations and the prediction can be done.

Once the best-fit distribution is determined, the cumulative distribution function (CDF) of the fitted distribution was used to predict exceedances and the return period. Exceedances is defined as the condition when ozone concentration level exceeds or greater than the given government standard ($H = 0.1$ ppm). In environmental problems, return period is considered as the average number of days between exceedances. Thus to compute the return period (T) of exceedances of Ozone concentration (x) using CDF, the formula is given by:

$$T = \frac{1}{1-F(x_{\text{exceedances}})} \quad (1)$$

where $F(x_{\text{exceedance}})$ is the CDF of the ozone concentration and $x_{\text{exceedance}}$ is when $x \geq H$.

$$1 - F(X) = P(x_{\text{new}} > x_{\text{exceedances}}) \quad (2)$$

According to Malaysia Air Ambient Quality Guideline (MAAQG), the standard for exceedances is whenever the hourly ozone concentration exceeds 0.1 ppm.

3. Results

The hourly ozone concentration for Southwest monsoon is reported in Table 1 shows the descriptive statistics for O_3 concentration for Shah Alam and Putrajaya areas. The statistics minimum O_3 for the areas are about the same with level of 0.000 ppm while the maximum O_3 for Shah Alam and Putrajaya are 0.1410 ppm and 0.1150 ppm respectively.

The results indicate that that Shah Alam has higher O_3 concentration compared to Putrajaya. Besides, the mean values were higher than median values for the three areas. The result indicates that there exist outlying data. The statistic standard deviation for Putrajaya is larger than Shah Alam. The results indicate that Putrajaya has higher variability in the Ozone concentration level. For all areas, the coefficient of skewness and kurtosis are greater than zero. The result shows that the Ozone distribution at the locations are positively skewed, if the model were to be fitted to the Ozone data, a skewed model is more appropriate. Parameter estimates for four distributions according to Southwest Monsoon season data are shows in Table 2. These

parameter estimates have been obtained by using maximum likelihood estimators (MLE). Results in Table 2 shows that for both Gamma and Weibull distribution, Putrajaya has larger value of estimated parameters compared to Shah Alam.

Table 1 Descriptive Statistics for O3 Concentration for Southwest Monsoon

	Value (ppm)	
	Shah Alam	Putrajaya
Minimum Value	0.0000	0.0000
Maximum Value	0.1410	0.1150
Mean	0.0164	0.0200
Standard Deviation	0.0177	0.0181
Median	0.0100	0.0150
Skewness	1.6018	1.3189
Kurtosis	5.9446	4.6863

Table 2 Parameter Estimate for Southwest Monsoon

Distribution	Parameter Estimate	
	Shah Alam	Putrajaya
Gamma	$\alpha = 0.8007$ $\beta = 0.4893$	$\alpha = 1.0638$ $\beta = 0.5331$
Lognormal	$\alpha = 1.4736$ $\beta = -0.2487$	$\alpha = 1.3047$ $\beta = 0.1520$
Normal	$\alpha = 1.7699$ $\beta = 1.6361$	$\alpha = 1.8074$ $\beta = 1.9951$
Weibull	$\alpha = 0.8647$ $\beta = 1.5208$	$\alpha = 1.0539$ $\beta = 2.0359$

Table 3 Goodness of Fit for Southwest Monsoon

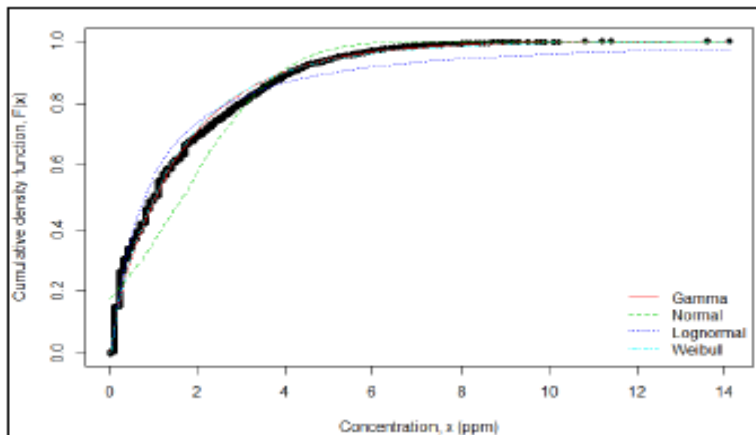
Distribution	Statistics	Shah Alam	Putrajaya
Gamma	Kolmogorov-Smirnov	0.1045	0.0572
	Cramer-von Mises	9.1923	3.1828
	Anderson-Darling	74.4673	25.2371
Average Score (Gamma)		28.2547	9.4741
Weibull	Kolmogorov-Smirnov	0.1055	0.0611
	Cramer-von Mises	9.1044	3.1661
	Anderson-Darling	74.5250	25.5210
Average Score (Weibull)		27.9116	9.5827

Table 3 presents the result of goodness of fit test. Based on the average value of the test statistics; the Kolmogorov-Smirnov, Cramer-von Mises and

Anderson-Darling, the study provides the evidence that O₃ level can be best modelled by Weibull and Gamma at Shah Alam and Putrajaya sites respectively.

The results of CDF plots are shown in Figure 1. The black line is the empirical cumulative distribution function (ecdf) and the colour lines are CDFs plots from different distributions which are Gamma, Lognormal, Normal and Weibull. Figure 1 also shows that Normal and Lognormal distributions are not good fits to the data because the lines are quite far away from the ecdf line. The worst distribution is Lognormal since the lines are very far from the ecdf line. Since Gamma and Weibull distributions are closest to the ecdf line, it indicates that Gamma and Weibull distributions give a good fit to the data at Shah Alam and Putrajaya.

(a) Shah Alam



(b) Putrajaya

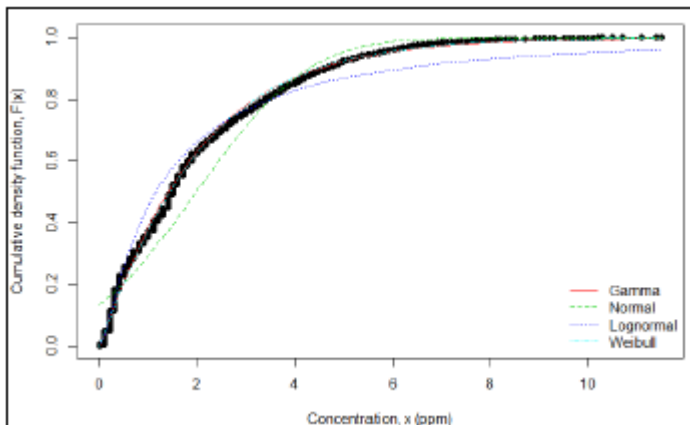


Figure 1 Cumulative Distribution Function Plot in (a) Shah Alam and (b) Putrajaya for Southwest Monsoon

Table 4 Model Performance and Estimated Return Period

	Best Distribution	R²	Return Period (days)
Shah Alam	Weibull distribution	0.9293	6.8306
Putrajaya	Gamma distribution	0.9928	7.4405

Based on the best model obtained, Table 4 shows the results to assess the performance of the model. High value of the coefficient of determination (R^2) of 0.9293 for Shah Alam data and 0.9928 for Putrajaya indicates a very strong capability of the model for prediction since the observed and predicted value is proven to have a strong linear relationship. Using formula given by equation (1) and (2) with the threshold of 0.1 ppm and the scaled standard, $X_{standard\ new}$ of 10 (i.e. with respect to the modified scale data used in the analysis mentioned in methodology section) the study has estimated that the return period of O_3 exceedance at both locations is about 7 days.

4. Conclusion

An accurate tool or model is vital for prediction of pollutant level. Thus, this study has contributed a suitable probability model that can be used to predict O_3 level and its exceedances. Weibull model is found the best for Shah Alam while Gamma model is the best for Putrajaya. The findings from this study are important to help the responsible body to manage and mitigate air pollution problem due to O_3 emission since the occurrence of exceedance is expected to arrive within the cycle of 7 day period. The results of this study can also be used to facilitate further studies.

Acknowledgment

The authors would like to thank the Department of Environment Malaysia for providing the data.

References

1. Department of Environment. (2014). Malaysia Environmental Quality Report 2014.
2. Felzera, B.S., Cronina, T., Reilly, J.M., Melillo, J.M., & Wang, X. (2007). Impacts of Ozone on trees and crops. *C.R. Geoscience* 339, 784–798
3. Ghazali, A. N., Yahaya, S. A., & Mokhtar, Z. M. (2014). Predicting Ozone concentrations levels using probability distributions. *ARPN Journal of Engineering and Applied Sciences, Vol.9*

4. Lee, S.B., Bae, G.M., Lee, Y.M., Moon, K.C., & Choi, M. (2010). Correlation between Light Intensity and Ozone Formation for Photochemical Smog in Urban Air of Seoul. *Aerosol and Air Quality Research*, 10, 540–549
5. Ling, Z. H. Ling, Guo, H., Lam, S. H. M., Saunders, S. M., & Wang, T. (2014). Atmospheric photochemical reactivity and ozone production at two sites in Hong Kong: Application of a Master Chemical Mechanism–photochemical box model. *Journal of Geophysical Research: Atmosphere*, 119 (10), 567–582
6. Rafia, A., Hassan, M. N., & Ibrahim, N. A. (2002). Review of air pollution and health impact in Malaysia. *Environmental Research*, 71–77.
7. Ramli, N.A., Ghazali, N.A., & Yahaya, A.S. (2010). Diurnal fluctuations of ozone concentrations and its precursors and prediction of ozone using multiple linear regressions. *Malaysian Journal of Environmental Management*, 11 (2), 57–69
8. Shaadan, N., Roslan, R.R., Deni, S.M. & Ismail, A. *The diurnal Ozone dynamics profile at several locations in Selangor, Malaysia: A functional data analysis approach*. International Conference on Mathematics, Engineering and Industrial Applications 2018 (ICoMEIA 2018). AIP Conference Proceedings 2013, 010001 (2018).
<https://doi.org/10.1063/1.5054199>
9. Siti, R. A., Sharifah, N. S., Muhammad, F., Mohd, T. L., Emilia, Z. A., & Praveena, S. M. (2015). The assessment of ambient air pollution trend in Klang Valley, Malaysia. *World Environment*, 1–11.
10. Verma, N., Satsangi, K., & Kumari, K.M. (2015). Prediction of Ground Level Ozone Concentration in Ambient Air using Multiple Regression Analysis. *J. Chem. Phys. Sci.* 5(4), 3685–3696



Collaboration on SDG-data between national stakeholder groups



Maciej Truszczynski

Statistics Denmark, Copenhagen, Denmark

Abstract

Four years after the Sustainable Development Goals (SDGs) have been adopted, a growing number of stakeholders recognize that the fulfilment of the SDGs will only be possible, if all stakeholder groups, public and private, play an active role in implementing the 2030-agenda. At the centre of the implementation of the SDGs lie the data on which performance will be measured and evaluated and new initiatives taken, irrespectively of whether you are a nation, a NGO or a private sector representative. A number of initiatives focusing on SDG data have been taken which either is focusing on governments or private sector performance reporting on the SDG. Amongst other, initiatives have been taken by the UN World Data Forum, UNCTAD-ISAR, UN Global Compact, the SDG Index and many others. In the spirit of SDG 17, a number of initiatives on SDG-data have been taken. In Denmark, Statistics Denmark has initiated collaboration between all stake-holders in Denmark on Partnership for SDG data with the aim of co-creating the solutions needed. Some of the key challenges of the current initiatives on SDG-reporting are:

1. Focus is on reporting the SDG indicators on national level by governments or solely on specific stakeholder group SDG-reporting – we argue that the UN process should cater for all stakeholder groups voices as part of the review of the progress towards meeting the SDGs through reporting.
2. There is no differentiation between reporting on do no harm and positive pursuits – we argue both should be present for balanced reporting.
3. Comparable data for all stakeholder groups are not available – we argue that focus should be on progress on national or company level – comparison should focus on how well nations and companies developed

The event will provide an opportunity to discuss experiences and learnings from working cross-collaborative between all stakeholder groups in order to secure a focus on both do no harm and potential positive pursuits, in order to ensure a balanced approached and good governance, so that the reporting on sustainable results towards 2030 builds on trust between governments, NGOs and the private sector.

Keywords

SDG; Sustainable Development; Indicators; Reporting Platform; Private Sector

1. Introduction

Four years after the Sustainable Development Goals (SDGs) have been adopted, it becomes more and more clear to a growing number of stakeholders that the fulfilment of the SDGs will only be possible, if all stakeholder groups, public and private, play an active role in the implementation of the 2030 Agenda.

At the centre of the implementation of the SDGs lie the data on which performance will be measured and evaluated and new initiatives taken, irrespectively of whether you are a government official, an NGO or a private sector representative. Statistics Denmark has from the outset shown engagement and willingness to become a major voice in the debate on the statistical follow up on the 2030 Agenda in Denmark. The first step in this direction was developing a data platform (national reporting platform) to present the Danish follow-up on the global indicators. The development of the platform was organized as a project with a project manager and a steering group. In its current form, the data on the platform is mainly based on the data sources from Statistics Denmark's own production system and is compiled in collaboration with different units within the office. But it is our ambition to develop the platform so that it shows development in other SDG relevant domains, such as private sector's activities and contribution to the SDG, and civil society's activities. In order to address those ambitions, the existing data in Statistics Denmark is not sufficient. It is therefore necessary to reach out to other data producers and sources, and investigate what data can be applicable. In this regard, Statistics Denmark is conducting a number of initiatives focusing on SDG data and is in dialogue with various stakeholders, be it private sector, civil society, or municipalities and is considering how to include data from different sources in the statistical follow-up on the SDGs.

Among those initiatives, the goal of following up on private sector activities relating to the 2030 Agenda is the focus of this paper. Other initiatives, such as data disaggregation, use of data from non-official sources, coordination of data streams for the global reporting, and compilation of national indicators will also be described in this paper, though to a lesser degree.

As a concluding remark, in the spirit of SDG 17, Statistics Denmark has established a Partnership for SDG data. The primary idea was collaboration between all stakeholders to ensure the best possible reporting on the UN indicators and a dialogue between data producers, data users and stakeholders. Furthermore, through a dialogue with stakeholders, a parallel goal was to ensure that it is data that are topic for debate and not the methodology or data sources. And finally, a dialogue with a wide range of stakeholders draws our attention to other data sources than those already used in Statistics Denmark.

Beginning with description of the work with the follow up on private sector activities, the text below sheds some light on different issues Statistics Denmark has experienced during the work on the follow up on the SDGs:

2. Methodology

Private sector was on the 'SDG board' right from the adoption of the 2030 Agenda. In Denmark, the measurement of contribution of the private sector to the global goals was and is a heavily discussed topic. The activities of the private sector were and still are followed by various stakeholder groups that request more and comparable data about them. There are various challenges in measuring contribution of the private sector to the 2030 Agenda in a statistical way. Firstly, the 2030 Agenda itself does not have many indicators that are directly applicable to the activities of the private sector. In consequence, there are parallel attempts to propose new indicators with relevance for the follow up on the private sector, including activities conducted by UNCTAD or GRI. Indicators proposed here pose new demands to data availability which to a large extent are not yet met by the current statistical production. Besides data gaps, the comparability of company CSR reporting is lacking. This challenges the NSOs in compiling statistics, as the CSR reports provide one of the best information sources for approximating the intentions of the 2030 Agenda.

The ultimate ambition of following up on private sectors' activities regarding 'doing no harm' and 'positive pursuits' is far from accomplished. Statistics Denmark is currently conducting a pilot project that aims to present figures on private sector that complement the existing follow-up on the 2030 Agenda. The project focuses on company strategies regarding the SDG and on some selected targets. To reach this goal, Statistics Denmark has prepared a specific 'company' questionnaire and has plans of complementing the replies with figures on UNCTAD core indicators. The full results of the pilot are to be found on Statistics Denmark SDG web site.

The pilot project mainly addresses non-agricultural private sector and consists of the following elements:

- a. Presentation of enterprise statistics that complements the ongoing compilation of global indicators and the Danish follow-up on the SDGs.
- b. Presentation of statistical results on how and in what areas Danish enterprises contribute to the fulfilment of the global goals. This part includes a questionnaire.
- c. A number of companies will get a possibility to present themselves and their contribution to the fulfilment of the global goals on our SDG platform in a specific section for enterprises – within the boundaries of Statistics Denmark acting as a neutral 'communicator' of data.

AD a.

The main purpose is to build on existing business statistics and develop an additional statistical information split up by industries to give a good first picture of SDG related activities of the Danish companies. This additional information can be about decomposing existing SDG indicators to a more detailed level, but also about supplementary indicators that can shed some extra light on activities with relevance for the SDGs, such as UNCTAD indicators. It is our aim that the information provided will go beyond actual business statistics and include green national accounts, business enterprises' expenditures on supplementary education, gender balance, pay gap, etc. The following indicators can be mentioned among those considered:

- Contribution to GDP
- Research and Development share of GDP
- Turnover for environmental goods and services
- Corporate taxes
- CO2 emissions compared to value added
- Energy intensity
- Women's pay as a percentage of men's pay
- Gender balance in boards and managements of companies

The final selection of indicators will be agreed upon by a 'follow-up' group consisting of representatives from various stakeholders, such as Danish Business Authority, Danish Business Association, or Danish Industries.

AD b.

Besides the general, existing, statistics, there is a need for new statistics to present what global goals private sector companies are contributing to – and if possible to demonstrate why, how and to what extent. In order to address this question, Statistics Denmark compiled a voluntary qualitative questionnaire. In the course of the discussions on the questionnaire it was decided to focus on Danish companies with more than 250 employees. The choice was determined by the fact that it was too extensive to contact all Danish companies. Besides, the companies with over 250 employees cover widely the different industries and furthermore, they also employ 1/3 of the workforce in the private sector. There were expressed some considerations that selection of companies with over 250 employees could bias the results of the survey, as it is expected that larger companies are more aware of 2030 agenda than smaller ones. The questionnaire was built around the following themes/questions:

1. Does your company focus on fulfilment of one or more SDGs?
2. What is the current policy/practice regarding the SDGs (after 2015)
3. Why does your company focus on the SDGs?
4. What is the current contribution to the fulfilment of the SDGs (since 2015)

5. Way ahead – how is your company going to contribute to the SDGs in the course of the next 2-3 years?

One of the challenges, particularly regarding ‘question 4’ was selection of relevant targets/indicators for the questionnaire. As only few, if any, indicators can be directly linked to the areas of work of private companies, it was decided only to select relevant targets, as their application can be extended to also include private sector activities. Here, 21 targets were selected. The goal is also to group the responses by industries and also, where possible, to disaggregate industries into more detailed levels.

One of the important challenges/considerations that could be mentioned, was whether companies should be asked about additional actions regarding the SDGs or whether it was sufficient to refer to already existing operations. The main concern behind this consideration was, whether we only should measure how the 2030 agenda impacts the activities of companies or also include general operations of companies as SDG relevant. In this line, other consideration was whether the questionnaire will address future activities or also past/ongoing. Here, it was decided to measure all activities that could relate to the 2030 agenda after it was adopted.

The overarching question is also who will be the users of the data on private sector as a whole – would it be general public or would data also have appeal to other users, such as investors? The latter would set new demands on data, including answers to the challenging issue of data confidentiality and privacy of statistical units.

AD c.

The project will open for a possibility for private sector companies to link to their descriptions of SDG activities. On the one hand, this could encourage companies to participate in the questionnaire and on the other, to complement the results of the questionnaire with information that may not be strictly perceived as statistical, but still considered as relevant by the users.

3. Results

Statistics Denmark has asked around 600 larger Danish enterprises about their activities relating to the SDG. A total of 178 major Danish companies have responded to the survey on the business community's commitment to the UN's world goals. The result can hardly be perceived as representative for smaller enterprises and the extent to which they are representative for larger enterprises can be discussed.

Of the 178 companies, nearly 120 responded that they think the global goals into the company's activities. This corresponds to approx. 2/3 of the

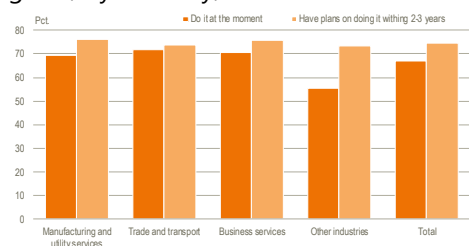
participating companies. The picture is pretty much the same, both for industrial companies, commercial, transport, or service companies.

Generally, it can be observed that a big part of larger Danish companies indicate awareness of the Global Goals. On the other hand, less than 50 per cent of larger Danish companies indicate concrete actions. It cannot be concluded against this background, how the situation is among medium and small Danish companies. The ambition of taking a higher social responsibility is among top scorers together with the ambition of ensuring a higher degree of equal pay.

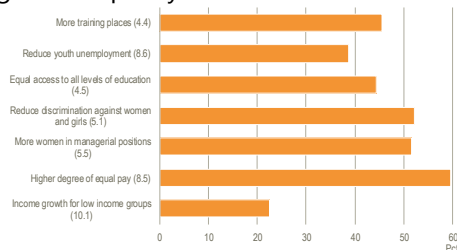
Among the 'low scorers' is the explanations that the purpose of company activities on Global Goal is to attract investors, together with that companies focus their activities on 'reducing food waste' and 'cheap medicine in developing countries'. But both 'reduce food waste' and 'cheap medicine in developing countries' can be industry specific and dependent on the number of companies working within those industries. Finally, only a smaller share of companies has answered that they are prioritizing income growth by income groups.

Some of the survey's results are presented in the figures below.

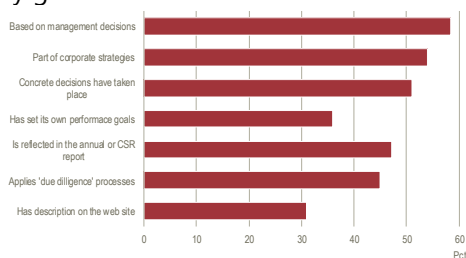
Enterprises that focus on the global goals, by industry, share



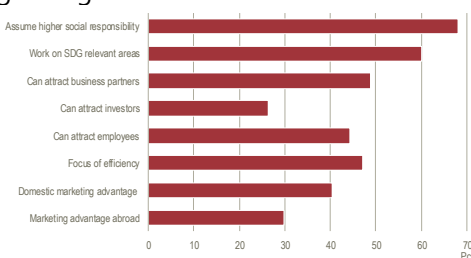
Global goals by target, inclusion and gender equality



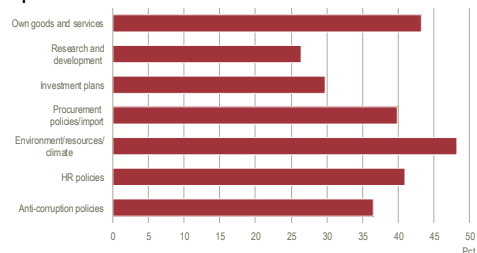
Enterprise actions on the global goals by general actions



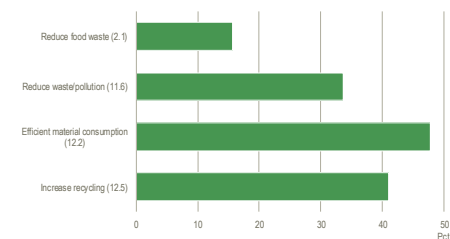
Enterprise explanations for focus on the global goals



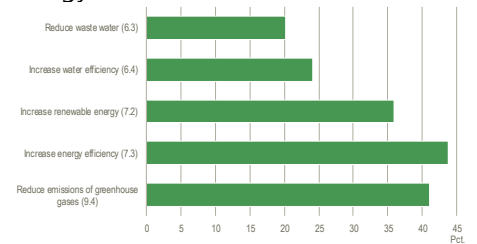
Enterprise action on the global goals by specific efforts



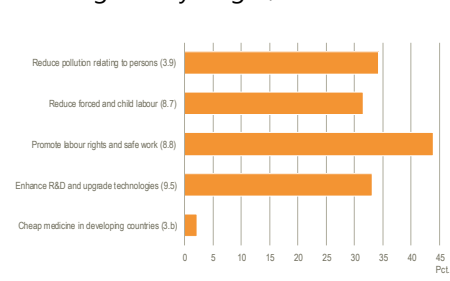
Global goal activities by target, resources and waste



Global goal activities by target, Co2, energy and water



Global goals by target, others



4. Discussion and Conclusion

There is still a need for extensive work and consultations on the shape of the SDG follow up on private sector activities. This follow up is very complex both due to the (lack of) data availability and the differences between companies. The question to be posed here is what can be changed in order to improve data availability. One of the possibilities considered could be to discuss whether it is possible to streamline company CSR reports, so that they live up to a specific set of requirements and are comparable at least to some extent. Furthermore, it was also considered whether existing (not only statistical) data streams could be captured at different sources, e.g. from company reports to stock exchanges? And finally, how can the follow up on company activities relating to the SDG be shaped in a way that is also has appeal to investors?

The next issue is the number of stakeholders in the attempts to follow up on private sector activities' relating to the SDG. The coordination between the areas of different stakeholders, in the view of the author of this paper, does not seem predominant which could result in either fragmented or overlapping follow up.

The final issue is to what extent pursuing of a goal of 'doing no harm' and 'positive pursuits' can be followed up on, taking the existing data availability into account? And if not, is it possible to create comparable data across individual industries?

5. Annex: Other aspects of work on the follow up on the Global Goals

Disaggregation of data. In the context of the SDGs, disaggregation of data can have many dimensions. Due to the wide use of administrative data in Danish statistical production, disaggregation by sex, age, income, education etc. does not pose difficulties. Disaggregation by religion on the other hand is more challenging, as it is not legal to ask persons about their religious beliefs. Disaggregation by people with disabilities is also a challenging aspect as data availability is limited and only in few cases we can achieve a fragmentary follow up on the global goals. Finally, geographical disaggregation. Here again, with data based on administrative registers, the challenges, at least relating to the social statistics, aren't substantial. However, due the size of population in Denmark, many of the indicators cannot be disaggregated by geographical levels, as this would interfere with statistical confidentiality.

Data from non-official sources – this is a heavily discussed topic in the statistical community. The first issue to be addressed in this context is what types of data we talk about. Generally, it could be dis-cussed whether it is data showing a 'section' of a bigger picture, such as Big Data, but also data from civil society could fall under this term. There is willingness from Statistics Denmark to receive and publish data from civil society, and there is willingness from civil society to deliver data. However, in the Danish case, data provided by civil society seldom fully corresponds to the requirements to the indicators and in consequence the indicator is only partially covered. The challenge is then not only to explain the figures that are not fully covering the indicator but also to complement them with additional data. Which again poses a question of comparability of data covering the indicator.

Coordinating data streams from the national statistical system – Statistics Denmark is coordinating national reporting on the global goals. Due to its complexity, the reporting is a challenging task. Firstly, the complex nature of many questionnaires makes it difficult to find relevant contributors. This implies a time consuming follow up with various governmental agencies. Furthermore, the coordination easily gets multi-layered – first a questionnaire comes in, then it is considered in Statistics Denmark, then it gets sent to a presumably right ministry. A frequent reply from the ministries is that it is not the right one or that it can only answer parts of the questionnaire. In such cases, the work starts from the beginning and the questionnaire can be repeatedly circulated among ministries, which requires some goodwill from the involved staff. From the NSO side, the challenge is to balance the perceivably 'over dimensioned' questionnaires with the integrity of the coordination process.

Finally, **compilation of national indicators for the national follow up on the SDG.** In this case, Statistics Denmark will coordinate a project on selection of national indicators. The first step will be a 'nationwide'

consultation in order to give a solid fundament for the work and ensuring that national indicators address issues that are identified as important by as wide group of stakeholders as possible. The list of the selected indicators will be presented to the Danish government as a possible input to both national baseline and national action plan for the SDGs.



The development of national reporting platform for global SDG indicators in Finland



Jukka Hoffrén
Statistics Finland

Abstract

The Finnish national reporting platform (NPR) of Agenda 2030 Sustainable Development Goal (SDG) –indicators was published in February 2019. At the moment database contains national data for 131 indicators of the total of internationally agreed set of 243 indicators i.e. some 53 per cent of the indicators. The national statistical office of Finland, Statistics Finland, is in charge of national compilation of global SDG -indicators.

The Finnish SDG development project for national compilation of SDG –indicator data was carried out during 2018 at statistics Finland. Project received the mandate and financing from the Prime Minister’s Office. The tasks of development project included two mapping rounds of possible indicator data producing institutions and organizations. This was followed by compilation of existing data, metadata and translation of indicator names into Finnish. An internet based national reporting platform was established and published on February 12, 2019. The Finnish NPR is internet –based database that is realized by Px-web software.

In order to extend and update the SDG -indicator data in NPR, Statistics Finland has set up a co-operation network of main data producers and Prime Minister’s office in April 2019. This network ensure and coordinate continued development efforts of the database. In total there are some 11 institutions in Finland that provide data for the database NPR. For the running activities statistical office has been issued additional allowance in state budget for next five years i.e. from 2019 to 2023. Also an option till 2030 have been preliminary agreed with the Ministry of Finance.

For statistical office SDG indicator compilation has been major challenge as it has brought up totally new data validation, methodological, conceptual, metadata and cooperation issues as well as methodological and metadata assistance to other national data producing institutions, not common tasks in common statistics compilation work. Within the development project several important lessons were learned. Firstly, continued political, institutional and financial support for the national production of the SDG indicator work is highly important. Secondly, communication and networking as well as interagency co-operation of experts play a major role in the compilation of NPR data. Thirdly, establishment of the SDG -indicator set requires that several methodological and conceptual challenges need to be addressed on national

level. Fourthly, the fact is that statistical data for only some 50 - 70 per cent of the indicators can presently be acquired although there is public and political pressure to raise the coverage up to 100 per cent.

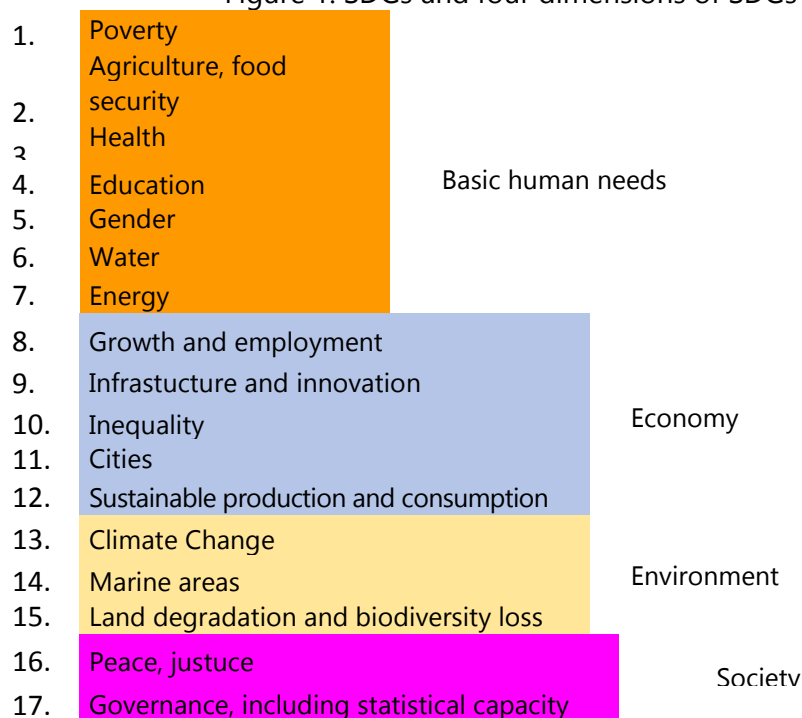
Keywords

Agenda 2030; SDGs; indicators; reporting platform; united nations

1. Introduction

United Nations' General Assembly (UNGA) adopted in September 2015 the "Transforming our world: the 2030 Agenda for Sustainable Development" document. The core of the 2030 Agenda is a list of 17 Sustainable Development Goals (SDGs) and 169 related targets to end poverty, protect the planet and ensure prosperity and peace. Governments are expected to take ownership and establish national frameworks for the achievement of the 17 SDGs. Monitoring of the SDGs is foreseen to take place at various levels – national, regional, global and thematic. The High-Level Political Forum (HLPF) is the United Nation's (UN) central platform to follow up and review the 2030 Agenda and the SDGs at the global level. UN member countries are encouraged to conduct voluntary national reviews (VNRs) of progress towards the SDGs. (Eurostat 2018) Figure 1 presents how 17 SDGs cover the four dimensions of sustainable development (Palm V. 2017).

Figure 1. SDGs and four dimensions of SDGs



The 2030 Agenda foresees establishing a set of global indicators to follow up and review the goals and targets. The Inter-Agency and Expert Group on SDG indicators (IAEG-SDG) was created to carry out this task, under supervision of the United Nations Statistical Commission (UNSC). UNGA adopted global indicator list of 232 different indicators in July 2017. These indicators cover all 169 targets of Agenda 2030. As some indicators are used to monitor more than one target, the list overall includes 244 indicators. Only 40 per cent of indicators are ready to use and classified as tier 1 by the UNSC. For further 31 per cent data are available for only limited number of countries worldwide and they are classified as tier 2. For the remaining part of the methodology still has to be agreed and they are classified as tier 3. Tier classifications are under constant reclassifications as the global statistical system evolves. The UNSC anticipates the possibility of yearly refinements to global indicator list, with two comprehensive reviews in 2020 and 2025. The IAEG-SDG is working to fully implement the global indicator list and to improve it further. This includes supervising the methodological work to develop tier 3 indicators and the extension of data coverage as well as identifying possible additional indicators to include in a comprehensive review of the indicator set in 2020. At national level each country is committed to producing their national data on these global UN indicators (Eurostat 2018)

Guidelines of Conference of European Statisticians (CES) to monitor progress towards SDGs and targets should be the result of close collaboration between statisticians and policy-makers. Statisticians should ensure that the monitoring of SDGs is consistent with relevant existing conceptual frameworks. NSOs play key role in measuring the achievement of SDGs and NSOs also serve as national focal points for statistics for SDGs as well as the national coordinating bodies on SDGs. NSOs also play a key role in providing data for global SDG indicators. (UNECE 2017)

According to CES recommendations national reporting platforms (NPR) for SDG indicators can have three components (i) a data collection or submission portal to different data providers, (ii) a production database and (iii) a dissemination portal for users containing tables, texts and publications. Countries should aim to present all SDG indicators available at the national level in their NPRs regardless of their data sources (official statistics as well as data from other data providers). Metadata on data sources should be presented together with data. Ensuring data validation and quality control are essential. In case of data from other sources NSOs do not have direct authority to apply quality assurance mechanisms, but NSOs should ask data providers to document data quality and the methods used to produce the data. (UNECE 2017)

2. Methodology

National governments play a key role in the implementation of the 2030 Agenda. In Finland national implementation is guided by Government Programme “Vision for Finland in 2025” and the Society’s commitment to Sustainable Development “The Finland we Want by 2050” constitutes Finland’s national interpretation of 2030 Agenda. These national goals are in line with global SDGs. Finland has also established a national set of sustainable development indicators to monitor progress towards SDGs although the data compilation for these indicators is only to be started in late 2019. Finland has already completed first VNR for HLPF in 2016 and second one is due in 2020. (Prime Minister’s Office 2017) As Finland has been one of the forerunners in implementing sustainable development principles, the Prime Minister’s Office was one of the key actors behind starting the compilation of Finnish NPR on global sustainable development.

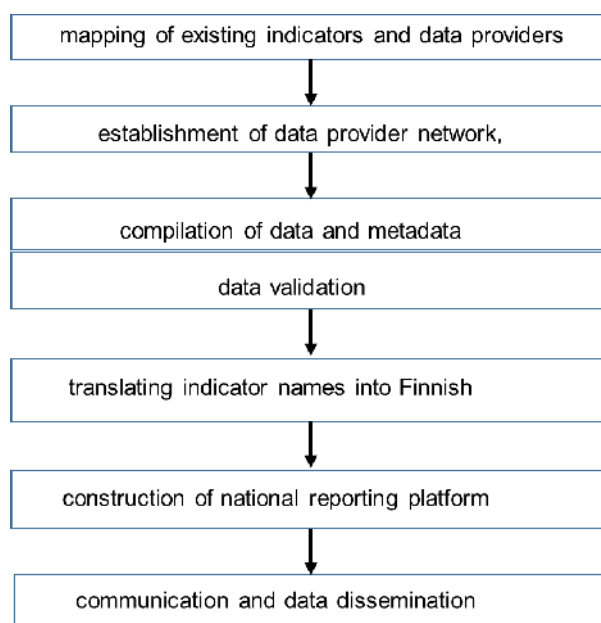
The Finnish SDG development project for national compilation of SDG – indicator data was carried out during 2018 at the National Statistical Office (NSO), Statistics Finland. Project received the mandate, steering and financing from the Prime Minister’s Office. The project employed two experts from NSO and the allowance for the tasks was 150,000 euros. Despite funding from PMO the experts worked independently following statistical codes of conduct and ethics.

The project methodology involved eight distinctive steps: 1. systematic mapping of indicators and data providers, 3. establishment of data provider network, 4. compilation of data and relevant metadata, 5. data validation, 6. translating indicator names into Finnish, 7. construction of national reporting platform as well as 8. communication and data dissemination. Figure 2 presents these seven phases as a waterfall project model utilised in the establishing NPR for SDG indicators.

The first task of development project included two mapping rounds of possible indicator data producing institutions and organizations. Project identified data providers that had already existing indicator data. Possible providers that had data under development or had not calculated the relevant indicator data, were excluded from the data providers’ list at this point. They will be revisited later. As a result an existing data providers and contact persons network was established. Next Statistics Finland delivered official request for SDG indicator data and relevant metadata to these organisations and contact persons. Replies for inquiry resulted not only official statistics data but also much data that had to be validated by Statistics Finland. This validation was quite general at this stage, but it included the assessments of relevance and correctness of data as well as assessments of methodological soundness.

Compilation of existing data and metadata was followed by translation of indicator names into Finnish. The translation was seen highly important as English language and terminology are common to Finnish people. Finally, an internet based national reporting platform was established and published in early 2019. (Statistics Finland 2019) The Finnish NPR is internet –based database was realized by Px-web software.

Figure 2. Seven phases of establishing Finnish NPR for SDG -indicators



3. Results

During the project it was possible to define responsible provider organisation for over 90 per cent of the SDG indicators. The biggest provider was Statistics Finland responsible for 85 indicators of which data is available for 59 indicators. The Finnish Environment Institute could provide 12 out of 24 indicators, Ministry of Foreign Affairs 9 out of 12, National Institute for Health and Welfare 6 out of 12, Natural Resource Institute 1 out of 9 and Ministry of Social Affairs and Health 2 out of 4 indicators. Other important providers include The Finnish National Agency for Education, Prime Minister’s Office, Finnish Immigration Service, Finnish association to promote sustainable business and Ministry for Agriculture and Forestry.

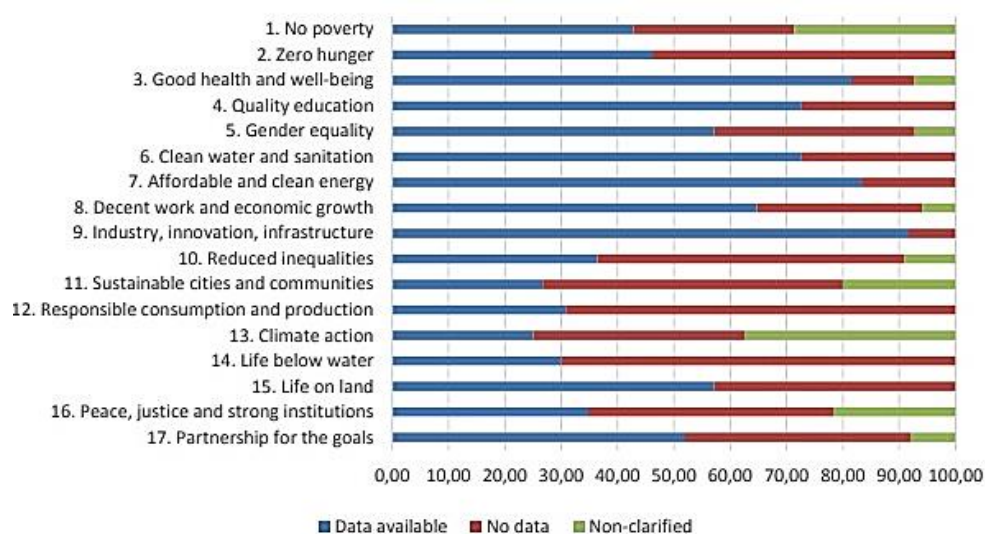
The Finnish national reporting platform (NPR) of Agenda 2030 Sustainable Development Goal (SDG) –indicators was published on February 12th, 2019. At the moment database contains national data for 131 global indicators of the total of internationally agreed set of 243 indicators i.e. some 53 per cent of the indicators of which 40 per cent came from domestic providers and 10

per cent from databases of international organisations. Indicators time-series should ideally range from 1990 to 2017, but in many cases only data for some recent years is currently available. Currently data for 91 indicators is not currently available and definitions for 22 indicators are not yet clarified. The SDG indicator set is currently available Finnish, Swedish and English. Table 1 summarises the current data availability of SDG indicators in Finnish NPR.

The national statistical office, Statistics Finland, is currently in charge of national compilation of global SDG -indicators. In order to extend and update the SDG -indicator data in NPR, Statistics Finland set up a co-operation network of main data producers and Prime Minister's Office (PMO) in April-May 2019. This network ensures and coordinates continued development efforts of the database. In total there are some 11 institutions in Finland that provide data for the database NPR. For the running activities statistical office has been issued additional allowance in state budget for next five years i.e. from 2019 to 2023. Also an option till 2030 have been preliminary agreed with the Ministry of Finance.

In future Statistics Finland will continue national collection, update and publish of SDG indicator data. Statistics Finland also reports follow-up indicator data concerning Finland in international inquiries. On methodological side, Statistics Finland also continues to supplement the existing data by exploring new data sources and proxy indicators. Also possibilities to replace data from international organisations by domestic providers' data is under examination. International development work is also made to complement the data sources. Future challenges include further disaggregation by gender, age, disability and geographical area.

Table 1. Current SDG data availability in Finnish National Reporting platform



4. Discussion and Conclusion

The SDGs are ambitious step for NSOs as the SDG indicator set takes a much broader view of progress and state of sustainability than earlier. The practical challenges of data compilation require previously unseen collaboration of NSOs with diversity of actors. Instead of just producing new social, economic and environmental statistics also more intense conceptual and methodological work for SDG indicator framework is required from NSOs. One vital task for NSOs is the validation of SDG data provided by non-statistical government agencies, research institutes, private companies and NGOs. Development of common and unified codes of practices are highly important to ensure data quality and comparability. Assistance to other national data producing institutions as well as international support, assistance and statistical capacity building are emerging new areas

In Finland within the development project several important lessons were learned. These are traditionally not common tasks for NSOs. Instead these steps demand new kind of approaches and require a lot new kind capabilities and skills from statisticians. The Finnish findings can be summarised as follows.

- i. Firstly, continued political, institutional and financial support of prime ministers' office for the national production of the SDG indicator work was highly important. The national compilation of data for SDG indicator set requires additional resources and high level support that justifies the priority of SDG indicator work.
- ii. Secondly, collaboration and networking as well as interagency expert co-operation played a major role in the compilation of NPR data. Beside high level political support open and confidential cooperation between statisticians and other experts is crucial. This collaboration relates especially to the indicator concepts and data classifications, definitions and compilation methodologies.
- iii. Thirdly, establishment of the SDG -indicator set required that several methodological and conceptual challenges needed to be addressed on national level. In Finland the statistics are compiled in accordance with ESS classifications that occasionally differ from those implied by UN. One important challenge was the validation process to ensure data quality of non-NSO data.
- iv. Fourthly, maximum statistical data that can be acquired is currently only some 50 - 70 per cent of the indicators. SDG indicator framework hints that data should be already available since it has been included to the indicator set by the UN. The absence of large share data poses normative political and public pressures on NSO to increase the data coverage. In Finland PMO set a target to raise the coverage up to 100 per cent.

Future challenges include further disaggregation of data as the collected data in most cases is optimised to minimum samples because of cost-savings and to minimise response burden. Thus, in many cases current data does not necessary allow the in-depth disaggregation recently emphasised by UN organisations. As the SDG indicator set disaggregation seems to deepen over time, the basic statistical data collection samples have to be reconsidered. Development of data for indicators currently undefined, with non-existent data or in need of development are major challenges. In some cases, indicator data could be calculated from existing data bases with considerable additional work. In most cases considerable amount of resources and methodological assistance from NSO are needed.

For policy makers to SDGs embrace progress that is to be managed to achieve multiple objectives. Consequently, NSOs are in demand to produce assessments of progress towards SDGs indicator by indicator as they owe expertise on the indicators. However, political and normative goals are in many cases non-existent and thus NSOs face a difficult task. There also exists even more demanding future necessities to develop aggregate indicators which evaluate the relative contribution of each SDG and their interaction with each other, in order to evaluate general progress and to provide more clear insights and proposals for social reforms and changes.

References

1. Eurostat. Sustainable in the European Union 2018. Monitoring report on progress towards SDGs in an EU context. 2018 edition. European Union: Luxembourg. Printed in Belgium.
2. Palm, Viveka 2017. The response of official statistics to Sustainable Development Goals as seen from the process in the IAEG-SDG. Statistics Sweden.
3. Prime Minister's Office 2017. Government Report on the implementation of the 2030 Agenda for sustainable Development. Sustainable development in Finland – long-term, Coherent and Inclusive Action. Prime minister's Office Publications 11/2017. Helsinki.
4. Statistics Finland 2019. UN indicators for sustainable development (Agenda 2030). http://www.stat.fi/tup/kestavan-kehityksen-yk-indikaattorit-agenda2030_en.html [15 Apr 2019]
5. UNECE 2017. Conference of European Statisticians. Road map on Statistics for Sustainable Development. Geneva: United Nations.



Coordinated communication for better follow-up of the 2030 Agenda in Sweden



Sara Frankl, Karin Hansson
Stockholm, Sweden

Abstract

The purpose of the 2030 Agenda is to achieve transformation into a socially, environmentally and economically sustainable society in all countries and at all levels by the year 2030. It is a long-term social change, described in the Agenda as 17 goals and 169 sub-goals.

- **Communication as a strategic tool**

This major transformation in society requires a change in both approach and behavior, which, in turn, requires knowledge and commitment to sustainable development. One strategic tool to achieve this is communication that informs, engages and leads to action. Communication is thus a key component in Sweden's implementation of the 2030 Agenda.

- **Statistics Sweden's role: coordinator of the communication**

The Swedish government has commissioned Statistics Sweden with coordinating the development, production and availability of the statistical follow-up in Sweden, on both the global and national level. As the coordinating agency, Statistics Sweden's role is to make sure that the communicative efforts that are linked to the 2030 Agenda must be aimed at contributing to the goal of a sustainable society. Furthermore, the communication on how Sweden lives up to the goals and targets of the Agenda by providing data and statistics that are trustworthy and objective to support evidence based decision making and to facilitate accountability.

- **A complex change process with many actors**

The breadth and complexity of the Agenda means that many different producers of statistics and data will need to contribute to the statistical follow-up. These may include organizations within the Swedish system of official statistics as well as other authorities and civil society organizations. Statistics Sweden has proposed developing a specific web page where the statistical follow-up of the Agenda can be gathered and displayed. In addition to this, the intended audience of the communication is very diverse. Basically, it consists of all those actors that need to be involved in achieving the goals of the 2030 Agenda: civil servants, politicians, actors within the civil society, businesses, and the general public.

- **First results**

So far, we have accomplished three interesting results. First, Statistics Sweden has proposed a list of indicators that has attracted a lot of attention and we have had several opportunities to present it in different contexts. Introducing this proposal to various audiences is a great way to start communicating about the 2030 agenda.

Second, we have produced films about Sweden's work with two of the Agenda goals: quality education for all and responsible consumption and production, as well as teaching material to go with them. Third, we have published web articles on our web site, aimed at non-expert readers.

- **This presentation**

The Swedish presentation during the session will focus on how the work on the statistical follow-up and digital solutions for its communication continues in Sweden. It will give an account of challenges, implemented and planned communication activities and the coordination and collaboration efforts that underpin results.

Keywords

SDG; global goals; collaboration; communication strategy

1. Introduction

The aim of this paper is to present aspects of communication such as challenges, underlying conditions and opportunities present in Statistics Sweden's work with coordinating the statistical follow-up of the 2030 Agenda in Sweden. The complexity and the multiple actors that need to be involved in the transformation poses many challenges to the implementation as well as to the follow-up and communication of results.

The purpose of the 2030 Agenda is to achieve transformation into a socially, environmentally and economically sustainable society in all countries and at all levels by the year 2030. It is a long-term social change, described in the Agenda as 17 goals and 169 sub-goals.

This major transformation in society requires a change in both approach and behaviour, which, in turn, requires knowledge and commitment to sustainable development. One strategic tool to achieve this is communication that informs, engages and leads to action. Thus, communication is a key component also in the statistical follow-up of the 2030 Agenda.

The agenda is ambitious but also somewhat abstract in its goals; carefully planned and executed follow-up helps to make it more concrete and relevant for the inhabitants in Sweden. However, follow-up is effective only when it is communicated in a clear way. This is Statistics Sweden's main task. By showing how Sweden is doing, we can make people aware of what sustainable

development entails and what is being done to fulfil the sustainable development goals. This is necessary to keep engagement high on all levels of Swedish society.

At Statistics Sweden, our aim is to create a follow-up that spans many sectors and subject areas, with focus on synergies. We need also to highlight challenges that cannot be solved in a particular sector alone and show where collaboration is necessary.

2. Methodology: Our work so far and plans for next steps

The Swedish government has commissioned Statistics Sweden with coordinating the development, production and availability of the statistical follow-up in Sweden. As the coordinating agency, Statistics Sweden's role is to make sure that the communicative efforts linked to the 2030 Agenda are aimed at contributing to the goal of a sustainable society. A central starting point is that the statistical follow-up needs to be an integrated part of the overall follow-up. This is illustrated in figure 1.

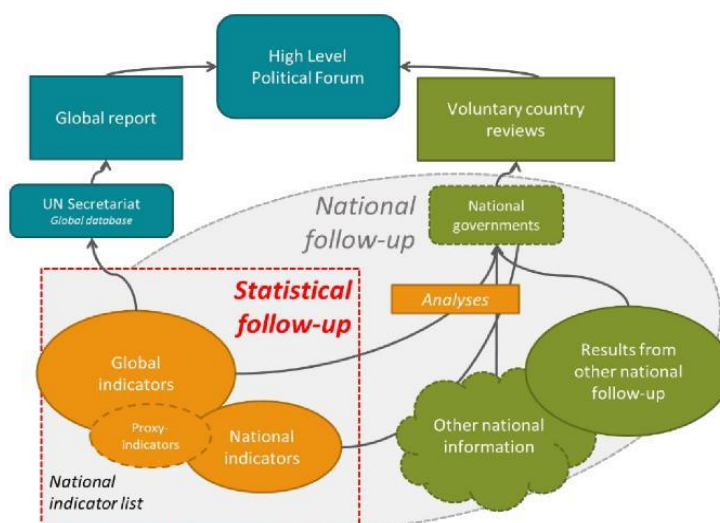


Figure 1. Overview of the national follow-up of the 2030 Agenda in Sweden

How do you set up the statistical follow-up of an agenda that requires such a thoroughgoing change and spans so many areas and aspects of society? An additional challenge for us is this: The government assignment was in many ways unclear and purposefully vague to leave room for adapting the ambitions to the fact that no additional funding was provided.

From the start, it was clear to us that there are many possible ways of implementing the tasks. A very narrow interpretation of coordination could mean only keeping track of who provides data and statistics to the global level of the follow-up. At the other extreme of the scale, a broad interpretation could

mean setting up a fully developed national coordination system with a nationally developed indicator framework, adapted to the national context. This framework could also include an elaborate digital data and communication platform, allowing for automation of global data flows and interactive use of quality checked data from multiple producers, inside and outside of the national statistical system.

Statistics Sweden has applied a step by step approach aiming for the more elaborate coordination and communication but by necessity allowing for temporary, ad hoc solutions and for allowing the development of the statistical follow-up to take time.

2.1 National list of indicators and base line report

The first step of our work was to create a national list of indicators for the follow-up. We decided early on that we needed to do this in broad consultation with stakeholders, for example, actors in the statistical system, the potential data producers, and users in the form of government agencies and civil society organisations. We sought input primarily on identifying domestic indicators needed to complement the global ones to adapt the follow-up to a national context.

The final list contains the global indicators and is complemented with 50+ domestic indicators. Statistics Sweden has also produced a baseline report on how Sweden lives up to the goals and targets of the agenda.

The base line report and list of indicators have received a lot of attention and we have had several opportunities to present it in different contexts. We have found that introducing the contents of these documents to various audiences is a great way to start communicating about the 2030 agenda.

2.2 Material for schools and the general public

The base line report and list of national indicators were intended mainly for specialised users and the stakeholders involved in the process. However, our communication must also reach non-experts. Therefore, we have produced material for schools and the general public. So far, we have made two films and teaching material to go with them about statistics related to two of the Agenda goals: quality education for all and responsible consumption and production. We have also published a number of articles on our web site about national and international efforts relating to statistics for the global goals.

2.3 Collaboration with the commission for the 2030 Agenda for sustainable development

The commission for the 2030 Agenda for sustainable development was established in March 2016 and in March 2019 it delivered its final report. This national committee was set up to support and stimulate the work on

implementing the Agenda in Sweden. The commission's task included to propose an overarching action plan for the government, to promote sharing of information and knowledge and to secure support for the implementation of the 2030 Agenda through a broad dialogue with various actors in the society.

Statistics Sweden has had ongoing dialogue with the commission during its work. One of our focuses was on communication. The first step was to establish a common understanding on overlaps, synergies and boundaries of our respective tasks. The conclusion of this was the common view that the follow-up must consist of more than the statistical indicators and statistical analyses. Statistics Sweden and the commission agreed that the follow-up needs to include more qualitative information, for example, on actions and implementation efforts across society, as well as results from other, more sectoral follow-up, statistical and otherwise.

2.4 Next step: focus on strategic communication

To date, our communication work has been mainly on an ad hoc basis. The next step is to plan for more strategic communication, aiming specifically to support change by informing and creating engagement. We also plan to continuously monitor the effects of our efforts and make necessary adjustments. In our role as coordinator, we need to make sure that the communication contributes to Sweden's transformation towards the 2030 Agenda the way intended. The strategic communication plan will be developed in close collaboration between statisticians and communication specialists. The plan will include, among other things:

- Options for digital data platform solutions with different levels of ambition
- Analyses of important user groups
- Visibility in the media
- Social media presence (Instagram, Facebook, Twitter)
- Statistics Sweden's presence at national, regional and local events with links to the 2030 Agenda and sustainable development.

3. Results

As for the results of the communication work, it is of course too early to measure effects on behaviour. However, there are two important results related to the indicator list and the base line report.

One of the main components of the coordination of the follow-up and the communication work around is involving stakeholders. As already mentioned, the breadth and complexity of the Agenda means that many different producers of statistics and data will need to contribute to the statistical follow-up. The broad involvement in the development of the follow-up is vital for the

success of Statistics Sweden's assignment as a means to ensure wide use and create trust and engagement.

One of the main components of the coordination of the follow-up and the communication work around is involving stakeholders. As already mentioned, the breadth and complexity of the Agenda means that many different producers of statistics and data will need to contribute to the statistical follow-up. The broad involvement in the development of the follow-up is vital for the success of Statistics Sweden's assignment as a means to ensure wide use and create trust and engagement.

The first tangible result of the broad collaboration deployed in the production of the first baseline report and indicator list is the interest it sparked among a wide variety of actors. The interest is most notable in the amount of requests we have received to present the results in different contexts. Statistics Sweden's experts have also been asked to do interviews in sector media channels as well as in broad publicly publicised newspapers and radio.

The second tangible result of the efforts is the far reaching use of the content of the base line report in the Sweden's national voluntary report to the High-level Political Forum on Sustainable Development in July 2017.

4. Discussion and Conclusion

During the course of our work with the assignment to coordinate the statistical follow-up we have sought to engage many actors within and outside of the statistical system. This has created high expectations that the follow-up will cover different perspectives, most notably from the other producers of statistics and data and from civil society organisations. Naturally, we are pleased that the output of the work seems to have contributed to engagement. However, coupled with the lack of resources, this also poses a real challenge for Statistics Sweden as the coordinator of the process.

- **Step-by-step approach**

With scarce resources it is of course important to focus on the right things. Therefore, we would like to stress the importance of continuous coordination and collaboration with a wide range of actors for the follow-up to develop in a successful direction. This is what we mean by a step-by-step approach. For example, it is important that the actors know that we always consider their input even if it does not show in the tangible output immediately. We also invite and expect stakeholders to take part in the development of the processes, as well as in the statistical production of the indicator used in the follow-up.

- **Diverse target group calls for broad involvement**

Another challenge is that the intended audience of the follow-up and is very diverse. Basically, the target group consists of all those actors that need to be involved in achieving the goals of the 2030 Agenda: civil servants, politicians, civil society organisations, businesses, and the general public. This further stresses the importance of broad involvement.

A key factor to succeed in informing and sparking engagement is the close collaboration between statisticians and communication specialists at Statistics Sweden. This way, we can make sure to adapt the communication to different user groups and for a wide target audience.

- **Engage with the users**

Another important aspect is user engagement. In the step by step approach that Statistics Sweden applies, we have been able to use the fact that the Swedish statistical system is highly decentralised to our advantage. One effect of the decentralisation is that producers and users of official statistics often reside within the same government agencies and that statistics producers often have a thorough understanding of the subject matter areas they are involved in.

Looking ahead Statistics Sweden plans to invite the most important users to work together with representatives from the production side to develop the first follow-up of the baseline report. This work is planned for October 2019. Our intention is to develop this interaction further and broaden the scope to other user groups in the future.

5. Conclusions

In conclusion, we see great benefits with also coordinating the communication as part of our assignment to coordinate the follow-up of the 2030 Agenda. Only then can our communication work contribute to Sweden's actions towards reaching the targets and goals of the 2030 Agenda. In our experience so far, some important requirements for creating well-coordinated, effective communication are:

- data and statistics that underpin evidence based decision making *but also* supports and facilitates actors possibilities to exact accountability;
- trustworthy and objective data and statistics;
- broad involvement across different actors and levels of society;
- collaboration between professionals in different areas such as statisticians and communication specialists.



The need to develop labour accounts in Malaysia: An assessment



Siti Asiah Ahmad, Noraliza Mohamad Ali, Nur Layali Mohd Ali Khan, Nur
Hurriyatul Huda Abdullah San, Nurfarahin Harun
Department of Statistics Malaysia

Abstract

Labour statistics in Malaysia are available from various sources. These statistics are compiled using different approaches, at different intervals or frequencies and at times adopting different concepts and definitions. This is due to the difference in objectives and purposes of the data collection. Often, these statistics can be found scattered across agencies. The presence of these statistics offers opportunity for Malaysia to develop labour account. The statistics can provide information on the transition from employment, unemployment and inactivity in four quadrants which presents statistics on jobs, persons, volume and payments. Therefore, this paper aimed to study the feasibility to develop labour accounts in Malaysia as well as to study all the possible data sources needed to materialize the accounts. The Malaysian Labour Account results are important for statistical offices and also data users because with labour account, contradictory result between data sources can be eliminated and a comprehensive overview of the situation on labour market is possible.

Keywords

Labour account, jobs, persons, volume, payments

1. Introduction

Labour is an important factor for economic growth and the country development. In Malaysia, labour supply is measured based on Labour Force Survey (LFS) conducted by Department of Statistics Malaysia (DOSM) since 1982. This survey used household approach. Besides household approach, DOSM also conducted censuses and surveys on establishment which provide labour demands statistics among others Economic Census (EC), Quarterly Employment Survey (QES) and Annual Economic Survey (AES). Labour market information are also available in other agencies for example, Ministry of Human Resource (MOHR) and Malaysian Employers Federation (MEF).

Consolidation and compilation of labour statistics in Malaysia is quite challenging considering various data sources. Among the challenges faced in consolidating and compiling labour statistics are coverage, concept used and reference period. Thus, it is important for Malaysia to harmonize labour market information through development of labour account.

Labour Account was brought up as early as 1980s, as a result of discussions at international level concerning the linkages of labour statistics with other areas of statistics namely social and demographic statistics and economic statistics as organized by the System of National Accounts (SNA) (Hoffmann, 2000). This was developed into logical framework for obtaining the key labour market variables during 15th International Conference of Labour Statisticians (ICLS) in 1993 (Buhmann et. al, 2002). In addition, labour account provides a time series of estimates of the number of employed persons, number of jobs, hours worked and income earned for each industry in one coherent framework.

Several countries are at different stages of developing labour account such as Australia, Denmark, the Netherland and Switzerland. Australia has benefited from the development of labour account for instance, it has resulted in consistent estimates of key labour market variables over time. For the first time, statistics on total number of employed persons for Australia is available from the industry perspective in a time series. This could be used to better assess policy changes targeting a particular industry, providing a more complete picture of the number of people impacted by the change (Australian Labour Account Fact Sheet 6, 2018).

In order to ensure that a country is able to manage labour supply efficiently, countries should carefully measure the labour market situation through reliable data. Thus, considering all benefits mentioned above, Malaysia is planning to develop labour account based on data collection by DOSM and others agencies as well as from administrative records. Labour account offers a framework to bring together Malaysia's labour market data from multiple statistical sources into coherent and consistent set of labour statistics which consist of four components namely filled job, employment, labour cost and hours of works.

2. Methodology

The ILO lists six key elements in labour statistics which are i) employed persons and jobs; ii) unemployed and underemployed persons; iii) job vacancies; iv) hours of work and full-time equivalents; v) income from employment and labour costs; and vi) organisation of the labour market i.e. statistics on collective labour agreements, industrial disputes and trade-union memberships (Australian Labour Account, 2018).

At present there is no international standard for developing labour account. However, the ILO has documented a guideline comprising four basic steps in the development of labour account involved statistical integration. First, is definition of the model and the identity equation. Second, is harmonization of definitions and classifications in source statistics,

achievement of full coverage. Third, is minimization of measurement errors and the final steps is balancing.

ILO describes two approaches in compiling labour account which are cross-sectional and longitudinal. Cross-sectional approach involves confrontation and reconciliation of key labour market measure. Longitudinal approach incorporates changes to population and labour force via births, deaths and net migration (Buhmann et. al, 2002).

Australia, Denmark, the Netherland and Switzerland have followed these approaches in developing a labour account system in their respective countries. Australia and the Netherland adopt cross-sectional approach while Denmark and Switzerland adopt both cross-sectional and longitudinal approaches

Australia

The Australian Labour Account framework has been designed aligns with United Nations System of National Accounts as applied in Australian System of National Accounts (ASNA). This framework consists of four quadrants: Jobs, Persons, Volume and Payments that covers all types of employment including employees, self-employed and contributing family workers (Australian Labour Account, 2018). The main identity relationship can be defined as follows:

Table 1: Identity relationship of Australian Labour Account

Jobs quadrant Total jobs Filled jobs	Filled jobs + Job Vacancies Number of main jobs + Number of secondary jobs
Persons quadrant Labour force Employed person Underutilised persons	Employed persons + Unemployed persons No. of main jobs (total economy level) Unemployed persons + Underemployed persons
Volume quadrant Available hours of labour supply Average hours worked per job Hours sought but not worked Hours paid for	Hours actually worked + Hours sought but not worked Hours actually worked / Filled jobs Hours sought by unemployed + Additional hours sought by underemployed Ordinary time hours paid for + Overtime hours paid for
Payments quadrant Total labour cost Total labour income Average labour income per employed persons Average cost per hour worked / Average cost per hour paid	Total labour income + Employment related costs + Payroll tax – Employment Subsidies Compensation of employees + Labour income from self-employment Total labour income / Employed persons Total labour cost / (Hours worked / Hours paid)

However, Australia is facing limitation in terms of conceptual, content, data source, methodologies and other quality limitation includes timeliness, data availability and accuracy during development of labour account (Australian Labour Account, 2018). For example, due to content limitation, analysis of jobs, persons, hours and payments by gender or age does not include in Australian Labour Account.

Denmark

Currently Denmark has two separate accounts called Labour Market Account (LMA) and Working Time Accounts (WTA). LMA provides a complete overview of population's labour market status while WTA provide time series on jobs, employment, hours worked and compensation of employees. The long-term goal is to merge WTA and LMA into one micro based system.

Through LMA, Denmark able to compile the population labour market data in terms of full-time persons i.e. person who work for 37 hours per week. Data sources for LMA including e-income register, central business register, income statistics, statistics on persons receiving social benefits and the central population register. The main objective of LMA is to construct a micro-register covering the population (Statistics Denmark, 2015).

WTA published figures on employment, jobs, hours of worked, full-time equivalents and compensation of employees according to SNA definitions. Some sources that was used for development this account is based on the Register of Employment Statistics, Statistics on Earnings, Labour Force Survey and Labour Cost Survey. Table 2 shows the main relational equations for employees that satisfied the WTA (Buhmann et al., 2002).

Table 2: Identity relationship of Danish Working Time Accounts

Jobs	Employed persons – Employed persons on leave + Secondary jobs
Total hours of work	Number of jobs * Actual hours per job
Total compensation	Number of jobs * Compensation per job

The estimate on the total hours of works was based on the average number of jobs from the Register of Employment Statistics multiplied by data by wages statistics on total hours worked per job. The model for self-employed in Danish WTA is simpler because it only covers jobs, employment and actual hours worked.

The Netherland

In the Netherland, there still distinction in employment data between pre-revision data (up to 1997) and post-revision data (from 1995 onwards). Total of employment in post-revision data can be presented by gender, status of employment and industry, while pre-revision data which is a breakdown by

weekly hours worked can be presented in terms of persons employed by occupation, educational level and nationality of origin (Buhmann et al., 2002). The identity relationship underlying the Dutch Labour Accounts as shown in Table 3.

Table 3: Identity relationship of Dutch Labour Account

Employee labour	
Jobs	Persons employed + Secondary jobs
Annual hours per job	Weekly hours per job * Number of weeks – Holiday leave – Festivity leave
Total hours	Jobs * Annual hours per job
Full-time equivalents	Total / Annual hours per full-time job
Self-employed labour	
Jobs	Persons employed + Secondary jobs
Total hours per week	Jobs * Weekly hours per job
Full-time equivalent	Total hours per week / Weekly hours per full-time job
Various concepts of hours	
Hours paid	Contractual hours + Hours overtime
Hours actually worked	Hours paid – Hours paid but not worked
Labour cost and labour income	
Labour cost	Hours actually worked * Labour cost per hour
Employee cost	Hour actually worked * Employee cost per hour
Hour actually worked	Jobs * Hours worked per job
Hours paid	Jobs * Hours paid per job
Earnings	Hours paid * Hourly earnings
Earnings	Jobs * Annual earnings
Annual earnings	Hours paid per job * Hourly earning

The relation between employed persons and jobs are made between data from household survey, both persons employed and jobs while establishment survey for jobs. Annual earnings data are available in Annual Survey on Employment and Earnings while for wages data are available in social security files. The number of jobs could be higher than number of employed person because employed persons can have more than one job at a time.

Switzerland

Switzerland has two different accounts, Labour Market Indicator (LMI) and Labour Accounts. LMI followed cross-sectional approach while Labour Accounts followed longitudinal approach. LMI used four different sources of employment that is Population Census, Swiss Labour Force Survey, Establishment Survey and Statistics on Jobs (JOBSTA). All four sources use slightly different definitions of employment and harmonization between these

sources need to be done according to common definition. In Swiss Labour Accounts, the priority identity relationship are as follows:

Table 4: Identity relationship of Swiss Labour Account

Non-monetary dimension	
Total population	Persons employed + Persons unemployed + Persons outside labour force
Jobs	Persons in employment + Vacancies
Actual hours of work	Actual volume of work / Average annual hours worked in full-time job
Full-time equivalents	Actual hours of work
Paid volume of work	Actual hours of work of person – Unpaid actual hours of person + Absences from work of person
Monetary dimension	
Labour costs sum	Paid volume of work * Earning per year (employees and self-employed)
Labour costs sum	Earning per year (employees and self-employed) + Social contributions + Other labour costs

Both persons employed and jobs are linked between data from household and establishment surveys as well as the relationship between the number of employed persons, hours worked and full-time equivalents. Gross changes in the population such as immigration, emigration, death and births also are taken into consideration. The monetary dimension likes earning and wages still not in part of Swiss Labour Account (Buhmann et al., 2002).

Australia, Denmark, the Netherland and Switzerland agreed on the main principle of developing labour account as it gives them new opportunities to complement, present and improve existing labour statistics (Buhmann et al., 2002). At the first stage of development, these four countries have to deal with the challenges in terms of differences results between sources, unclear overview of the labour market situation due to various sources. The cut-off time difference between the surveys shows that the labour market situation may vary and the relationship between labour data with other statistical system are not clear especially in relation to population and education statistics.

In this early stage of developing Malaysian Labour Account, standards and guidelines of ILO and System of National Account (SNA) will be referred as well as the four countries experienced. On top of that, Malaysia also need to consider the availability of complete and readiness of labour data as a base for the development of labour account. Malaysia also has to face the same issues and challenges faced by these four countries.

The source from Table 5 shows different indicators, approaches and frequencies due to its different objectives and purposes of data collection. The

presence of these data gives Malaysia an opportunity to develop labour account. The data from different sources will be harmonized and adjusted for minimizing the errors through integration process.

Table 5: Malaysia Labour Force Information Sources

Sources/Frequency	Labour force Employed	Unemployed	Underemployed	Main jobs	Secondary jobs	Job vacancies	Filled jobs	Hours actually worked	Hours paid for	Overtime hours paid	Total labour income	Employment subsidies	Compensation of -----	Employment related -----	Payroll tax
Household Approach															
Labour Force Survey (<i>Monthly, Annually</i>)	✓	✓	✓	✓	✓	✓	✓	✓							
Salaries and Wages Survey (<i>Monthly, Annually</i>)								✓	✓	✓	✓	✓			
Informal sector workforce survey (Biennially)		✓													
Household Income Survey (<i>Twice in 5 years</i>)		✓		✓			✓								
Establishment Approach															
Economic Census (<i>5 years</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census of Distributive Trade (<i>4 years</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Annual Economic Survey (<i>Annually</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Quarterly Employment Statistics (<i>Quarterly</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Quarterly Services Survey (<i>Quarterly</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Monthly Manufacturing Survey (<i>Monthly</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Monthly Distributive Trade (<i>Monthly</i>)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Derived Statistics															
Labour Productivity (<i>Quarterly</i>)		✓						✓							
GDP Income Approach (<i>Annually</i>)													✓		
Administrative Approach															
Malaysian Employer Federation (MEF)								✓			✓	✓	✓		
Employees Provident Fund (EPF)		✓					✓						✓		
Social Security Organisation (SOCSSO)		✓					✓						✓		
Inland Revenue Board of Malaysia (LHDN)		✓					✓				✓				✓
JobsMalaysia and Jobstreet		✓	✓				✓	✓							

Malaysia is planning to develop labour account from the sources available in Table 5. Selections of data which present the primary source need to be made for each component. For the combination of data sources, the definitions, reference periods and frequencies need to be harmonized to a common definition based on ILO definition.

3. Results and Findings

Total employment defined as total persons engaged in production. Data of employed persons, unemployed person and persons outside labour force are available in LFS which may consider as primary source in persons component. However, comparison also can be made between data sources available. For example, in EC, breakdowns of employed persons are working proprietors, unpaid family workers, full-time and part-time paid employees whereas in LFS the breakdown of employed persons consist of employer, employee (private and government), own account workers and unpaid family workers. On the other hand, the elements of persons employed can be integrated through harmonization of the definitions and the classifications into common definitions as recommended by ILO.

Jobs component consist of number of filled jobs and jobs vacancies. Number of employed persons from persons quadrant also can be referred as number of filled jobs as a job is a position held by a person that involves work and responsibility. Jobs component may have combination of primary data sources as QES represent private data while Public Service Department provide public service information. The number of jobs could be higher than the number of employed persons as one person can have more than one job. Filled jobs compose of main jobs and secondary jobs which are available from LFS.

Total hours worked refer to hours actually worked by persons in employment. Statistics of hours actually worked should include normal periods of work, overtime, time spent at the place of work on work, time spent at the place of work waiting or standing-by and time corresponding to short rest periods at the workplace. Nevertheless, hours paid for but not worked, meal breaks longer than 30 minutes and time spent on travel from home to work and vice versa need to exclude from total hours actually worked. Total hours actually worked can be done by harmonizing the two main sources in developing these account that is LFS and EC and also can take into consideration data from Labour Productivity Report.

Data on payment of labour which includes total labour income, compensation of employees, employees' subsidies and others related to labour payments are widely being used from the administrative data which collect and manage the information of labour in Malaysia. Besides that, data from Salaries and Wages Survey that consists of cash, payment in kind and

overtime payment breakdown can be harmonized with data from EC which breaks into salaries and wages includes overtime, allowances, bonuses, commission and also director's salaries, allowances, fees, bonuses and commissions. Data on earnings from Household Income Survey (HIS) that consist of wages and salaries, allowances, bonuses, other cash (e.g.: commissions, overtime), free/concessional food, free/concessional lodging, free/concessional consumer goods and services, other payments in kind received and employer's contribution also can be served as primary source for payments element. Besides, data for employment costs are also available from EC and other government agencies.

For now, Malaysia still in the process of strengthening the Labour Market Information (LMI) using administrative records from various agencies and sources. Through this process, Malaysia would be able to obtain real time LMI and lead to granular and timely data. Reference from the other countries like Australia, Denmark, the Netherland and Switzerland are taken into consideration when developing the labour accounts even the developments of each country are at different stages. Based on the sources of data that Malaysia have and the identity relationship that were used by Australia, Denmark, the Netherland and Switzerland, the components in developing the labour account can be grouping into four main quadrants that is persons, jobs, payments and volume. Figure 1 shows the fundamental components in each quadrant.

Figure 1: Four main quadrants in Malaysian Labour Accounts

PERSONS	JOB
Labour force Employed Unemployed Underemployed Underutilised	Main jobs Secondary jobs Job vacancies Filled jobs
PAYMENTS	VOLUME
Total labour cost Employment related costs Payroll tax Compensation of employees Employment subsidies Total labour income Labour income from self-employed	Hours actually worked Hours paid for Overtime hours paid for Ordinary time hours paid for Hours sought but not worked Hours sought by unemployed Additional hours sought by underemployed

Each component has data available from different sources and approaches. Selection of data need to make first in which sources serve as primary sources or combination of sources also can be appropriate with the harmonization

procedure of their definition and classification. Malaysia is planning to begin with jobs' quadrant as the concept of jobs is the central of the framework as mentioned during 15th ICLS. This will follow by persons' quadrant which data on employment are available from various sources including time series data. Moreover, employed variable in persons' quadrant were largely used in the relationship with the other quadrants.

According to Table 5, employed variables are widely available in Malaysia labour force sources information. Besides, the components in persons' and jobs' quadrant showed the most completed data that available from household, establishment and administrative data. The primary source for each component needs to be selected or the combination of the sources of data also can be done based on ILO guidelines and standards.

4. Discussion and Conclusion

Labour account provides a time series of estimates of the number of employed persons, the number of jobs, hours worked and the income earned for each industry in one coherent framework. It helps in identifying the inconsistency and gaps in the process of developing labour account. The statistics obtained from the development of labour account should provide a reliable information which mostly being used by people engaged in the use of labour statistics in macro-economic analysis, forecasting and in policy related research.

Consolidation and compilation of labour statistics in Malaysia is quite challenging considering various data sources. The data sources available consist of primary and secondary data. Primary data was collected through census and surveys while secondary data is data that readily available and recorded. These two types of data have different intentions and purpose of collection where primary data are more reliable than secondary data. Furthermore, the approach of every census and surveys also different, household approach and establishment approach. For example, household approach covers entire labour force including formal and informal sector while establishment approach only covers formal sector which is employees in registered company.

The reference period for each source are different. For example, LFS is conducted every month while EC is conducted once in five years. The comparability between each source are not suitable because it does not represent the same situation between the data sources.

Furthermore, the concept, classification and definition of every sources are different. This is due to the different objective and aim of every census, surveys and research conducted. Thus, the objective to develop labour account is to compile the available data sources in Malaysia including administrative

records based on ILO guidelines and SNA concepts also the four countries mentioned as best practice.

Thus, the idea to establish Malaysian Bureau of Labour Statistics (MBLS) is to strengthen the demand data as MBLS is intended as a production and monitoring center for Malaysian labour market information on a regular basis for the strategic planning and socio-economic development of the country.

Ideally, by having a labour account, will helps to reduce the costs of data collection. Through labour account the same question related to labour will not be asking twice in different surveys in order for quality control. Labour account also can improve the data quality checks. After the adjustment of the data which have different definitions, concepts and coverage, the remaining discrepancies within these definition relationships will reveal the magnitude of error in one or more sources used (Buhmann et al., 2002).

References

1. Australian Bureau Statistics. (2018). Australian Labour Account: Concepts, Sources and Methods 2018 [cat. No. 6150.0].
2. Australian Bureau Statistics. (2018). Labour Statistics Fact Sheet Series. Australian Labour Account Fact Sheet 6.
3. Australia Bureau of Statistics. (2017). Media Released: 6150.0.55.001- Labour Account Australia, Experimental Estimates, July 2017.
4. Buhmann. B, Leunis. W, Vuille. A, Wismer. K. (2002). Labour Accounts: A Step Forward to a Coherent and Timely Description of the Labour Market.
5. Hoffman. E. (2000). Developing Labour Account Estimates: Issues and Approaches.
6. ILO (1993), General Report: Fifteenth International Conference of Labour Statisticians, Geneva.
7. Leunis. W.P. Verhage. K.G. Labour Accounts, Core of the Statistical System on Labour.
8. Netherlands Economic Institute. (1996). Labour Market Studies Netherland.
9. Peichl. A. Sieglöcher. S. (2010). Accounting for Labour Demand Effects in Structural Labour Supply Models.
10. Statistics Denmark. (2015). Documentation of Statistics for Labour Market Account 2015.



Labour supply statistics: Challenges and way forward



Noraliza Mohamad Ali, Nur Layali Mohd Ali Khan
Department of Statistics Malaysia

Abstract

From the perspective of labour supply, Labour Force Survey (LFS) has been conducted in Malaysia since 1972 to cover selected regions in the country. Since 1982, the coverage has been extended to produce national annual estimates of labour force statistics. With economic and social developments and transitions nationally and globally, the frequency and disaggregation was further improved to quarterly and later monthly estimates at national and state levels. Ever since the inceptions of the LFS, concepts, definitions and classifications has been reviewed and adopted accordingly with the recommendations of the International Labour Organisation (ILO) and best practices of other National Statistical Offices. Considering the frequency of disseminations of LFS statistics, at times, the information has been overly utilised and has been the subject of misinterpretation when used as a replacement or proxies to other unavailable labour market the statistics of labour supply through household approach to cater and complement the other dimensions of labour market statistics.

Keywords

Labour Force Survey; Labour statistics; Labour market information; Labour market dynamics

1. Introduction

Sound evidence-based policy making relies on comprehensive demographic, economic and social statistics. As much as the well-being of an economy is evaluated through the shift in the population structure and changes of the social landscape, labour plays an important part in understanding the world of work, such as the relationship between employment and growth, wage formation, the importance of human capital, migration and labour market regulations (ILO, 2013). Bean (2018), International Labour Organisation (ILO) (2017a) and KPMG Economics (2016) perceived labour statistics as the fundamental fragment of any labour market information system which is in turn very critical for research and policy formulation.

ILO (2017a) defined labour statistics as groups of official statistics relating to work, productive activities, workers, the characteristics of the labour market

and the way it operates. These statistics comprised of a wide range of topics and link to many other bodies of official statistics, such as economic, education and health (ILO, 2017a). The dynamics of labour statistics encompassed both demand and supply (ILO, 2017a). Among the labour demand statistics are the number and characteristics of enterprises, jobs, vacancies as well as the costs of hiring. Statistics about labour supply deal with the size, structure and characteristics of working-age population, and more specifically, information on employment, unemployment, and persons outside the labour force (ILO, 2017a).

In terms of approach of data collection, these statistics are available from various sources ranging from surveys and censuses using household or establishment approach as well as administrative records. The production of these labour indicators adopted different methodologies, which adhered to the corresponding international standards (ILO, 2017a; ILO, 2017b). From the perspective of labour supply statistics, the usual source is household approach through the conduct of censuses and surveys. Although not as popular as the other data sources, administrative record can provide information of labour supply as well.

In most countries, the statistics on labour supply are largely dependent upon the Labour Force Survey (LFS). Eurostat (2019) stated that LFS is a long-standing survey in Europe, going back to the 50s or 60s in some of the European countries. France was the first European country to carry out LFS in 1950, followed by Germany in 1957 (European Communities, 2003). The United Kingdom's first ever LFS was conducted in 1973 (Office of National Statistics (ONS) UK, 2017). In the United States of America (USA), Current Population Survey which is the equivalence of LFS is conducted by the U.S. Census Bureau to gauge labour supply since 1948 (U.S. Bureau of Labor Statistics and U.S. Census Bureau, 2016). Australian Bureau of Statistics (ABS) implemented the quarterly LFS for the country since November 1960 and later improved the frequency of data collection to every month beginning February 1978 (ABS, 2003). Closer to home, that is in the ASEAN regions, Singapore has reached its 41st edition of the LFS in 2018 (Ministry of Manpower, Singapore, 2019), while the undertaking of this survey in Thailand begun in 1963 (National Statistical Office Thailand, 2017).

In Malaysia, the LFS was conducted by the country's national statistical office, Department of Statistics, Malaysia (DOSM) since 1974 at irregular interval covering Peninsular Malaysia. The survey coverage was expanded to cover the whole country in 1982 and was conducted every year since then with

exception for the year 1991 and 1994¹. LFS also housed the supplements of other data collection modules i.e. Salaries & Wages Survey, Migration Survey and Survey of Manpower in the Informal Sector.

Considering the overarching role of the labour supply statistics within the national labour market information system in Malaysia, it is important that LFS as the primary source of these statistics is assessed objectively; and practical and realistic strategies is proposed to improve the production of labour supply statistics in the country.

2. The Labour Force Survey in Malaysia

The LFS in Malaysia remained as one of the key statistics in the labour market information framework. Initially conducted for the region of Peninsular Malaysia in 1974, the nationally representative survey took off in 1982. Due to the growing demands for regular and timely statistics, the survey frequency was increased to quarterly interval in 1998; and subsequently monthly survey took off since 2004.

Pen-and Paper Interviewing (PAPI) approach is employed for Malaysia's LFS in which trained enumerators visited households in selected living quarters (LQs) to collect demographic particulars of all household members and detailed labour force particulars of all members aged 15 years and over. A total of twenty-five per cent of the monthly allocated samples are repeated for the next quarter using Computer Assisted Telephone Interviewing (CATI). The survey population is defined to cover persons who live in private LQs and hence excludes persons residing in institutional LQs such as hotels, hostels, hospitals, prisons, boarding houses, and construction work site. The sample for LFS is drawn from Malaysia Statistical Address Registry (MSAR) which is made up addresses of LQs, composed into Enumeration Blocks (EBs) of 80 to 120 LQs each. The core reference material used to define concepts, definitions and classifications are as proposed by the ILO through the conventions and recommendations; and guidelines.

The estimates for the specific characteristics in the survey population are acquired through inflating the sample by the combination of adjusted weight and external weight. The adjusted weight is used to adjust for non-response in the survey, while the external weight i.e. the up-to-date population estimates is divided into specific characteristics of state, sex, age group, citizenship and ethnic group and compared to the sample of similar characteristics. On the basis of ratios of these distributions, correction factors or weights are derived which, when applied to the sample cases, make the

¹ The absence of LFS for the two years was due to resources constraint as the organisation prioritized the implementation of Population and Housing Census in 1991 and the Agriculture Census in 1994.

sample distribution conform to the external benchmark. The combination of these weights is then applied to the LFS sample data to obtain estimates of labour force statistics (DOSM, 2019).

Since the LFS is designed to be representative at the geographical areas of states as well as urban and rural areas, disaggregation of the estimates by numerous socio-demographic and economic characteristics must be interpreted with cautions and subject to relative standard error. In the meantime, statistics that are not published as well as the micro data are provided upon request with considerations to the reliability of the related statistics. The disaggregation of LFS statistics which are usually made available to users by frequency of data collection are as in Table 2.1.

Table 2.1: Frequency and disaggregation of Malaysia's LFS indicator

Frequency	Indicator	Disaggregation
Annual	Labour force participation rate	Sex, Age, Ethnic group, Educational attainment, Highest certificate obtained, State, Urban/rural area
	Employment-to-population ration	Sex, Age group, Ethnic group, Urban/rural area
	Labour force	Sex, Age group, Ethnic group, Marital Status, Educational attainment, Highest certificate obtained, State, Urban/rural area
	Employed	Sex, Age group, Ethnic group, Marital status, Educational attainment, Highest certificate obtained, Industry, Occupation, Status in Employment, State, Urban/rural area, Mean and Median hours worked
	Unemployed	Sex, Age group, Ethnic group, Marital status, Educational attainment, Highest certificate obtained, Working Experience, Duration of unemployment, State, Urban/rural area
	Unemployment rate	Sex, Age group, Ethnic group, Educational attainment
	Outside labour force	Sex, Age group, Ethnic group, Educational attainment, Highest certificate obtained, State,

Frequency	Indicator	Disaggregation
		Urban/rural area, Reasons for not seeking work
Quarterly	Labour force participation rate	Sex, Age group, Ethnic group, Educational attainment
	Labour force	Sex, Age group, Ethnic group, Educational attainment
	Employed	Sex, Age group, Ethnic group, Marital status, Educational attainment, Highest certificate obtained, Occupation, Status in Employment, State, Urban/rural area, Mean and Median hours worked
	Unemployed	Duration of unemployment
	Unemployment rate	Sex, Age group
	Outside labour force	Sex, Age group, Reasons for not seeking work
Monthly	Labour force participation rate	None
	Labour force	
	Employed	
	Unemployed	
	Unemployment rate	
	Outside labour force	

3. Assessment of Malaysia's LFS as the source of national labour supply statistics

Being one of the longest running national household surveys, the LFS is one of the most convenient sources in providing a rather long time series of the annual, quarterly and monthly principal labour force statistics. The annual statistics goes back as far as 1982, while the quarterly and monthly series begin in 1998 and 2004 respectively. Table 3.1 shows labour supply during selected years beginning 1982 leading up to the most recent year. It is observed that as the population grow, so does the employed person. However, there's a noticeable increased in the share of non-citizens that is 10.3 per cent of total populations in 2018 as opposed to 4.3 per cent of total populations in 1992. Table 3.2 shows the share of Gross Domestic Products (GDP) by sector during similar period, wherever available. It is observed that the share of GDP in agriculture sector reduced considerably while the services sector dominated with majority share since 2012.

Table 3.1: Population estimates and principal labour force statistics, Malaysia, selected years

Year	Unit	1982	1992	2002	2012	2018
Population	('000)	14,651.1	19,067.5	24,542.5	29,510.0	32,382.3
Citizens	('000)	14,651.1	18,205.1	22,942.1	26,961.7	29,059.6
Non-citizens	('000)	n.a.	862.4	1,600.4	2,548.3	3,322.7
Labour force	('000)	5,431.4	7,319.0	9,886.2	13,221.7	15,280.3
Employed	('000)	5,249.0	7,047.8	9,542.6	12,820.5	14,776.0
Unemployed	('000)	401.9	515.0	516.7	589.3	437.5
Employed less than 30 hours	('000)	182.4	271.2	343.5	401.2	504.3
Outside labour force	('000)	2,944.6	3,783.6	5,473.8	6,927.4	7,094.4
Labour force participation rate	(%)	64.8	65.9	64.4	65.6	68.3
Unemployment rate	(%)	3.4	3.7	3.5	3.0	3.3

Note: n.a. Breakdown for citizenship is not available

Source: Current Population Estimates, Malaysia, 2018-2019; Labour Force Survey, Various Years, DOSM

Table 3.2: Share of GDP by kind of economic activity at constant price, Malaysia, selected years (%)

Year	1992	2002	2012	2018p
Price	(1987=100)	(2000=100)	(2010=100)	(2015=100)
Kind of economic activity				
Agriculture	14.6	8.3	9.8	7.3
Mining and Quarrying	8.6	10.2	9.5	7.6
Manufacturing	25.1	29.0	23.2	22.4
Construction	3.8	3.9	3.8	4.9
Services	38.8	42.2	52.5	56.7
Less: Undistributed FISIM	5.0	4.5	-	-
Plus: Import Duties	4.0	1.7	1.1	1.2
GDP at Purchasers' Prices	100.0	100.0	100.0	100.0

Note: Starting 2005, FISIM has been distributed to all activities

Source: National Accounts, Various Years, DOSM

Within the view of the policy makers and regulators, this lengthy time series is very convenient for monitoring purposes, looking at historical points to identify trends, cycles and structures and ultimately formulate the suitable strategies and initiatives. This features also appeal to the academia in conduction labour market and human capital development studies.

Beyond the compulsory questions to identify labour force status of the population, Malaysia's LFS which canvassed all members in the selected households is one of the few national surveys that asks detailed demographic characteristics of the population. Examples of the questions are marital status, educational attainment, highest certificate obtained and field of studies. This is a cost-effective method to ensure regular updates as a substitute to the population and housing census. This features also make it the most commonly used survey to ride additional modules and supplementary questionnaire regularly or on ad hoc basis. This corroborates with the findings of ILO (2017a) and European Communities (2003) that LFS offers a consistent framework to study employment, unemployment and persons outside the labour force concurrently, with rich and extensive dimensions for disaggregation, as well as provides venue to study parts of informality through the informal sector employment. In line with the European Communities (2003) observation, the Malaysia's LFS also to some extent facilitates the opportunity to obtain labour supply information across all sectors of the economy in a consistent manner.

LFS in Malaysia can be considered as one of the more matured and established data collection activity in the national statistical system. Sound methodology is in place since it adheres to ILO's international standards in terms of compilation and dissemination. This allows for international validity, consistency, accuracy, reliability, timeliness and comparability. To date, the statistics derived from Malaysia's LFS is consistently used to update the ILOSTATS statistical database².

As LFS presents numerous advantages, it is common to witnessed its usage across various research and studies. The use of LFS to mobilize efforts on human capital development is most apparent in the medium term strategies documents, i.e. five-year Malaysia Plans, the most recent being Mid-Term Review of the Eleventh Malaysia Plan. Additionally, short term policies documents that utilised the LFS especially within the scope of human capital development are Bank Negara Malaysia (BNM) Annual Report and Ministry of Finance (MOF) Economic Outlook. The statistics of the LFS is also the main features in the studies on labour market position employed by independent and government-backed research agencies such as Khazanah Research Institute. At the international forefront, labour market studies by the World

² The ILOSTATS is a website maintained by the ILO Department of Statistics which offers access to data tables for key indicators, statistical briefs, concepts and methods.

Bank and Organisation for Economic Co-operation and Development (OECD), among others, mostly depend on the LFS as the major data sources.

In spite of its many plus points, the LFS is not without faults and limitations. As far as reliability and quality of the estimates goes, LFS being a household survey is subjected to sampling errors and non-sampling errors. The sampling errors occurrence is especially true when the estimates are disaggregated for small groups or areas which are under-represented in the sample (DOSM, 2019; ILO, 2017a). Although the sampling error can be reduced by increasing the number of observations sampled, it is not the most financially smart solution in the long run. The non-sampling error in LFS might prevail due to misleading comprehension of definitions and concepts either by enumerators or respondents; or defective methods of data collection. Unlike the sampling error, this error may rise with the increase in sample size. Banda (2003) emphasised that this type of error can be more detrimental for large-scale household surveys in the absence of proper control mechanism.

Another issue to consider is the use of proxy respondents i.e. one household member providing the required information on all the members of his or her household. Since 75 per cent of the sampled households in the national LFS currently uses PAPI where enumerators visited households to obtain information, more often than not, households are not fully occupied due to members being at work, school or other places. According to the ILO (2017a), this may also hamper the precision of the response.

Due to its reputation as the most cost-effective and frequent data collection activity, the LFS is often ridden on for testing new data collection instrument in addition to the regular supplements of Migration Survey and Salaries & Wages Survey. At times, these added loads might compromise the quality of responses for labour-related fields in the questionnaire. Furthermore, this also may add to respondents' burden and eventually cause the response rate to decline.

Considering the sample design which does not take into account economic activity and occupation of household members, certain information is obtained indirectly and is perceived as the by-product of the LFS. The obvious instances would be estimates of employment across economic sectors or occupation categories. Although both might produce statistically reliable estimates at major groups, disaggregation at detail subsectors or occupations may not be able to offer nationally representative estimates. This is also customary for other variables in the LFS such as educational attainment, highest certificate obtained and field of studies. The sampling base on which such estimates would depend would be too small, and the degree of variability correspondingly high (European Communities, 2003).

Being a regular survey with multiple demographic and socioeconomic variables, LFS is sometimes the subject of misused. The short term difference

in the number of employed persons i.e. the employment change is often interpreted as jobs created and denotes as labour demand statistics. Since LFS adopts household approach, the interpretation might provide the wrong signal to the market and the overall economy. Similarly, when used to measure graduate employability, the LFS might not offer the most accurate results since it is not designed to cater for potential labour supply or track graduates across the labour market.

4. Way forward for the labour supply statistics in Malaysia

LFS definitely has a special position as the source of labour supply in the spectrum of labour market information. Nevertheless, as we move ahead within the realm of demographic transition towards ageing population and urbanised households on one hand, and the rapid technological change and revolutionised world of work and economic landscape on another, there is a pressing need to alter and improve the current method for production of labour supply statistics.

Firstly, it is important to extend the coverage of the LFS' sample to take into account all population including those living in institutional LQs. Semi-skilled employed persons within selected subsectors, especially non-citizens in the agriculture and construction sectors often reside in communal houses. Thus, this recommendation is to ensure a more comprehensive coverage, and essentially increase the accuracy of the labour supply estimates in the market.

The later generation of respondents are technology savvy, value privacy and non-intrusive. As far as the mode of data collection for LFS is concern, it is timely that we transform towards a more respondent friendly self-completion mode through drop-off and pick-up of questionnaire and e-survey. In keeping with this modernization, the content of the questionnaire should be reviewed and simplify where necessary to cater for the primary objective of the LFS which is to determine the labour force status of the population as either employed, unemployed, underemployed and outside labour force.

Further than taking up surveys, it is high time that the national statistical system ventures into a more strategic source i.e. administrative records. As indicated by ILO (2017a), the nature of administrative records which is created and maintained by the corresponding agency is an economical source with real-time information and exhaustive coverage. UK leveraged the administrative record of unemployment insurance since 1920s to complement surveys and censuses data (Bean, 2018). MOM, Singapore and BLS, USA also incorporate information from various sources to produce comprehensive monthly updates on labour market situation. In this respect, Malaysia's Employment Insurance System (EIS) which took off contributions from employees since January 2018 and offers insurance to retrenched employees

since January 2019 is one very potential source of labour supply information. Besides providing immediate financial benefits and upskilling opportunities to workers who have lost their jobs, the other objectives of EIS also includes to provide up-to-date and comprehensive labour market information to policymakers (PERKESO, 2019).

As the economic structure diversifies and the technology changes rapidly, the world of work is becoming more and more dynamic. As such, no one particular source should be over-utilised to form a comprehensive labour market information and analysis framework. Malaysia is geared towards improving the labour market statistics in keeping up with these changes so as to provide labour statistics that can be linked to the economy and social backdrop.

References

1. Australian Bureau of Statistics. (2003, April 14). *Labour Force*. Retrieved from Australian Bureau of Statistics (ABS): <https://www.abs.gov.au/AUSSTATS/abs@.nsf/DOSSbyTopic/139689E1A84FE4F0CA256BD00028B0E5?OpenDocument>
2. Banda, J. P. (2003, December). Nonsampling errors in surveys. *In Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys, United Nations Secretariat, New York*. (pp. 3-5).
3. Bean, R. (2018). *International labour statistics: A handbook, guide, and recent trends (Vol. 3)*. Routledge.
4. Department of Statistics, Malaysia. (2019). *Labour Force Survey, Malaysia, 2018*. Putrajaya: Department of Statistics, Malaysia.
5. European Communities. (2003). *The European Union labour force Survey*. Luxembourg: Office for Official Publications of the European Communities.
6. Eurostat. (2019, April 17). *eurostat Statistics Explained*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey_%E2%80%93_development_and_history#Development_of_the_EU-LFS
7. International Labour Organization. (2013). Edited by Cazes, S., & Verick, S. *Perspectives on labour economics for development*. Geneva: ILO.
8. International Labour Organization. (2017a). Quick Guide on Sources and Uses of Labour Statistics. Geneva, Switzerland. Retrieved from www.ilo.org/publns
9. International Labour Organization. (2017b). *Visualizing Labour Market: A Quick Guide to Charting Labour Statistics*. Geneva: ILO.
10. KPMG Economics. (2016, February). *KPMG Research Paper: The Role of Capital and Labour in Driving Economic Growth in Australia*. KPMG.

11. Ministry of Manpower, Singapore. (2019). *Labour Force in Singapore, 2018*. Singapore: MOM, Singapore.
12. Office for National Statistics, UK. (2017). *User Guide (Vol.1): LFS Background and Methodology 2016*. ONS, UK.
13. Pertubuhan Keselamatan Sosial (PERKESO). (2019, July 29). *Employment Insurance System*. Retrieved from Social Security Organization : <https://www.perkeso.gov.my/index.php/en/mengenai-sip/our-objective>
14. U.S. Bureau of Labor Statistics and U.S. Census Bureau. (2016). History of the Current Population Survey. In *Current Population Survey Design and Methodology Technical Paper 66* (pp. 2-1 -2-7). U.S. Bureau of Labor Statistics and U.S. Census Bureau.



Managing unstructured data through Big Data Analytics towards intelligent insights



Sasongko Yudho
eBdesk, Malaysia

Abstract

Big Data Analytics (BDA) has been the in-trend technology for almost ten years. The capability in managing unstructured internet data makes BDA to be the benchmark in any data processing implementation. The scalability characteristic and being developed under opensource community, make BDA become cost-effective for any organization to implement. The robust architecture of BDA allows various research, especially in artificial intelligence areas which mostly work based on unstructured data and non-linear analytical models. The rise of BDA contributes significantly in the development of learning based analytical research which not easy to be implemented using the traditional system architecture and framework. The parallel processing capability offered by BDA enable many organizations to embark in massive processing with lower cost of implementation. This paper also describes the implementation of BDA in two domain of case study, the media intelligence and the marketplace intelligence. Both domains have been the major recent requirements in various organization in establishing strategy and action plan. Government need to monitor the public perception in media, private sector needs the marketplace intelligence to position the product marketing strategy, industry requires real-time insight that able to describe the dynamic of market, etc, are problem statements that become the factors in embarking BDA implementation.

Keywords

Big Data, Media Intelligence, Unstructured Data, Price Intelligence

1. Introduction

Data has been the foundation in any organisation as important component in decision making process. Any policy established should be based on data driven study to give better decision quality and result. There are less organisations that are still using assumption and tradition as the benchmark in establishing decision today. Those that are still implementing traditional approach will be left behind and losing the competitive advantage.

As the world is becoming more integrated whereby anything can influence each other, the need of massive data analytics become a compulsory. Stock market is no more rely only on technical and fundamental analysis. They also

need to measure the market sentiment based on people opinion and transform it into scoring variable to be calculated together in the stock predictive model.

1.1 Unstructured Data

Data can be categorised into structured and unstructured based on its characteristic. Structured data is data established in a structured format and pattern that has been defined properly. Tables, worksheets, are typical format that are used to store the data. For so many years organizations have been putting more attention to the structured data. They invest million of dollars developing framework to manage the structured data. Relational database, data warehouse, business intelligence, are frameworks have been used by organization in managing structured data.

The unstructured data is data without pattern and definitive format. It can be a narrative text or binary code without any descriptive provided. Content of books, content of documents, memos, contracts, images, movies are sample of unstructured data. This kind of data receives less attention from the organization for data processing as the characteristic of the data makes them challenging to process. The common processing done to the unstructured data are indexing and searching.

Unfortunately, 90% of data in the world is in unstructured format. It has been a challenge for many organizations to engage with unstructured data processing. Putting the context on the data, extracting the object based on language grammar, clustering the content for similarity, are research conducted as part of unstructured data processing. It will be a great impact for organization if they are able to manage the unstructured data, so better knowledge can be acquired, and better decision making can be established.

1.2 The Internet Era

Internet has been a commodity since the booming of dot com during end of nineties. The cost required by public is getting more affordable every year. Internet penetration reports are also showing positive trend in many countries. With the smart phone as everybody's communication platform, internet access become a basic need for everyone.

The social media also contributes significant impact on digitalization of asset. More content is produced every day in digital format and store them in the internet-based platform like clouds. It makes the internet become the giant platform of data sources which can be one of organization directions in seeking more sources for information.

2. Methodology

In the principle, Big Data can be defined as an all-encompassing term for any collection of data sets so large and complex that it become difficult to process using traditional data processing applications. Of course, the definition has been updated from time to time according to the development of the technology supporting the Big Data ecosystem.

2.1 A Need for Massive Processing

The original intention in big data research is the need of parallel processing. Based on computer architecture, the components performing processing are CPU and RAM, while data is stored in disk. The technology in CPU and RAM development is not as fast as technology in developing disk for storage. As result we can see the size that disk has is growing very fast compare the speed of CPU and the size of RAM.

A computing process is started with transferring some data from disk into memory, and then CPU will perform the processing. There is an issue of performance with that kind of approach as the data transferred into RAM will be limited following to the size of the RAM itself. So, instead of performing total in-memory processing, the platform is only able to perform batch processing. Some of processes are still acceptable to be performed using batch processing. Unfortunately, this kind of process is not suitable for some analytics processing, such Machine Learning as part of Artificial Intelligence framework.

The rise of Big Data Analytics (BDA) contributes significantly for Artificial Intelligence (AI) research development. AI works based on learning algorithm whereby training data set is established for the machine to learn and to understand the context of the knowledge. Some of popular algorithm in AI is Neural Network. This algorithm works by loading all data into memory and perform repetitive process in order to establish the model. The quality of the model will be based on the volume and variety of data, and the number of repetitions.

This methodology requires a massive parallel processing to achieve the best accuracy model. It is very costly to perform such processing using traditional technology framework and infrastructure. It is also the answer why neural network was not popular, until the development of Big Data Analytics has reached the maturity from the perspective of implementation.

2.2 Scalable Distributed Resources

BDA offers a scalable resources architecture. It allows user to have multiple resources from different machines to be aggregated as a single processing cluster. Scalability is the key of BDA implementation which has been adopted by many organizations especially the Social Media Provider such as Google,

Facebook, Twitter, etc. It has been proven as robust platform which able to manage massive data processing.

Scalability is the ultimate benefit of BDA, be it from storage, memory and CPU processing perspective. Aggregating resources into single cluster enable user to perform a total in-memory processing as machines can be added to the cluster easily. For example, to perform in-memory 5TB data size, we can deploy 20 machines with 256GB size of RAM each. BDA combine the power of RAM from those 20 machines virtually as single memory for processing purpose. The physical distribution of data into each of RAM will be done by the BDA software without complex setting and configuration by the user.

2.3 Opensource for Cost Efficiency

BDA also offer great benefits in term of investment for organization. The main component of BDA is opensource, developed under Apache community. It license-free which allows organization to use with no limit of capacity. Since it is opensource, there are continuous development for new features and architecture by the community. Various forum also has been activated for community to discuss any issues and potential development which can be used as references.

3. Results

We have successfully established BDA architecture focusing in two domains, Media Intelligence and Marketplace Intelligence. Both case studies are leveraging on unstructured data captured from internet using crawling methodology.

3.1 Media Intelligence

3.1.1 Public Perception Matters

Media has the capability to affect the world by playing around with public opinion be it from people on the streets, government officials and decision makers. It is common that financial investor evaluating company performance and credibility can be swayed by positive or negative perception about the said companies from the media, which directly will impact its economy standing. Changes in market price are likely the responses to some events reported in the news. It is not impossible for investors to act based on these perceptions that they read in newspaper. In the long run, perception of issues can cause massive chaos when it is not well managed.

3.1.2 Natural Language Processing

All data captured from the media are in the format of narrative text. It is written in human language based on specific grammatical rules. We use Natural Language Processing (NLP), a sub-field of Artificial Intelligence

methodology in order to perform various processing of the natural language data in the media content with objective to establish some context of the data.

The NLP objective is to extract objects in the content and classify them into grammatical role. Starting from identifying the object as NOUN or VERB, until the process to identify the function of the sentences, such main-clause and sub-clause. We also perform the NLP for advanced computational linguistics by classify the name of person, name of location, people statements, summarization, etc.

Hybrid approach between machine learning and dictionary based are used as the framework of NLP model. We leverage on the BDA to perform in-memory processing in establishing the NLP model using hundreds of thousand data training. We manage to establish NLP model for English, Malay and Indonesian language, to be used in media analysis.

3.1.3 Issue Analysis

After the unstructured data has been processed using the NLP framework, statistical analysis is performed to develop issue analysis based on the context of media characteristics. Combining both liner and non-linier analysis, we perform various data story-telling such as:

- **Topic Management:** Clustering the contents into topic of interests
- **Analysis Dashboard:** Providing summary of analysis for certain topic
- **Influencer Analysis:** Detecting list of people that likely having the ability to influence perception
- **Sentiment Analysis:** Providing perception sentiment on positive/negative
- **Ontology Analysis:** Generating relationship among people based on specific context and problem definition
- **Topic Similarity:** Establishing the similarity of one issue with others and develop the similarity behavior and characteristics
- **Social Network Analysis:** Developing the social network analysis based on conversation in social media



Picture 1. Media Intelligence

3.2 Marketplace Intelligence

3.2.1 The Virtual Mall

Online marketplace is a platform that products or service information is provided by multiple third parties, whereas transactions are processed by the marketplace operators. Put it into a traditional context, online marketplace is a virtual department store.

Established 19 years ago, Lelong claims that it differentiates itself from other E-commerce by providing personalized store and offering free classroom that covers e-Commerce tips and challenges, product photo taking tips, product sourcing, shipping as well as providing real life case studies as references for those who want to be a success seller.

3.2.2 Characteristics and Behaviour

• Item Category Distribution

There are 800,000 goods listing from 1st of December 2018 to 31st of January, 2019 by unique URL, and the biggest catalogues of goods in Lelong are electronic and electrical appliances. This is a common situation in E-commerce, where electronic and beauty products always take the lead.



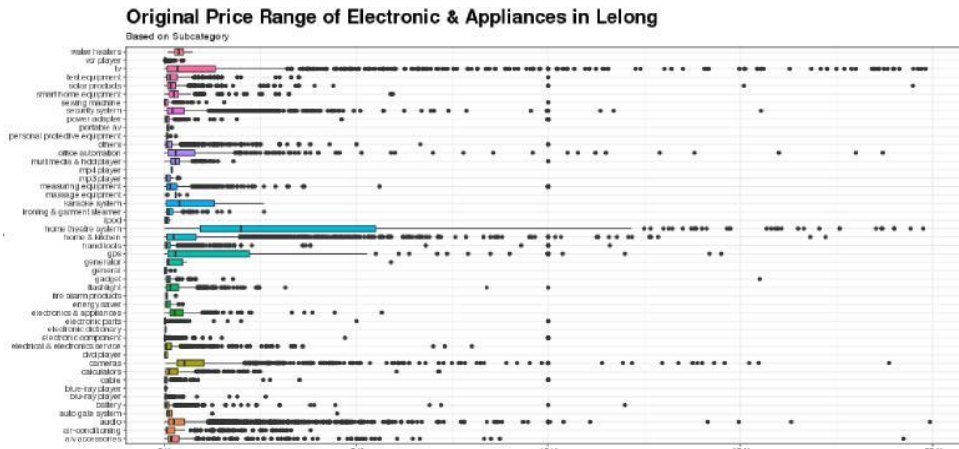
Picture 2. Item Category Distribution

- Seller Profile**

There are 3,191 sellers that sell their products in Lelong on the said period. From the pie chart, about 88.7% sellers ship their products from Peninsular area, 9.1% ship overseas and only about 2% sellers ship from East Malaysia. Peninsular sellers dominate all field. Oversea sellers are more prominent in the “sport & outdoor” category, while not being active in B2B, industries products and office equipment. Besides, only Peninsular sellers are selling parts, bracelets and bathing & skincare products.

- Price Characteristics**

The variance of price is very high in the market. Some products can go as much as RM40,000. We did some cleaning by eliminating products that diverted too much from the skew bar. Still, we see a lot of products price far greater than the typical range of pricing. TV, security system, home theatre system, and hand tools are some categories with pricing that are too diverse.



Picture 4. Lelong.my Price Distribution

4. Discussion and Conclusion

BDA plays important role in managing the unstructured data. The case study shows a better insight in understanding the internet data as part of organisation decision-making process to establish strategy and actions. The massive unstructured data set also allows the research in non-linear statistical model, especially for artificial intelligence, to be explore further.

There are many areas in parallel processing research that can be enhanced through Big Data framework and architecture. The opensource ecosystem enable the technology to be developed together by the community.



MyHarmony: Generating statistics from clinical text for monitoring clinical quality indicators



Md. Khadzir Sheikh Ahmad¹, Mohd Syazrin Mohd Sakri¹, 'Ismat Mohd Sulaiman¹, Syirahaniza Mohd Salleh¹, Dickson Lukose², Omar Ismail¹, Abd Aziz Latip³, Muhammad Aiman Mazlan³

¹Health Informatics Centre, Planning Division, Ministry of Health, Malaysia

²GCS Agile Pty. Ltd.

³MIMOS Bhd.

Abstract

The Ministry of Health developed MyHarmony with MIMOS Bhd. as part of the Malaysian Health Data Warehouse (MyHDW) initiative. MyHDW aims to be a trusted source of truth of comprehensive health data structured for analysis. MyHarmony fulfills that criteria by enabling data and information from unstructured form to be mined, such as texts, images, and sound. MyHarmony's first deliverable is the ability to mine clinical texts using Natural Language Processing (NLP) with SNOMED CT as its knowledge-base of clinical terms. MyHarmony is the engine that retrieves data and information into computer processable form by assigning SNOMED CT codes, which can then be further analysed statistically. MyHarmony is able to recognise and harmonise different terms that mean the same. It also understands context for a more accurate coding; such as negations (no, not known, unknown) and conditionals (past history, symptoms of, previous). Using SNOMED CT, MyHarmony's ability is further advanced by using subsumption technique for a more comprehensive statistical results. This study will present a use case where clinical text from anonymized hospital discharge summaries can generate clinical indicators using MyHarmony for health managers. An added benefit to the operational (hospital) staff is the ability to produce such indicators in an efficient and timely manner by reducing workload for data collection and submission. MyHarmony could be the new and improved way to provide important statistical measures for evidence-based health planning, leading to improved healthcare services and health as a whole.

Keywords

MyHarmony; MyHDW; Text mining; Quality Indicators; SNOMED CT

1. Introduction

The Ministry of Health developed MyHarmony with MIMOS Bhd. as part of the Malaysian Health Data Warehouse (MyHDW) initiative. MyHDW aims to be a trusted source of truth of comprehensive health data structured for analysis. MyHarmony is an application in the Malaysian Health Data Warehouse (MyHDW) that aims to analyse semi-structured and unstructured

data. The unstructured data can be in the form of free-text, visual, audio and machine generated data. Unstructured data does not have predetermined values and not stored in an organized manner to be analysed by a conventional data warehouse. Therefore, other techniques need to be applied. MyHarmony aims to address this and be included as part of MyHDW.

There were three (3) major deliverables in the conceptual stage. The first part refers to the development and implementation of health terminology standards, namely SNOMED CT, which will be the knowledge bases for MyHarmony. The second part was harmonization of the medical terminology to SNOMED CT terms by way of mapping. The last part was about the development and implementation of MyHarmony to show that the application can codify relevant terms in free-text using Natural Language Processing (NLP) technique. The SNOMED CT codified data can then be analysed for information generation.

2. Methodology

The development was first started in 2014 with the development of Cardiology Refset. Cardiology Refset was the terminology reference for the MyHarmony engine during the harmonisation/mapping and codification process. Cardiology Refset (version 1.0) was completed and released in 2014. It is a simple reference set [1] containing about 600 terms related to Cardiology speciality including signs and symptoms, diagnoses, procedures, body structures, medical devices and medications. It was delivered in time to be tested on MyHarmony standalone system to generate National Cardiovascular Disease (NCVD) registries.

The draft Refset and method was presented during IHTSDO meetings and Expo in succession on September 2013, October 2013, and April 2014 to gain feedback from experts in the international community. The finalized method was presented during SNOMED CT Expo, October 2014 [2].

The Cardiology Refset was then expanded to include all cardiology related terms and Cardiology Refset v1.1 was completed in July 2016 containing more than 6000 concepts. First, more than 300,000 SNOMED CT concepts (Fully Specified Names) were extracted and reviewed by PIK using eyeballing technique. About 12,000 concepts that were believed to be related to Cardiology specialty was given to the clinicians for review. The clinicians reduced the number of concepts to about 6,000. Additionally, the Refset included local terms and common abbreviations which were mapped to existing concepts.

Next, the team utilise MyHarmony to generate the analysis. There were 4 main functions in MyHarmony:

1. Terminology Management– to allow user to upload the SNOMED CT International Release content into MyHarmony, and upload SNOMED CT Refset in reference to the SNOMED CT International Release.
2. Data Management– to allow user to upload the data that will be harmonized and codified. This function also allows user to view the content of the data.
3. Codification Management– to allow user to codify the dataset according to the selected SNOMED CT Refset and view the codified data for validation purpose.
4. Query Management– to allow user to explore the data by generating queries using Structured Query Language (SQL).

The functions were arranged according to the work process. First, the SNOMED CT International Release, SNOMED CT Cardiology Refset, and the dataset needs to be uploaded. Then, the dataset is codified and saved. Using the Query Management, the codified data can then be explored via data profiling and query generation.

For initial analysis, SNOMED CT International release version 20160731 was used as the Cardiology Reference Set was developed using this version of SNOMED CT.

The team received a set of database from a hospital with cardiology service which consists of 16224 discharge summaries from year 2017. The database was then uploaded into MyHarmony. The personally identified information (patient names, ID, and street address) were anonymised prior to codification and analysis. The output is a codified dataset, which enable information processing and analysis by machines.

Using the Query Management, the codified data was then be explored via data profiling and query generation.

3. Result

The team conducted several data profiling queries to ensure that MyHarmony were able to capture the data correctly. For example, the number of records by month between Raw data (MyHarmony without SNOMED CT) and Harmonized data (MyHarmony with SNOMED CT) should return the same result. Other examples of data profiling queries are the number of records by gender, by specialty, and by ethnicity.

Next, the team developed queries required by the National Cardiovascular Disease (NCVD) registries and compare the results with published registry reports. For example, querying the number of Ischaemic Heart Disease (IHD) by gender shows 1:4 female to male ratio, which is a similar ratio in the registry reports. Furthermore, the query also shows that Harmonized data captures

more result compare to Raw data due to SNOMED CT relationship structure, thus capturing all the subtypes of IHD and its synonyms or ways of writing.

The registry, however, only captures three (3) diagnosis due to its structured format, which are ST Elevation Myocardial Infarction, Non-ST Elevation Myocardial Infarction, and Unstable Angina. This trend and pattern comparison allow validation by the Clinicians and gains their buy-in in using MyHarmony.

The team also tried to generate more queries required by the NCDV registry. However, it was limited by the documentation in the discharge summary. Registry queries requires more detail information that may often not documented in a discharge summary, such as information on smoking status and complications of procedures.

After that, the team was tasked to generate National Cardiology Key Performance Indicators (KPIs). MyHarmony are able to generate 7 out of the 8 KPIs (KPI 2 to 8). The first KPI was excluded because the data are available at the clinic and not documented in inpatient discharge summaries. The Health Information Framework (HIF) was developed for the 7 KPIs, which detailed out the inclusion and exclusion criteria, the target, the formula, the terms used by MyHarmony, and query, and lastly a section for additional notes.

Preliminary manual validation on the completeness and accuracy of codified data shows 90% precision and 70% recall. The content of those records is complex as it does not follow grammar rules, and contains a large number of short forms, abbreviations, acronyms and analogous terms (e.g., synd, ACS, CCS IV, NYHA 2). One example of record is "2VD with RCA culprit lesion - Ad hoc PCI DES to RCA and LAD" which is challenging to codify using approaches based on strict grammar. The revised version of MyHarmony uses a different approach based on shallow parsing and the consideration of multiple suitable combinations of words in a sentence. With further iterations and improvement in the mapping, these challenges were overcome[3].

From the SNOMED CT codified database, the system was able to show a more accurate result during analysis . This is because MyHarmony capitalises on the existing SNOMED CT relationships structure between concepts. In this case, when querying "Number of Ischaemic Heart Disease cases per year", MyHarmony search the code and term for "Ischaemic heart disease", its synonyms and accepted abbreviations, and all the subtypes of Ischaemic heart disease such as all subtypes of "Myocardial infarction" and "Angina". MyHarmony aggregates these records resulting in a more accurate analysis.

Usually, the result where MyHarmony utilise SNOMED CT's relationship structure would show more records. This is because Mi-Harmony was able to aggregate data not just through String match, but also utilize the IS-A hierarchy structure in SNOMED CT. For example, querying "Ischemic heart disease" will gather clinical records with synonymous terms like "Ischaemic

heart disease" and "IHD"; as well as clinical records containing all the subtypes of "Ischemic heart disease" such as "Myocardial Infarction" and "Unstable angina".

Context awareness such as negation and past events were also applied. For example, the term "No chest pain", "No known history of diabetes mellitus", and "Symptoms of heart failure" will not be coded as the presenting condition. Additionally, terms like "Previous history of", "Previous admission of", and "Family history of" within the same sentence as a clinical condition will not be coded as the current condition for the record.

4. Discussion and Conclusion

When showcasing these abilities to the clinicians, the team agreed that MyHarmony was able to:

- i. Generate more information from free-text utilising the SNOMED CT structure, thus, reducing the effort needed to collect data in a structured manner such as in a registry and indicator reports;
- ii. Able to generate new information by retrospectively running new queries on old discharge summary records; thus, reducing the effort and time to collect data in a prospective manner when new questions arise, such as for indicator reports that often change on a yearly basis;
- iii. Able to deliver information in a timelier fashion; thus clinicians and health managers are able to plan and take action without waiting for a 1 to 3 yearly report;
- iv. Improve documentation of clinicians when they are aware of MyHarmony's ability during roadshows.

Generating indicators for monitoring and evaluation can be a burden even for healthcare facilities equipped with EHR. Conventionally, collecting data for indicators requires multiple data entries in aggregated manner, with manual submission to central agencies, where the results are only published on a yearly basis. Introducing MyHarmony may reduce these burdens. Capturing data from the source in an automated way, i.e. free text documented by doctors, would reduce duplication of work and the amount of resources to capture the data into manual form. Having the data in granular form would allow a more dynamic analysis and prevents dishonesty. Information required, whether old or new information, can be formulated and disseminated back to the clinicians and health managers in a timelier fashion.

MyHarmony has the potential to expand further in its implementation and technology. However, there are still challenges to be addressed. Currently, MyHarmony has been developed to mine free-text for Cardiology via a back-end approach. It uses a single version of SNOMED CT International. The team is still researching the best approach to manage SNOMED CT versions and its

codified data, which may impact the resulting analysis in an inconsistent way. The team are also seeking international experience for this matter.

Other challenges include researching a more efficient and effective method to develop SNOMED CT Refsets. The initial method by referencing terms required by registries or indicators has been established. However, expanding the SNOMED CT Refsets to include relevant terms for a specific clinical specialty or domain needs to be refined further. Eye-balling technique to search the entire SNOMED CT content have its strength and weaknesses. Even though it is a very thorough method, there are possibilities of missed terms and very time consuming. Despite these challenges, the journey in developing MyHarmony and the lessons learnt has allowed the team to refine the methods and processes to expand the use of MyHarmony to other clinical specialties.

Analysis from unstructured data would hope to complement analysis from structured data (like census and registries), with the additional benefit workload reduction to provide timelier, trusted, and dynamic information.

References

1. SNOMED CT Simple Reference Set:
<https://confluence.ihtsdotools.org/display/DOCRFSPG/5.1.+Simple+Reference+Set>
2. Mohd Sulaiman I, Sheikh Ahmad MK. SNOMED CT Cardiology Reference Set Development, Malaysia. Proceedings of the SNOMED CT Implementation Showcase [Internet]. Amsterdam: IHTSDO; 2014. Retrieved from:
http://ihtsdo.org/fileadmin/user_upload/doc/showcase/show14/SnomedCtShowcase2014_Abstract_14058.pdf
3. Abdul Manaf NA, Mohamed K, Lukose D. Harmonizing EHR Databases with SNOMED CT. Proceedings of the SNOMED CT Implementation Showcase 2014 [Internet]. Amsterdam: IHTSDO; 2014.
<https://confluence.ihtsdotools.org/display/FT/SNOMED+CT+Implementation+Showcase+2014>



A copula approach to spatial econometrics with applications to finance



Hideatsu Tsukahara

Faculty of Economics, Seijo University, Tokyo, Japan

Abstract

Traditional models in spatial econometrics utilize a spatial weight matrix as a means to express spatial dependence, but its choice is quite arbitrary. Besides, it imposes a linear structure between dependent variables; in its simplest form, a dependent variable at one spatial unit is a linear combination of dependent variables at other spatial units. When the underlying disturbance distribution is assumed to be Gaussian or elliptical in general, the model does not allow asymmetry in dependence structure and tail dependence for spatial interactions. These restrictions are too strict in some applications, for example, to financial data. In this study, therefore, we generalize existent models to allow for some nonlinear and tail dependence in disturbance distribution by applying the copula approach which somehow reflects the spatial dependence indicated by spatial weight matrix. After discussing some properties of the resulting model, we develop an estimation method assuming (semi)parametric copula. Simulation results illustrate the applicability of our procedure, and some real applications to financial data will be given.

Keywords

Spatial dependence; tail dependence; asymmetry

1. Introduction

Suppose that there are N spatial units. Let

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}, W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{22} & \dots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N1} & \dots & w_{NN} \end{pmatrix}$$

For $i = 1, \dots, N$, y_i is a dependent variable at spatial unit i , and ε_i denote a disturbance term. Components of a *spatial weight matrix* W satisfies $w_{ij} \geq 0$ for all $i, j \in \{1, \dots, N\}$ and $w_{ii} = 0$ for all $i \in \{1, \dots, N\}$. The rows and columns correspond to the cross-sectional observations. Its (i, j) -element w_{ij} represents the prior strength of the interaction between spatial units i and j . This can be interpreted as the presence and strength of a link between nodes (the observations) in a network representation that matches the spatial weights structure (Anselin et al., 2008). As such, pairwise dependence structure is of special importance in spatial econometrics.

The simplest model in this field is the spatial autoregressive process:

$$y = \rho W y + \varepsilon \tag{1.1}$$

The term $W y$ is called a spatial lag. It follows that $y = (I_N - \rho W)^{-1} \varepsilon$, where I_N is the identity matrix of size N . Thus, the dependence structure of random vector y is completely determined by $y = (I_N - \rho W)^{-1} \varepsilon$. If $\varepsilon \sim N(0, \sigma^2 I_N)$, as is often assumed, then y is also normally distributed, and it cannot possess neither asymmetric dependence nor tail dependence. It also lacks to incorporate a nonlinear structure.

We would like to build a model in which the disturbance vector is also spatially associated but not through a weight matrix, so that we can capture nonlinear, asymmetric dependence structure in the tail. These kinds of dependence structures have been empirically observed in the finance and insurance literature. To this end, one way is to employ the copula approach. Krupskii et al. (2018) have recently suggested a model based on factor copulas, but it seems somewhat too ad hoc.

2. Copula

Let $F = (x_1, \dots, x_d)$ be a d -dimensional distribution function, and F_i be the i th marginal distribution function of F . According to Sklar's theorem, there exists copula (a distribution function C on $[0; 1]^d$ with uniform marginals) such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad \text{for all } (x_1, \dots, x_d)$$

C is called the copula associated with F , and when F is continuous, the copula associated with F is uniquely determined and is given by

$$F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)), \quad 0 \leq u_i \leq 1, i = 1, \dots, d.$$

Copulas have recently been drawing some attention mainly as a tool to model various dependence among random variables, including the fields of financial risk management and multivariate survival analysis; see Nelsen (2006) and Joe (2015) for an introduction to the topic. Popular bivariate families of copulas are as follows.

- (i) Clayton family

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \theta \in [-1, \infty) \setminus \{0\}$$

- (ii) Gumbel-Hougaard family

$$C_\theta(u, v) = \exp\{-[(-\log u)^\theta + (-\log v)^\theta]^{1/\theta}\}, \quad \theta \geq 1$$

- (iii) Frank family

$$C_\theta(u, v) = \frac{1}{\theta} \log\left(1 + \frac{(e^{\theta u} - 1)(e^{\theta v} - 1)}{e^\theta - 1}\right), \theta \in \mathbb{R}$$

- (iv) Gauss family

$$C_\theta(u, v) = \Phi_\theta(\Phi^{-1}(u), \Phi^{-1}(v))$$

where Φ_θ is the distribution function of bivariate normal distribution with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}$, Φ is the distribution function of univariate standard normal distribution.

(i)-(iii) are examples of well-known Archimedean copula:

$$C(u_1, u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2)), \tag{2.1}$$

where its *generator* $\psi: \mathbb{R}_+ \rightarrow [0,1]$ is a convex and decreasing function satisfying $\psi(0) = 1$ and $\psi(\infty) = 0$.

Considering bivariate case ($d = 2$), one is often interested in dependence between two random variables X_1 and X_2 in the tail of their distributions.

Definition 2.1 The *coefficient of upper tail dependence* between X_1 and X_2 is defined by

$$\lambda_U := \lim_{u \uparrow 1} P(X_2 > F_2^{-1}(u) \mid X_1 > F_1^{-1}(u)),$$

assuming that the limit on the right-hand side exists. Similarly, the *coefficient of lower tail dependence* between X_1 and X_2 is defined by

$$\lambda_L := \lim_{u \downarrow 0} P(X_2 > F_2^{-1}(u) \mid X_1 \leq F_1^{-1}(u))$$

assuming that the limit on the right-hand side exists. When $\lambda_U \in (0; 1]$, X_1 and X_2 are said to be upper tail dependent, and when $\lambda_U = 0$, they are said to be asymptotically upper tail independent. And similarly for λ_L .

If F_1 and F_2 are continuous, these coefficients of tail dependence depend only on the copula C associated with F , and it holds that

$$\lambda_U = \lim_{u \uparrow 1} \frac{\bar{C}(u, u)}{1 - u}, \quad \lambda_L = \lim_{u \downarrow 0} \frac{C(u, u)}{u},$$

where $\bar{C}(u_1, u_2) := 1 - u_1 - u_2 + C(u_1, u_2)$ is a survival copula of C .

- For Frank and Gauss family, $\lambda_U = \lambda_L = 0$;
- For Gumbel-Hougaard family, $\lambda_U = 2 - 2^{1/\theta}, \lambda_L = 0$;
- For Clayton family, $\lambda_U = 0, \lambda_L = 2^{-1/\theta} (\theta > 0), \lambda_L = 0 (\theta \leq 0)$.

3. Methodology

We shall assume that values of dependent variable are observed repeatedly overtime $t = 1, \dots, T$, so that we have a spatial panel structure; see Anselin et al. (2008), and LaSage and Pace (2009).

$$y_t = \rho W y_t + \varepsilon_t, t = 1, \dots, T \tag{3.1}$$

We consider the following copula families for the distribution of ε_t :

(1) Elliptical copula: This is just a copula associated with a nelliptical distribution, and so it has a symmetric dependence structure.

(2) Skew t -copula: Let $Z \sim N_d(\mathbf{0}, \Sigma)$ and $W \sim \text{Ig}(\nu/2, \nu/2)$ be independent ('Ig' stands for inverse Gaussian). Then $X = W\gamma + \sqrt{W}Z$

has a skew t -distribution (Demarta and McNeil, 2005). The associated copula is called a *skew t -copula*. $\gamma = (\gamma_1, \dots, \gamma_d)'$ is a parameter representing the degree of "distortion".

(3) Nested Archimedean copula: The simplest extension of Archimedean copula (2.1) is given by $C(u_1, \dots, u_d) = \psi(\psi^{-1}(u_1) + \dots + \psi^{-1}(u_d))$, and its dependence structure is too symmetric. Thus the following nested Archimedean copulas have been introduced in the literature.

- *Fully nested Archimedean copula:* Starting with $C(u_1, u_2, \psi_1) := \psi_1(\psi_1^{-1}(u_1) + \psi_1^{-1}(u_2))$, define recursively for $d \geq 3$,

$$C(u_1, \dots, u_d; \psi_1, \dots, \psi_{d-1}) := \psi_1(\psi_1^{-1}(u_1) + \psi_1^{-1}(C(u_2, \dots, u_d; \psi_2, \dots, \psi_{d-1})))$$
- *Partially nested Archimedean copula*

$$C(\mathbf{u}) = C(C(u_{1,1}, \dots, u_{1,d_1}; \psi_1), \dots, C(u_{p,1}, \dots, u_{p,d_p}; \psi_p); \psi_0)$$

$$= \psi_0 \left(\psi_0^{-1} \circ \psi_1 \left(\psi_1^{-1}(u_{1,1}) + \dots + \psi_1^{-1}(u_{1,d_1}) \right) + \dots + \psi_0^{-1} \circ \right.$$

$$\left. \psi_p \left(\psi_p^{-1}(u_{p,1}) + \dots + \psi_p^{-1}(u_{p,d_p}) \right) \right)$$

See Joe (1997) or McNeil (2008) for exposition. Simple examples are

$$C(u_1, u_2, u_3) = \psi_1 \left(\psi_1^{-1}(u_1) + \psi_1^{-1} \circ \psi_2(\psi_2^{-1}(u_2) + \psi_2^{-1}(u_3)) \right),$$

$$C(u_1, u_2, u_3, u_4) = \psi_1 \left(\psi_1^{-1} \circ \psi_2(\psi_2^{-1}(u_1) + \psi_2^{-1}(u_2)) + \psi_1^{-1} \right.$$

$$\left. \circ \psi_3(\psi_3^{-1}(u_3) + \psi_3^{-1}(u_4)) \right).$$

Because of its apparent rooted tree structure, this class of copulas is suited for spatial data; prior knowledge on the spatial structure among individuals can be utilized to construct the tree structure of nested Archimedean copula. Conversely, one can in fact apply the method in Segers and Uyttendaele (2014) to estimate the tree structure and thereby find which kind of dependence are left over by the standard spatial autoregression.

3.1. Analyzing Procedure

We suggest the following procedure for the statistical analysis. Note that all three copula families we discussed above are parametric. For simplicity suppose that ε_t 's are i.i.d. We tacitly assume that the model (3.1) with $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_N)$ has been fitted, but the residuals do not show i.i.d. behavior.

- By looking at 2-by-2 scatter plot or any other means, we need to check whether significant tail dependence and/or asymmetric dependence is found. If there is, then elliptical copula could be eliminated from the candidate copulas.
- For all the above model (1)–(3), we can compute either maximum likelihood or pseudo-likelihood estimates (Genest et al., 1995), depending on the assumption on the one-dimensional marginals. Use AIC (or some other information criteria) to search for the best-fitted

copula. Alternatively, we can use the distance between the fitted copula and empirical copula. For nested Archimedean copulas, Clayton or Gumbel-Hougaard copulas are favorable candidate with tail dependence.

- (iii) Diagnostic analysis: We can carry out some goodness-of-fit test for copula using resampling techniques (Fermanian, 2013). Some graphical diagnostic method would be desirable.

We will present some results of (ongoing) empirical analysis of financial data.

4. Discussion

To incorporate heterogeneity, exogenous explanatory variables X with regression coefficients vector β could be introduced in the model. One could consider vine copulas as well although their interpretation is not easy.

References

1. L. Anselin, J. Le Gallo, and H. Jayet. Spatial panel econometrics. In L. Mátyás and P. Sevestre, editors, *The Econometrics of Panel Data*, pages 625–660. SpringerVerlag, Berlin Heidelberg, 2008.
2. S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review*, 73:111–129, 2005.
3. J.-D. Fermanian. An overview of the goodness-of-fit test problem for copulas. In P. Jaworski, F. Durante, and W. K. Härdle, editors, *Copulae in Mathematical and Quantitative Finance*, pages 61–89. Springer-Verlag Berlin Heidelberg, 2013.
4. C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82: 543–552, 1995.
5. H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997.
6. H. Joe. *Dependence Modeling with Copulas*. CRC Press, Boca Raton, 2015.
7. P. Krupskii, R. Huser, and M. G. Genton. Factor copula models for replicated spatial data. *Journal of the American Statistical Association*, 113:467–479, 2018.
8. J. LaSage and R. K. Pace. *Introduction to Spatial Econometrics*. Chapman & Hall/CRC, Boca Raton, 2009.
9. J. McNeil. Sampling nested archimedean copulas. *Journal of Statistical Computation and Simulation*, 78:567–581, 2008.
10. R. B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York, second edition, 2006.
11. J. Segers and N. Uyttendaele. Nonparametric estimation of the tree structure of a nested archimedean copula. *Computational Statistics and Data Analysis*, 72: 190–204, 2014.



Spatial extension of GARCH Models for high-dimensional financial time series



Takaki S.¹, Yasumasa Matsuda²

¹Advanced Institute for Yotta Informatics, Tohoku University, Sendai, Japan

²Graduate School of Economics and Management, Tohoku University, Sendai, Japan

Abstract

Autoregressive Conditional Heteroscedasticity (ARCH) models, which were originally proposed by Engle (1982), have been playing major roles in modeling volatilities in financial time series. The purpose of this study is a multivariate extension of ARCH models to evaluate volatility matrices for high dimensional multivariate financial time series. The critical difficulty in the multivariate extension is in the so-called curse of dimension caused by a larger number of parameters for a higher dimension of multivariate series. We introduce financial distances among components of multivariate series, which are different from the usual physical one but are based on the closeness of financial conditions, and apply dynamic panel data models by spatial weight matrices constructed by the financial distance. As a result, we propose spatial autoregressive moving average models with generalized autoregressive conditional heteroskedasticity processes (SARMA-GARCH models) that can identify volatility matrices for high dimensional financial time series. We conduct comparative studies by real financial time series and show empirical features of the SARMA-GARCH models in terms of the forecast of volatilities.

Keywords

Volatility model; Spatial weight matrix; High-dimensional statistics

1. Introduction

Volatility which is a conditional variance in a model is one of the most important concepts in financial econometrics because it is used in widely areas such as risk management, option pricing and portfolio selection. Financial market data often exhibits volatility clustering (i.e., volatility may be high for certain time periods and low for other periods) This means time-varying volatility is more common than constant volatility. Therefore, accurate modeling of time-varying volatility is important in financial econometrics.

The seminal work of Engle (1982) proposes autoregressive conditional heteroscedasticity (ARCH) models and the most important extension of the model is generalized ARCH (GARCH) models proposed by Bollerslev (1986). The models have been widely used to identify volatilities. After that, many extended GARCH models have been proposed. For example, integrated GARCH models (Engle and Bollerslev (1986)), exponential GARCH models

(Nelson (1991)), threshold GARCH models (Glosten, et al (1993)), GARCH in the mean models, and GJR-GARCH models are proposed.

Univariate volatility models are generalized to multivariate cases in many ways. One important problem which multivariate volatility models contain is the curse of dimensionality. We estimate a conditional covariance matrix which has $\frac{n(n+1)}{2}$ quantities for a n-dimensional time series in multivariate analysis. However it is difficult to estimate all quantities. Thus, we attempt to give a conditional covariance matrix some simple structures to reduce the number of parameters. There are many ways for generalization. For example, exponentially weighted moving average models, constant conditional correlation models (Bollerslev (1990)), BEKK models (Engle and Kroner (1995)), orthogonal GARCH models (Alexander (2001)), dynamic conditional correlation models (Tse and Tsui (2002)), dynamic orthogonal component models, and factor GARCH models are proposed.

The ideas of spatial econometrics have been applied to volatility models to reduce number of parameters in a covariance matrix and to extend the models to spatial models in recent years. Caporin and Paruolo (2008) and Borovkova and Lopuhaa (2012) have applied the ideas of spatial econometrics to time series multivariate GARCH models. Yan (2007) and Robinson (2009) have done spatial extensions of stochastic volatility models which are another kind of volatility models. Sato and Matsuda (2017, 2018) have extend time series GARCH models to spatial models.

This paper contributes to extend GARCH models to spatiotemporal models for high dimensional financial time series which we call spatial autoregressive moving average models with generalized autoregressive conditional heteroskedasticity processes, namely SARMA-GARCH models by using spatial econometrics ideas.

The model is characterized by a spatial weight matrix which express cross-section correlations between assets and used to reduce the number of parameters. A spatial weight matrix is usually determined by geographical information of spatial data. However, financial data doesn't include geographical information, therefore we need to consider a method to make spatial weight matrix from financial data. Here, we apply the multiple linear regression model to return series of assets to calculate spatial weight matrices with stepwise model selection procedures for selecting subsets of explanatory variables in the regression model.

Parameters in the SARMA-GARCH model are estimated by a two step procedure. First step is the estimation of spatial parameters and second step is the estimation of GARCH parameters. Spatial parameters which are scalar parameters reflecting the strength of spatial dependence between assets are estimated in first step. Conditional variances in the model follows GARCH

processes and unconditional variances in GARCH processes are constant. Therefore, we regard volatilities in the model as constant and apply the quasi-maximum likelihood method with a spatial panel model with heterogeneous constant variances. GARCH parameters are scalar parameters for specifying volatility behaviors. We calculate residuals by fitting the spatial panel model in first step after that we apply GARCH models with the residuals of each asset.

In real data analysis, We apply the SARMA-GARCH model to daily returns of the S&P 500 stock price data. We compare in-sample and out-sample performances of the SARMA-GARCH model with those of CCC models which is a multivariate volatility model and a baseline model in this study. First, we check the in-sample performances based on log-likelihood. The results show the log-likelihood of the CCC model is greater than that of the SARMA-GARCH model. This means model fitting of the CCC model is better than the SARMA-GARCH model because the number of parameters in CCC models is more than five times of those of SARMA-GARCH models. Secondly, we compare out-sample performances by using the quasi-likelihood loss function. The result shows the quasi-likelihood loss function of the SARMA-GARCH model are smaller than that of CCC models. Then, the out-sample performance of SARMA-GARCH models is better. This is because the CCC model may be over-fitting and it cause lower forecasting performances.

Moreover, stock prices in the U.S. market are volatile and correlation structures between stocks may change over time, so the characteristic of the SARMA-GARCH model which is that the model can capture dynamic correlation between assets may play an important role in analysis.

The rest of paper proceeds as follows. Section 2 introduces SARMA-GARCH models. The estimation procedures are described in section 3. Section 4 examines empirical properties of SARMA-GARCH models by applying the models to real data such as stock price in the U.S. market. Section 5 discusses some concluding remarks.

2. Methodology

Let $\{r_{i,t}\}, i = 1, \dots, N$ and $t = 1, \dots, T$, be return series of financial instruments. We shall define SARMA-GARCH models to describe volatilities of return series by

$$r_t = \lambda W r_t + u_t \tag{1}$$

$$u_t = \rho W u_t + \varepsilon_t \tag{2}$$

$$\varepsilon_{i,t} = \sigma_{i,t} f_{i,t}, \tag{3}$$

$$f_{i,t} \sim i.i.d(0,1),$$

$$\sigma_{i,t}^2 = \omega_i + \alpha_i \sigma_{i,t-1}^2 + \beta_i \varepsilon_{i,t-1}^2,$$

where $r_t = (r_{1,t}, \dots, r_{n,t})$, $\varepsilon_t = (\varepsilon_{1,t}, \dots, \varepsilon_{n,t})$, $\sigma_{i,t}$ is volatility, $f_{i,t}$ is an independent and identically distributed (i.i.d) random variable with mean zero and variance 1. The matrix W , n by n matrix, is called a spatial weight matrix

and pre-determined before analysis. For parameters $(\rho, \lambda, \omega_i, \alpha_i, \beta_i)'$, spatial parameters, ρ and λ , describes spatial interactions of return series and ω_i, α_i and β_i are GARCH parameters and specify volatility behaviors. The positivity of $\sigma_{i,t}^2$ is ensured by the following sufficient restrictions $\omega_i > 0, \alpha_i \geq 0, \beta_i \geq 0$, and the sum $\alpha_i + \beta_i < 1$ for stationarity. Moreover, we assume $|\lambda| + |\rho| < 1$ to guarantee the existence of the model.

Let us consider the volatility matrix for SARMA-GARCH models. From (3), the variance matrix of ε_t, Γ_t , is a diagonal matrix whose components are $\sigma_{i,t}^2$ that is $\Gamma_t = \text{diag}(\sigma_{1,t}, \dots, \sigma_{n,t})$. From equations (1) and (2),

$$r_t = (I - \lambda W)^{-1}(I - \rho W)^{-1}\varepsilon_t.$$

Therefore, the volatility matrix for SARMA-GARCH models, Σ_t , are

$$\Sigma_t = (I - \lambda W)^{-1}(I - \rho W)^{-1}\Gamma_t(I - \rho W)^{-1}(I - \lambda W)^{-1}, \quad (4)$$

where I is an identity matrix. Volatility $\sigma_{i,t}^2$ changes over time, so the volatility matrix for SARMA-GARCH models expresses time-varying volatility structures in financial instruments and can capture dynamic correlations. Moreover, the spatial weight matrix in (4) expresses cross-sectional correlation between observations and plays important role to reduce the number of parameters for cross-sectional correlation and to overcome the curse of dimensionality.

A spatial weight matrix is usually determined by geographical information of spatial data and predetermined such as first-order contiguity relation or inverse distance between observations. However, $\{r_{i,t}\}$ are financial data and don't include geographical information. Therefore, we need to determine financial distances to make a spatial weight matrix. Some author have proposed spatial weight matrix based on financial distance calculated from financial statement data such as dividend yields or market capitalizations. Here, we propose a method to make spatial weight matrices by multiple regression models with backward stepwise model selection procedures. It begins with the full least squares model containing all $n - 1$ explanatory variables.

$$r_{i,t} = \delta_0 + \sum_{i \neq j}^n \delta_j r_{j,t} + z_{i,t},$$

where $z_{i,t}$ follows an i.i.d standard normal distribution. Then, we iteratively removes the least useful explanatory variables, one-at-a-time. Details are given as follows:

1. Apply the OLS methods to the full model and obtain t-values for explanatory variables.
2. Remove the explanatory variable whose t-value is the minimum values in all t-values and not statistically significant.
3. Apply the OLS methods to the model contains all but one of the explanatory variables in step 2.

We repeat steps 2-3 until the minimum t-value is greater than a critical value, for example 1.96.

3. Estimation

We shall propose estimation of the parameters $(\rho, \lambda, \omega_i, \alpha_i, \beta_i)'$ in SARMA-GARCH models. Parameters are estimated by a two step procedure. First step is the estimation of λ and ρ and second step is that $\omega_i, \alpha_i, \beta_i$.

Now, let us derive quasi likelihood function by regarding $f_{i,t}$'s as Gaussian variables with mean zero and variance $\sigma_{i,t}^2$. Then, the likelihood function of SARMA-GARCH models is

$$\log L = T \log |I - \lambda W| + T \log |I - \rho W| + \sum_{t=1}^T \sum_{i=1}^n \left(-\frac{1}{2} \log 2\pi \sigma_{i,t}^2 - \frac{\varepsilon_{i,t}^2}{2\sigma_{i,t}^2} \right),$$

where $\varepsilon_{i,t}^2$ is the i-th element of $(I - \rho W)(I - \lambda W)r_t$. Here, the number of parameters are $3n + 2$ and optimization of all parameters simultaneously is a difficult task, so we adopt a two step procedure to reduce the number of parameters.

Parameters ρ and λ are estimated in first step. The parameters are estimated by the quasi-likelihood estimation method. Here, we regard $\sigma_{i,t}^2$ as constant heteroskedastic variances because GARCH processes are stationary processes, namely variances in the model are different according to assets but don't change over time. Gaussian likelihood function for first step estimation is derived by regarding $f_{i,t}$ as independent Gaussian variables with mean zero and variance σ_i^2 . Then the log likelihood function is

$$\log L = T \log |I - \lambda W| + T \log |I - \rho W| + -\frac{T}{2} \sum_{i=1}^N \log 2\pi \sigma_i^2 \sum_{t=1}^T \sum_{i=1}^n \left(\frac{\varepsilon_{i,t}^2}{2\sigma_i^2} \right) \quad (5)$$

The QML estimator $\hat{\lambda}$ and $\hat{\rho}$ maximizes the log likelihood function (5).

We move to estimation of GARCH parameters. We have already obtained estimate of spatial parameters, λ and ρ . The residuals are obtained by

$$\hat{\varepsilon}_t = (I - \hat{\rho}W)(I - \hat{\lambda}W)r_t,$$

where $\hat{\lambda}$ and $\hat{\rho}$ are estimates of spatial parameters in first step. we apply the GARCH (1, 1) model to the residuals. Let $f_{i,t}$ in (3) be Gaussian white noise with unit variance. Then $\varepsilon_{i,t}$ is an GARCH (1, 1) process if

$$\begin{aligned} \varepsilon_{i,t} &= \sigma_{i,t}^2 f_{i,t} \\ \sigma_{i,t}^2 &= \omega_i + \alpha_i u_{i,t-1}^2 + \beta_i \sigma_{i,t-1}^2. \end{aligned}$$

Here, we use residuals $\hat{\varepsilon}_{i,t}$ in stead of $\varepsilon_{i,t}$. Then, the conditional log likelihood function of GARCH models is given by

$$\log L = \sum_{t=2}^T \left\{ -\frac{1}{2} \log(2\pi \sigma_{i,t}^2) - \frac{\hat{\varepsilon}_{i,t}^2}{2\sigma_{i,t}^2} \right\},$$

where $\sigma_{i,t}^2 = \beta_{i,0} + \beta_i u_{i,t-1}^2 + \beta_{i,2} \sigma_{i,t-1}^2$ can be evaluated recursively. Maximizing this with respect to ω_i, α_i and β_i , we have the estimators of GARCH parameters.

4. Real data analysis

We examine empirical properties of SARMA-GARCH models by applying daily return data of the U.S markets to demonstrate practical performances of volatilities and co-volatilities identified by SARMA-GARCH models. Moreover, we show prediction performance and dynamic spillover effect of shock.

We apply the SARMA-GARCH model to daily returns of the S&P 500 stock price data, that is the returns $\{r_t\}$ are computed as $100(\log P_t - \log P_{t-1})$, where P_t is the closing price and t is the time index referring to trading day t . The sampling period starts on April 1st, 2002 and ends on July 4th, 2016 for a total of 3500 returns. Moreover, we sample data for prediction from July 5th, 2016 to December 30, 2016. The number of firms are 395. Spatial wight matrices are made in accordance with the manner written in section 2. Here, the critical value is 1.96.

We adopt constant conditional correlation (CCC) models as a benchmark. Let $r_t = (r_{1,t}, \dots, r_{n,t})$ be a n -dimensional vector process. CCC models are represented by the following equations

$$\begin{aligned} r_t &= \sum_t^{\frac{1}{2}} \varepsilon_t, \\ \Sigma_t &= \text{diag}(\sigma_{1,t}^2, \dots, \sigma_{n,t}^2), \\ \sigma_{i,t} &= \omega_i + \alpha_i r_{i,t-1}^2 + \beta_i \sigma_{i,t-1}^2, i = 1, \dots, n \end{aligned}$$

where Σ_t is a diagonal matrix with $\sigma_{i,t}^2$ as i th diagonal element, and ε_t unobservable random vector with mean equal to 0 and variance-covariance equal to $R_t = (\rho_t, i, j)$. CCC models assume the correlation matrix is constant.

Table 1 shows the estimated values of λ and ρ . Estimates of α_i and β_i are in the ranges [0.01, 0.59] and [0.27, 0.98], respectively. We find that $\hat{\lambda}$, the strength of interactions among return series, are significant. This suggests that asset returns tend to move together strongly.

Table 1: Estimated values of λ, ρ and GARCH parameters and their standard errors (s.e.) of λ and ρ in the SARMA-GARCH model applied to log returns of stock price data of the U.S financial market.

parameter	estimate	s.e
λ	0.9199	0.0006
$\hat{\rho}$	-0.3200	0.0017
$\hat{\alpha}_i$	[0.01, 0.59]	
$\hat{\beta}_i$	[0.27, 0.98]	

Table 2: Log-likelihoods and quasi-likelihood loss functions for the SARMA-GARCH model and the CCC model applied to log returns of stock price data of the U.S. financial market.

	in-sample log-likelihood	out-sample QLIKE
SARMA-GARCH	556	414
CCC	534	455

We compare the in-sample and out-sample performances of SARMA-GARCH models with those of CCC models. First, we check the in-sample performances based on log-likelihood. Table 2 shows the log-likelihood of CCC is bigger than that of SARMA-GARCH. This means model fitting of the CCC model is better. One reason is that the number of parameters in CCC models is more than five times of those of SARMA-GARCH models. Secondly, we compare out-sample performances. We calculate predicted volatility based on definition of the models. After that we calculate prediction error based on the quasi-likelihood loss function:

$$QLIKE = \frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} r_t' V_t^{-1} r_t + \log|V_t|,$$

where r_t is a vector of return series V_t is a volatility matrix made by predicted volatility and T_{pre} is the size of time dimension for prediction period. Table 2 shows out-sample performance of SARMA-GARCH models are better. This shows CCC models may be over-fitting and it cause lower forecasting performance. Moreover, CCC models which assume constant correlation between stock prices can't capture dynamic relations, but SARMA-GARCH models can capture dynamic correlation as volatility matrix. Therefore, the out-sample performance of the SARMA-GARCH model is better.

5. Conclusion

We have proposed a spatial autoregressive moving average models with generalized autoregressive conditional heteroskedasticity processes, namely SARMA-GARCH models to evaluate volatilities of financial instruments. We apply spatial weight matrices which is an important tool in spatial econometrics for multivariate volatility models to overcome the curse of dimensionality. we propose a two step procedure to estimate the parameters in SARMA-GARCH models. In the real data analysis of the U.S. markets, we detect SARMA-GARCH models have smaller prediction error than that of CCC models.

We complete the paper by describing challenging problem for future research. In the empirical analysis, we used the spatial weight matrix based on least-squares estimates with backward stepwise model selection procedures. However, The choice of spatial weight matrix is an important problem in

spatial analysis for financial data. Therefore, another spatial weight matrix can be more interesting to improve our volatility analysis. Another challenge is to establish asymptotic properties of estimators for the SARMA-GARCH model to investigate theoretical properties of proposed estimators.

References

1. Alexander, C. (2001). *Market models: A guide to financial data analysis*. John Wiley & Sons.
2. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307-327.
3. Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. *The review of economics and statistics*, 498-505.
4. Borovkova, S. & Lopuhaa, R., (2012). Spatial GARCH: A Spatial Approach to Multivariate Volatility Modeling. Available at SSRN: <https://ssrn.com/abstract=2176781>.
5. Caporin, M., & Paruolo, P., 2008, Structured multivariate volatility models, Available at SSRN: <http://ssrn.com/abstract=1318639>.
6. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.
7. Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric reviews*, 5(1), 1-50.
8. Engle, R. F., & Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric theory*, 11(1), 122-150.
9. Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5), 1779-1801.
10. Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347-370.
11. Robinson, P. M. (2009). Large sample inference on spatial dependence. *The Econometrics Journal*, 12(s1).
12. Sato, T. & Matsuda, Y. (2017). Spatial Autoregressive Conditional Heteroskedasticity Models. *J.Japan Statist. Soc.*, Vol. 47, 2
13. Tse, Y. K., & Tsui, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics*, 20(3), 351-362.
14. Yan, J. (2007). Spatial stochastic volatility for lattice data. *Journal of agricultural, biological, and environmental statistics*, 12(1), 25.



Analysis of regional economic growth against crisis: US-Japan statistical comparative study before-after Lehman's shock



Gigih Fitrianto¹, Shojiro Tanaka², Ryuei Nishii³

¹ Graduate School of Economics, Hiroshima University of Economics, Hiroshima, Japan

² Professor, Faculty of Media Business, Hiroshima University of Economics, Hiroshima, Japan

³ Professor, Office for Establishment of an Information-related School, Nagasaki University, Nagasaki, Japan

Abstract

This paper applied several statistical methods for detecting structural change, such as: F-test and Wald-test based on broken time-trend model (Greene, 2012) to analyze the effect of Lehman's crisis towards regional economic growth in US and Japan. This paper also observed the spatial neighboring factor toward the impact of crisis across regional economies. F-test performed better than Wald test for spatially-independent model. On the other hand, in spatially-dependent model Wald test produced better results. It endorsed the endogeneity (explanatory variable correlated with error term) in spatially-dependent model might imply the F-test become non-robust for autocorrelated disturbances (Krämer, 2003). In the US, it was revealed that the negative impact of crisis clustered in West Coast, Southeastern, and Great Lake regions. States in those regions with good manufacturing, construction, insurance, and finance as the main contributor for their gross regional product. On the other hand, states that relied on agriculture, forestry, fishing, hunting, and mining sectors had resilience, the results of which were obtained by the tests. In Japan, it was confirmed that the shock spread-out across all the regions. After the crisis, most prefectures showed recoveries. Several other prefectures showed negative trend, such as Nagasaki.

Keywords

Neighbor Effect; Spatial Analysis; Structural Change

1. Introduction

One of the biggest economic crises, the Lehman's shock began the fall of US housing prices in 2007 which cause sub-prime crisis. These defaults spread-out to the financial sector (Reinhart & Rogoff, 2008) and interbank markets from August 2007 (BIS Report, 2009). The crisis was transmitted globally from two main channels: global financial interconnections (Longstaff, 2010; Aloui, Aissa, & Nguyen, 2011); and trade flows (Ahn, Amity, & Weinstein, 2011; Cetorelli & Goldberg, 2011).

In Japan, the impact of Lehman's crisis heavily damaged the Japan economy from 2008. In 2008, Japan had a trade deficit and in 2009 the export

values reported declined from -3.5% in 2008 to -33.1% in 2009 (MOF Press Release, 2012).

Kawai and Takagi (2009) have shown that this occurred by the increasing of Japan's export to gross domestic product (GDP) ratio and trade openness. Due to the declining of demand from US, then Japan had faced lower total export. This decline is also affected financial institutions' core profitability from trade credit and financing as the regional capital is not kept up the increasing off risk asset (BOJ Financial System Report, 2008). Demirer et al. (2017) also show that the Japanese banks have strong connection globally impacted the spread volatility across bank stocks, especially with US banks.

In this research, we conducted further investigation for the impact of Lehman's shock towards US and Japan regional economic growth. By using this approach, we got a detailed illustration of how the shock transmitted within country.

2. Methodology

2.1 Data and Target Variable

We used the real gross regional product (GRP) based on production approach as the main dataset. The United States data were retrieved from regional statistics in Bureau Economic Analysis (BEA) and we apply quarterly data from 1st Quarter of 2005 until 2nd Quarter of 2018 for 51 states (including District of Columbia). On the other hand, for Japan we use annually GRP data from Economic and Social Research Institute, Cabinet Office for 47 prefectures from 2001 until 2014.

Based on Reinhart & Rogoff (2008) and Bank for International Settlement (BIS) Report (2009) we select the 1st quarter of 2008 as the breakpoint in the US. On the other hand, following Filardo et al. (2009) and Kawai & Takagi (2009), we select 2008 as the breakpoint for Japan regional data.

2.2 Empirical Model

2.2.1 Spatially-Independent Model

Our focus in this research is to observe the behavior of regional economic growth fluctuation. This approach allowed us to identify before and after shock conditions and observed which area had a recovery. Therefore, we analyze the regional transitory, by observing the time-trend of regional economic growth before and after crisis by using ordinary least-square (OLS) regression (Greene, 2012),

$$\ln\left(\frac{y_{i,t}}{y_{i,t-1}}\right) \equiv g_{i,t} = \mu_i + \tau_i t + (\mu_i^* + \tau_i^* t)I(t > t^*) + u_{i,t}, \quad u_{i,t} \sim iid N(0, \sigma_i^2) \text{ for } t \leq t^*$$

$$u_{i,t} \sim iid N(0, \sigma_i^{*2}) \text{ for } t > t^* \quad (1)$$

$y_{i,t}$ is GRP for states $i = 1, \dots, N$ at time period $t = 1, \dots, T$ and $\ln(y_{i,t}/y_{i,t-1})$ is regional economic growth, and t^* represents exogenously selected breakpoint. Here, μ_i are intercept and τ_i is time trend component before crisis. Similarly, $(\mu_i + \mu_i^*)$ and $(\tau_i + \tau_i^*)$ are defined after crisis.

Based on those models in (1), we conducted several methods for structural analysis: 1) F-Test; and 2) Wald-test. These methods use exogenously selected breakpoint, t^* . The F-test calculated based on a simple regression model for time trend with known t^* ,

$$F = \frac{[S_1 - (S_2 + S_3)]/K}{(S_2 + S_3)/(n_1 + n_2 - 2K)} \quad (2)$$

n_1 is the number of observations from before crisis group ($t = 1, \dots, t^*$), n_2 is the number of observations for after crisis group ($t = t^* + 1, \dots, T$). S_1 is sum of square residual from combined data ($n_1 + n_2$), S_2 and S_3 are sum of squared residual from n_1 and n_2 respectively. K is the number of parameters.

To provide further evidence for structural change at breakpoint t^* , we also adopt the Jouini & Boutahar (2004) and Greene (2012) approach by applying the Wald test,

$$W = (\hat{\theta}_1 - \hat{\theta}_2)'(\hat{V}_1 + \hat{V}_2)^{-1}(\hat{\theta}_1 - \hat{\theta}_2) \sim \chi_2^2 \quad (3)$$

where $\theta_1 = (\mu_i, \tau_i)'$ and $\theta_2 = (\mu_i + \mu_i^*, \tau_i + \tau_i^*)'$ are used for spatially-independent model in (1). \hat{V}_1 and \hat{V}_2 are asymptotic covariance matrices under null hypothesis: $\theta_1 = \theta_2$ and alternative hypothesis: $\theta_1 \neq \theta_2$ respectively.

2.2.2 Spatially-Dependent Model

To include the neighboring relations as a factor for the analysis, we extend equation (1) into spatial autoregressive model as follows,

$$g_{i,t} = \mu_i + \tau_i t + \rho_i \sum_{j \neq i} w_{ij} g_{j,t} + (\mu_i^* + \tau_i^* t + \rho_i^* \sum_{j \neq i} w_{ij} g_{j,t}) I(t > t^*) + u_{i,t},$$

$$u_{i,t} \sim N(0, \sigma_i^2) \text{ for } t \leq t^*$$

$$u_{i,t} \sim N(0, \sigma_i^{*2}) \text{ for } t > t^* \quad (4)$$

where ρ_i and $\rho_i + \rho_i^*$ represents the spatial regression coefficients before and after crisis. The spatial relationship calculated by conditions $w_{ii} = 0$, and $w_{ij} = 1$ if region i and j shares a common border, otherwise $w_{ij} = 0$. Due to this condition, then Alaska, Hawaii, and Okinawa became to have no neighbor.

We conducted similar structural change test by using F-Test and Wald test. Equation (4) include spatial factor ρ_i into the analysis, then we update the Wald test into,

$$W = (\hat{\theta}_{1,s} - \hat{\theta}_{2,s})'(\hat{V}_{1,s} + \hat{V}_{2,s})^{-1}(\hat{\theta}_{1,s} - \hat{\theta}_{2,s}) \sim \chi_3^2 \quad (5)$$

where $\theta_{1,s} = (\mu_i, \tau_i, \rho_i)'$ and $\theta_{2,s} = (\mu_i + \mu_i^*, \tau_i + \tau_i^*, \rho_i + \rho_i^*)'$. $\hat{V}_{1,s}$ and $\hat{V}_{2,s}$ are asymptotic covariance matrices under equation (4).

3. Result

3.1 Breakpoint Analysis of US Regional Data

3.1.1 By Spatially-Independent Model

Based on the assumption that the Lehman’s shock begun on the 1st Quarter of 2008 in the U.S, we will have several states that have negative impacts such as Virginia. On the other hand, South Dakota show resilience by positive growth rates before and after crisis. Even though North Dakota have positive trend of regional economic growth before crisis, but it showed declines after crisis. The data size before the crisis $n_1 = 12$, after the crisis $n_2 = 41$, respectively.

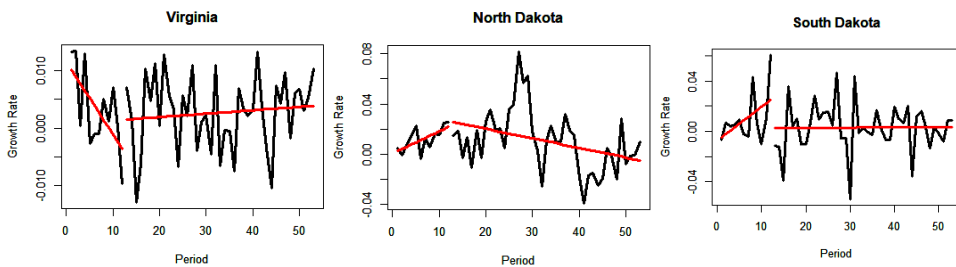


Fig 1. Regional Economic Growth Rate in U.S: Before and After Crisis

Virginia had a sharply declining trend before the Lehman’s shock. This state was suffered due to the declines in nondurable goods manufacturing and housing market. In **Fig 2**, the Southeastern, Great Lake, and West Coast region shows negative impact from the crisis. The main contributors to their GRP are goods manufacturing, construction, insurance, and finance (BEA Report, 2009).

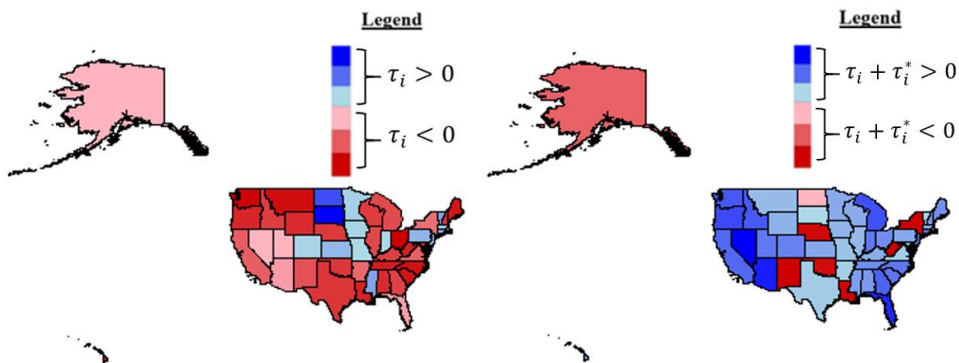


Fig 2. The regional trend comparison before (left) and after (right) breakpoint in US

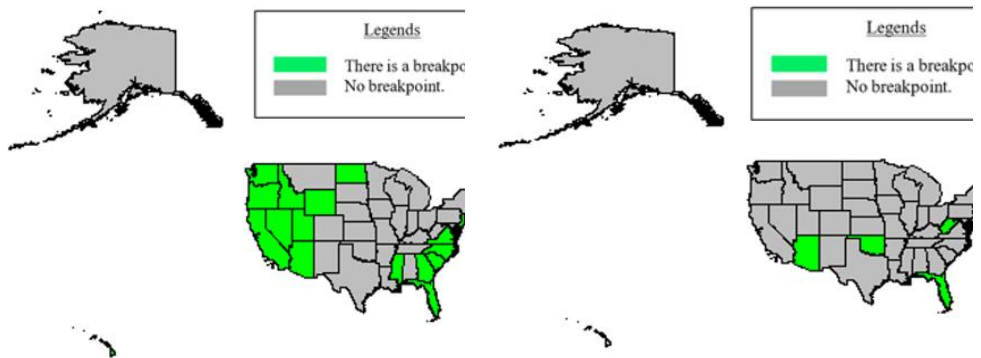


Fig 3. The F-test (left) and Wald Test (right) for Spatially-Independent model.

On the other hand, BEA (2009) also reported that several states in Plains and Rocky Mountain regions such as Kansas, Colorado, Iowa, South Dakota, and Missouri show resilience. Those states are mainly have agriculture, forestry, fishing, hunting, and mining sectors as the main contributor.

The F-test in **Fig 3** shows that mainly in the Southeastern and Far West region that observed the structural changes. Those states have good manufacturing, insurance, finance, and construction as their main contribution to their GRP.

3.1.2 By Spatially-Dependent Model

Based on **Figs 2** and **3**, we clearly observed the neighboring effects in the spread-out of Lehman’s shock in US regional economic growth. The negative trend clustered in several regions, such as Far West, Southeastern, and Great Lake regions.

Fig 4 shows the positive $\hat{\rho}_i$ that indicates there is correspondence between spatial neighboring factor and the effects of Lehman’s crisis in **Fig 2**. The positive $\hat{\rho}_i$ for states in Far West, Southeastern, and Great Lake indicates that the negative trend in the neighboring region affected the growth trend in those regions.

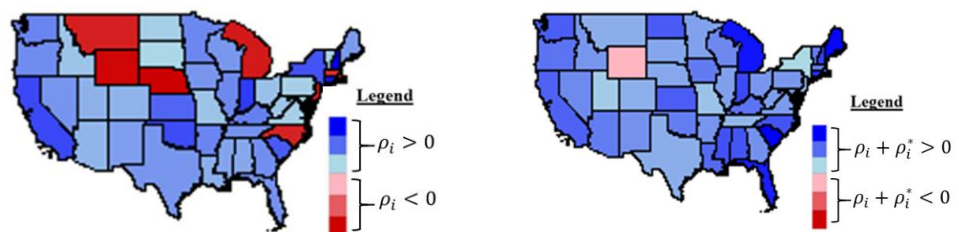


Fig 4. The spatial correlation before (left) and after (right) breakpoint in US

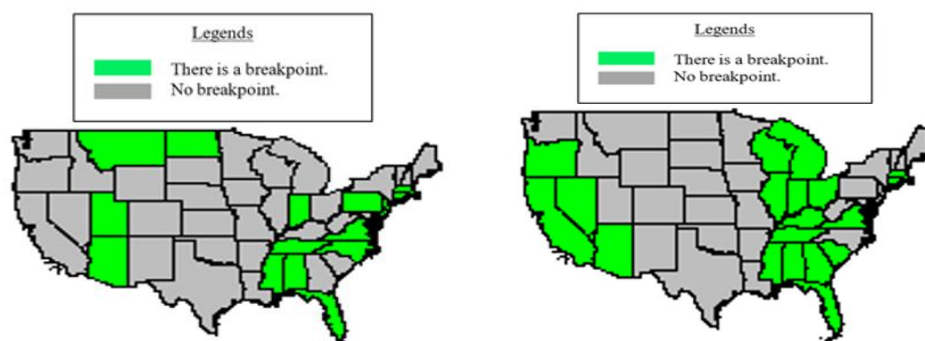


Fig 5. The F-test (left) and Wald Test (right) for Spatially-Dependent model.

As the next finding, we observed that the positive spatial correlation after crisis is higher than before crisis (see in **Fig 4**). Connecticut, Florida, Michigan, Maine, and South Carolina are states that have high spatial correlation ($\hat{\rho}_i + \hat{\rho}_i^*$) after crisis which indicates the recoveries from neighboring states have effect toward those states' regional economic recoveries. Wyoming become the only state that has $(\hat{\rho}_i + \hat{\rho}_i^*) < 0$.

Fig 5 shows the evidence that supports the structural break in US. The Wald test results shows more appropriate test compared to the F-test. The endogeneity (explanatory variable correlated with error term) in spatially-dependent model might imply the F-test become non-robust for autocorrelated disturbances (Krämer, 2003). On the other hand, the Wald test produce more appropriate results hence the test focus to evaluate the distance between the estimated parameters in constrained and unconstrained form (Greene, 2012).

3.2 Breakpoint Analysis of Japan Regional Data

3.2.1 By Spatially-Independent Model

By using 2008 as the breakpoint, Kagawa prefecture shows positive trend before crisis occurred (see **Fig 6** and **Fig 7**). However, this exceptional case was occurred due to the impact of the massive increase in manufacturing production for Kagawa in 2007 based on Bank of Japan (BOJ) Report (2008). The other prefecture showed a negative trend before crisis, such as Tokyo prefecture. However, we observed that Tokyo shows a strong recovery for regional economic growth after crisis.

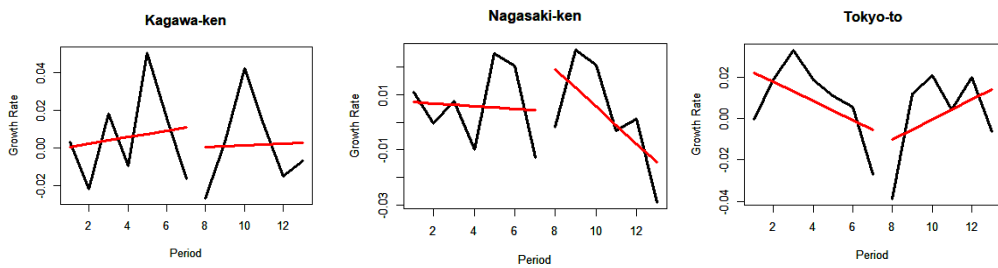


Fig 6. Regional Economic Growth Rate in Japan: Before and After Crisis

BOJ (2008) reported the declining in manufacturing production, business investment, and private consumption weakens regional economic growth. As the production and consumption declines, we had a further declining impact of the crisis. This spread-out affected almost all prefectures in Japan.

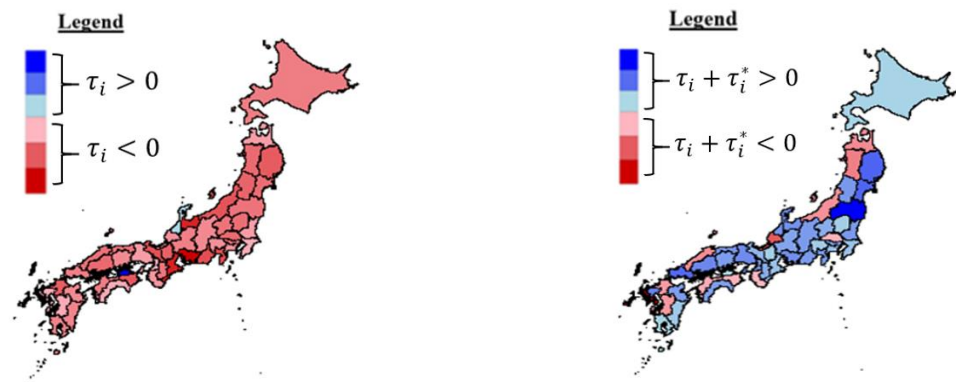


Fig 7. The regional economic trend comparison before (left) and after (right) breakpoint in Japan.

Based on **Fig 7**, most prefectures show recovery after the improvement in manufacturing production and business environment throughout Japan (BOJ Report, 2014). However, there are several prefectures that show negative trend, such as Fukui and Nagasaki prefectures. This trend in Nagasaki prefecture mainly affected by decline in the ship building industry, fisheries industry, and small enterprise (Miyao, 2014).

Even though we did not observe the structural change in Japan case based on F-test, but the widespread for the impact of Lehman's shock supported by the spatial correlation coefficients in **Fig 8**. The results indicate the growth rate in the neighboring prefectures induced each prefecture's economic growth in the same direction.

Wald test cannot be conducted in Japan data due to the small sample size (Greene, 2012). In our case, we only have sample size $T = 13$ with $n_1 = 7$ and $n_2 = 6$, which is too small for the requirement.

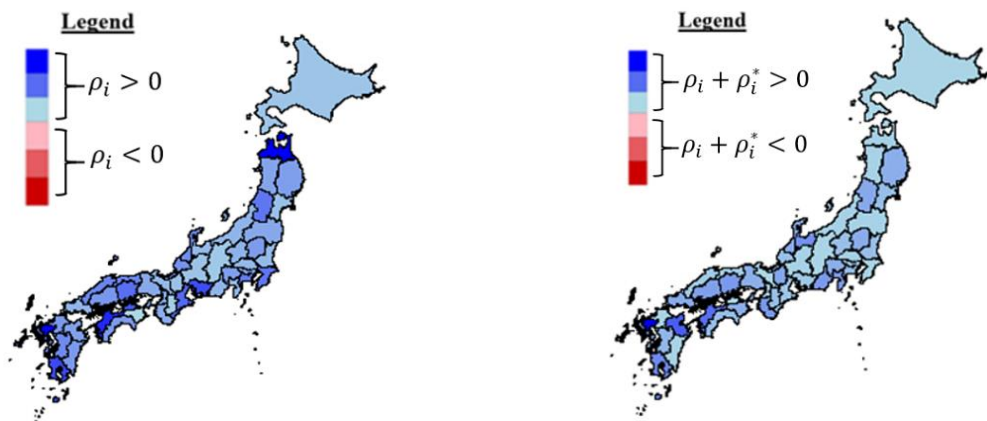


Fig 8. The spatial correlation before (left) and after (right) breakpoint in Japan.

4. Discussion and Conclusion

Several statistical methods for structural change, such as F-test and Wald-test based on broken time-trend model are used to analyze the effect of Lehman's crisis towards regional economic growth in US and Japan.

We observed that the F-test perform better compared to the Wald test for spatially-independent model. The Wald test produced better results for spatially-dependent model. We confirmed, the endogeneity in spatially-dependent model might imply the F-test become non-robust for autocorrelated disturbances (Krämer, 2003).

In the US, it was revealed that the negative impact of crisis clustered in West Coast, Southeastern, and Great Lake region. States in those regions with good manufacturing, construction, insurance, and finance as the main contributor for their GRP. On the other hand, states that relied on agriculture, forestry, fishing, hunting, and mining sectors had resilience.

In Japan, we observed a negative impact spread-out across all the regions. After the crisis, most prefectures showed recoveries, except several other prefectures, such as Nagasaki. The negative trend in Nagasaki mainly affected by decline in the ship building industry, fisheries industry, and small enterprise (Miyao, 2014).

References

1. Ahn, J.B., Amiti, M., & Weinstein, D.E. (2011). "Trade finance and the great trade collapse." *American Economic Review*, 101 (3), 298-302.
2. Aloui, R., Aissa, M.S.B., & Nguyen, D.K. (2011). "Global financial crisis, extreme interdependences, and contagion effects: the role of economic structure?" *Journal of Banking & Finance*, Vol. 35 (1), 130-141.
3. Bank for International Settlements (BIS) Report. (2009). *79th Annual Report: 1 April 2008-31 March 2009*. BIS Publications.

4. Bank of Japan (BOJ). (2008). *Regional Economic Report (Summary) (October 2008)*, BOJ Reports & Research Papers.
5. Bank of Japan (BOJ). (2008). *Financial System Report (March 2008)*, BOJ Reports & Research Papers.
6. Bank of Japan (BOJ). (2014). *Regional Economic Report (Summary) (January 2014)*, BOJ Reports & Research Papers.
7. BEA (Bureau of Economic Analysis) Report. (2009). "Economic slowdown widespread among states in 2008: Advance 2008 and revised 2005–2007 GRP-by-State Statistics." *BEA Press Release*
8. Cetorelli, N. & Goldberg L. S. (2011). "Global banks and international shock transmission: evidence from the crisis." *IMF Economic Review*, Vol. 59 (1), 41–76.
9. Demirer, M, et al. (2017), "Estimating global bank network connectedness," *NBER Working Paper No. 231340*
10. Filardo, A., et al. (2009). "The international financial crisis: timeline, impact and policy response in asia and the pacific," *BIS Papers No. 52*, 21-82.
11. Greene, W. H. (2012). *Econometric Analysis 7th edition*. Pearson: New York.
12. Jouini, J. & Boutahar, M. (2004). "Evidence on structural changes in US time series," *Economic Modelling*, Vol 22, 391-422.
13. Kawai, M., & Takagi, S. (2009). "Why was japan hit so hard by the global financial crisis?" *Asian Development Bank Institute (ADB) Working Paper No. 153*. Tokyo: Asian Development Bank Institute.
14. Krämer, W. (2003). "The robustness of the F-test to spatial autocorrelation among regression disturbances," *Statistica*, Vol. 63 (3), 435-440.
15. Longstaff, F.A. (2010). "The subprime credit crisis and contagion in financial markets," *Journal of Financial Economics*, Vol. 97 (3), 436-450.
16. Ministry of Finance (MOF) Japan. (2012). *Trade Statistics: Value of Export and Imports 2011 (Calendar Year) (January-December) (Fixed Annual)*. 2012 Press Release of Ministry of Finance, Japan.
17. Miyao, R. (2014). *Economic Activity and Prices in Japan and Monetary Policy*, Speech at a Meeting with Business Leaders in Nagasaki, Bank of Japan Speeches and Statement.
18. Reinhart, C. M, & Rogoff, K. S. (2008). "Is the 2007 US sub-prime financial crisis so different? An international historical comparison." *American Economic Review* Vol. 98 (2), 339-344.



On Gaussian semiparametric estimation for two-dimensional intrinsic stationary random fields



Yoshihiro Yajima

Tohoku University, Sendai, Japan

Abstract

We propose a Gaussian semiparametric estimator for semiparametric models of two-dimensional intrinsic stationary random fields (ISRFs) observed on a regular grid and derive its asymptotic properties. Originally this estimator is an approximate likelihood estimator in a frequency domain for long memory models of stationary and nonstationary time series (Robinson (1995); Velasco (1999)). We apply it to two dimensional ISRFs. These ISRFs include a fractional Brownian field, which is a Gaussian random field and is used to model many physical processes in space. The estimator is consistent and has the limiting normal distribution as the sample size goes to infinity. We conduct a computational simulation to compare the performance of it with those of different estimators.

Keywords

spatio-temporal models; local Whittle estimator; fractional Brownian field

1. Introduction

Let $\{X(s) : s \in \mathbf{R}^d\}$ be a random field. Throughout this paper, we specialize to the two-dimensional random field, $d = 2$. Whereas for the moment, for ease of description, we consider a general d -dimensional setting. If $\{X(s)\}$ satisfies that for any fixed $\mathbf{h}(\in \mathbf{R}^d)$, the increment $Z_{\mathbf{h}}(\mathbf{s}) = X(\mathbf{s} + \mathbf{h}) - X(\mathbf{s})$ is a stationary random field, $\{X(s)\}$ is called an ISRF. Then $\{X(s)\}$ is characterized by

$$\begin{aligned} E(X(\mathbf{s} + \mathbf{h}) - X(\mathbf{s})) &= 0, \\ \text{Var}(X(\mathbf{s} + \mathbf{h}) - X(\mathbf{s})) &= 2_{\gamma}(\mathbf{h}); \end{aligned}$$

where $2_{\gamma}(\mathbf{h})$ is the variogram function (see Chilés & Delfiner (2012); Cressie (1993)). Hereafter we also assume that $X(\mathbf{0}) = 0$.

For $\boldsymbol{\lambda}$ and \mathbf{h} , let $(\boldsymbol{\lambda}, \mathbf{h})$ be the inner product and $\|\boldsymbol{\lambda}\|$ be the norm.

Then if $2_{\gamma}(\mathbf{h})$ is a continuous function on \mathbf{R}^d satisfying $\gamma(\mathbf{0}) = 0$, it has the spectral representation

$$2_{\gamma}(\mathbf{h}) = \int_{\mathbf{R}^d} \frac{1 - \cos((\boldsymbol{\lambda}, \mathbf{h}))}{(2\pi)^d} G(d\boldsymbol{\lambda}) + Q(\mathbf{h}), \quad (1)$$

where $Q(\mathbf{h})(\geq 0)$ is a quadratic form and $G(\boldsymbol{\lambda})$ is a positive, symmetric measure such that $\|\boldsymbol{\lambda}\|^2 G(\boldsymbol{\lambda})$ is continuous at the origin and

$$\int_{\mathbb{R}^d} \frac{\|\boldsymbol{\lambda}\|^2}{1 + \|\boldsymbol{\lambda}\|^2} G(d\boldsymbol{\lambda}) < \infty \tag{2}$$

(See (see Chilés & Delfiner (2012); Cressie (1993); Solo (1992); Yaglom (1957).)

Hereafter we assume that $Q(\mathbf{h}) = 0$ and $G(\boldsymbol{\lambda})$ is absolutely continuous with density $g(\boldsymbol{\lambda})$. Then (1) and (2) reduce to

$$2_\gamma(\mathbf{h}) = \int_{\mathbb{R}^d} \frac{1 - \cos(\langle \boldsymbol{\lambda}, \mathbf{h} \rangle)}{(2\pi)^d} g(\boldsymbol{\lambda}) d\boldsymbol{\lambda},$$

and

$$\int_{\mathbb{R}^d} \frac{\|\boldsymbol{\lambda}\|^2}{1 + \|\boldsymbol{\lambda}\|^2} g(\boldsymbol{\lambda}) d\boldsymbol{\lambda} < \infty,$$

respectively.

An interesting special class of ISRF's that is often applied to empirical data analysis in space is a fractional Brownian field (FBF)(see Adler (1981); Mandelbrot& Van Ness (1968); Zhu & Stein (2002) and the references therein). A FBF is a Gaussian ISRF and has $2_\gamma(\mathbf{h}) = C \|\mathbf{h}\|^{2H}$, which is equivalent to

$$g(\boldsymbol{\lambda}) = \frac{CHK_H}{\|\boldsymbol{\lambda}\|^{d+2H}},$$

where

$$K_H = \pi^{d/2} 2^{2H+d} \Gamma(d + 2H/2) / \Gamma(1 - H), \quad 0 < H < 1.$$

(Yaglom (1957)). C is a scale parameter and H is a smoothness parameter with larger values corresponding to smoother surfaces.

Hereafter we assume that $d = 2$ and $\{X(s)\}$ is a two-dimensional Gaussian ISRF. Then we also denote $\boldsymbol{\lambda}$ by (λ_1, λ_2) and the spectral density function $g(\boldsymbol{\lambda})$ by (λ_1, λ_2) respectively.

In this paper we consider the following class of the spectral density function, which includes a FBF.

Assumption 1 $g(\lambda_1, \lambda_2)$ is expressed by

$$g(\lambda_1, \lambda_2) = \|\boldsymbol{\lambda}\|^{-2h-2} g_o(\lambda_1, \lambda_2), \quad 0 < H < 1,$$

where $g_o(\lambda_1, \lambda_2)$, is a nonnegative with $g_o(0,0) > 0$, symmetric, $g_o(\lambda_1, \lambda_2) = g_o(-\lambda_1, -\lambda_2)$, twice continuously differentiable function for $-\infty < \lambda_1, \lambda_2 < \infty$ and is bounded with bounded first and second order partial derivatives.

2. Methodology

We denote the sampling sites and observations by $s_{qr} = (q, r)$ and $X(s_{qr})(q, r = 1, \dots, n)$ respectively. Then the sample size is n^2 .

Definition 2.1(Data Tapers of order p). $\{h_t: t = 1, \dots, n\}$ is called a sequence of data tapers of order p if it satisfies the following conditions.

- (1) h_t is positive and symmetric around $t = n/2$ with $\max_{1 \leq t \leq n} h_t = 1$.
- (2) For any $n > 0$, there exists a constant $b, 0 < b < \infty$, which may depend on n so that $\sum_{t=1}^n h_t^n = bn$ holds.

(3) For $N = n/p$ (which we assume as an integer), the kernel $D_p = \sum_{t=1}^n h_t \exp(i\lambda t)$ satisfies

$$D_p(\lambda) = \frac{a_n(\lambda)}{n^{p-1}} \left[\frac{\sin(N\lambda/2)}{\sin(\lambda/2)} \right]^p,$$

where $a_n(\lambda)$ is a complex function, whose modulus is bounded and bounded away from zero, with $p - 1$ derivatives, all bounded in modulus as n increases for $\lambda \in [-\pi, \pi]$.

Then we define the discrete Fourier transform (DFT) $\omega_p(\omega_{j_1}, \omega_{j_2})$ and the periodogram $I_p(\omega_{j_1}, \omega_{j_2})$ by

$$\begin{aligned} \omega_p(\omega_{j_1}, \omega_{j_2}) &= \frac{1}{2\pi \sum_{t=1}^n h_t^2} \sum_{q,r=1}^n h_q h_r X(s_{qr}) \exp(i(\omega_{j_1, j_2}, s_{qr})), \\ I_p(\omega_{j_1}, \omega_{j_2}) &= |\omega_p(\omega_{j_1}, \omega_{j_2})|^2, \end{aligned}$$

respectively where ω_{j_1, j_2} is the bivariate Fourier frequency,

$$\omega_j = (\omega_{j_1}, \omega_{j_2}) = \left(\frac{2\pi j_1}{n}, \frac{2\pi j_2}{n} \right), \left[\frac{(n-1)}{2} \right], \leq j_1, j_2 \leq \left[\frac{n}{2} \right],$$

and $[x]$ is the integer part of x and $\{h_t\}$ is a sequence of data tapers of order p .

Next we define the normalized DFT by

$$\begin{aligned} v_p(\omega_{j_1, j_2}) &= \frac{\omega_p(\omega_{j_1}, \omega_{j_2})}{(G(\omega_{j_1}^2 + \omega_{j_2}^2))^{-(H+1)/2}}, \\ v_{pR}(\omega_{j_1, j_2}) &= \text{Re}(v_p((\omega_{j_1, j_2}))), \\ v_{pI}(\omega_{j_1, j_2}) &= \text{Im}(v_p((\omega_{j_1, j_2}))), \end{aligned}$$

where $G = g_0, (0,0)/(8\pi^2)$. H

Now we introduce the estimator. We denote by G_0 and H_0 the true parameters, and by G and H any admissible values. Then define the closed interval of admissible parameters of H_0 , $\mathcal{H} = [\Delta_1, \Delta_2]$, where Δ_1 and Δ_2 are numbers chosen such that $0 < \Delta_1 < \Delta_2 < 1$. We can choose Δ_1 and Δ_2 arbitrarily close to 0 and 1 respectively. Next consider the objective function

$$\begin{aligned} Q(G, H) &= \frac{1}{m} \sum_{(j_1, j_2) \in S_n} \{ \log(G(\omega_{j_1 p \xi}^2 + \omega_{j_2 p \xi}^2)^{-H-1}) \\ &\quad + \frac{(\omega_{j_1 p \xi}^2 + \omega_{j_2 p \xi}^2)^{H+1}}{G} I_p(\omega_{j_1 p \xi}, j_2 p \xi) \}, \end{aligned}$$

where

$$\begin{aligned} S_n &= \{ (j_1, j_2) \mid r_{L,n}^2 \leq \left(\frac{j_1 p \xi}{n} \right)^2 + \left(\frac{j_2 p \xi}{n} \right)^2 \leq r_{U,n}^2, 0 < j_1, j_2, b_L \leq \frac{j_2}{j_1} \leq b_u \}, \\ b_L &< 1 < b_u, \end{aligned}$$

and m is the cardinality of S_n . ξ plays an important role so that $I_p(\omega_{j_1 p \xi}, j_2 p \xi)$ and $I_p(\omega_{k_1 p \xi}, k_2 p \xi)$ are asymptotically independent if $(j_1, j_2) \neq (k_1, k_2)$. Then the estimator exists.

$$\hat{G}_n, \hat{H}_n = \arg \min_{0 < G < \infty, H \in \mathcal{H}} Q(G, H),$$

3. Result

Assumption 2 Define $\tilde{r}_{L,n} = \frac{nr_{L,n}}{p\xi}$, $\tilde{r}_{U,n} = \frac{nr_{U,n}}{p\xi}$.

- (1) $r_{U,n} \rightarrow 0$ and $\tilde{r}_{L,n} \rightarrow \infty$ as $n \rightarrow \infty$.
- (2) $\log n = O(\log(nr_{L,n}))$ and $r_{L,n}(\log\tilde{r}_{U,n}r_{L,n} + |\log(r_{L,n}/p\xi)| + |\log(r_{U,n}/p\xi)|)/r_{U,n} \rightarrow 0$ as $n \rightarrow \infty$.
- (3) For $p \geq 2, \xi \rightarrow \infty$.

Assumption 3 $nr_{L,n}^{-2}, r_{U,n}^2$ and ξ^{-p} are smaller order than $\tilde{r}_{U,n}^{-1}$ as $n \rightarrow \infty$.

Then we have the following asymptotic properties of the estimator.

Theorem 1 Under Assumption 1 and 2, \hat{H}_n converges to H_0 in probability as $n \rightarrow 1$.

Theorem 2 Under Assumptions 1-3, for $p \geq 2, m^{\frac{1}{2}}(\hat{H}_n - H_0)$ converges to $N(0, 1)$ in distribution as $n \rightarrow \infty$.

4. Discussion and Conclusion

This section examines empirical properties of the Gaussian semiparametric estimation (GSE) estimators for the parameter H in comparisons with those of Zhu& Stein (2002), which is a spatial domain method that assumes that g_0 in Assumption 1 is a constant. Our interests are in the comparisons between them when g_0 is not a constant.

To conduct the comparisons, we simulate spatial data on lattice points that satisfies Assumption 1 in the following three cases. In Case 1, we considered FBFs with $H = 0.5$, denoted as $X_{0.5}(s, t)$. Case 1 clearly satisfies Assumption 1 with g_0 being a constant. In Case 2, for iid noise $\varepsilon(s, t)$, we generate FBF contaminated with the noise, which is given by $X_{0.5}(s, t) + \varepsilon(s, t)$. In Case 3, we simulate the moving average of $X_{0.5}(s, t)$, given by

$$(1 + 0.3B_1)(1 + 0.3B_2)X_{0.5}(s, t),$$

for the backward shift operators defined by $B_1X(s, t) = X(s - 1, t)$ and $B_2X(s, t) = X(s, t - 1)$.

We simulated 100 sets of Cases 1, 2 and 3, for which we constructed the two kinds of estimators to examine the empirical comparisons between them. The sample paths of cross section over $s = 1$ for the three cases are shown in Figure 1. It should be noticed that the each of three sample paths is a realization of ISRF with $H = 0.5$,

We calculated the two kinds of estimators by GSE and Zhu& Stein (2002) for the three cases, where we chose the filter 1 with $M = 2$ for $k = 1, 2, 3, 4, 5$ to conduct OLS for the latter estimator.

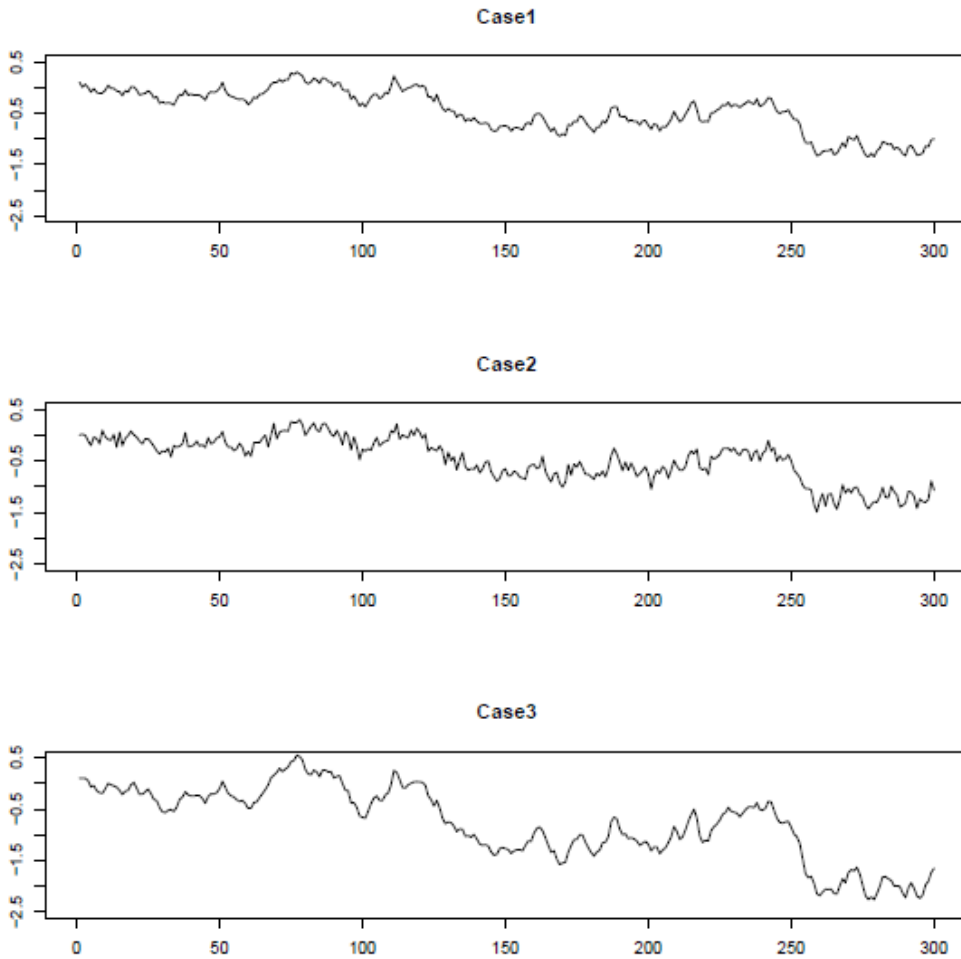


Figure 1: Sample path of the cross-section for the three cases of random fields over 300×300 grid points. For fractional Gaussian fields $X(s_1, s_2)$ with $H = 0.5$, Cases 1, 2 and 3 are, respectively, $X(s_1, s_2)$, $X(s_1, s_2)$ plus iid noise $\varepsilon(s_1, s_2)$ and the moving average $(1 + 0.3B_1)(1 + 0.3B_2)X(s_1, s_2)$ for the backward shift operators B_1 (horizontal) and B_2 (vertical).

To construct GSE, we designed the data taper as the simplest one of $p = 1$. In Figure 3, we show histograms of the estimators of GSE and Zhu & Stein (2002).

We find from Figure 3 that the estimator of Zhu & Stein (2002) are negatively and positively biased for Cases 2 and 3, respectively, while GSE cured the biases although the variances are larger.

In Case 1, Zhu and Stein has no bias with smaller variance. The comparisons for Cases 1-3 demonstrate that GSE can cure the bias occurring for Zhu and Stein (2002) when g_0 is not constant. GSE, which estimates H depending on low frequency components, can avoid bias that comes from the non-constant g_0 , while Zhu and Stein, which estimates on high frequency components, is

influenced directly by non-constant g_0 , that results in biased estimation. GSE works better to estimate H when g_0 is not constant than Zhu & Stein (2002).

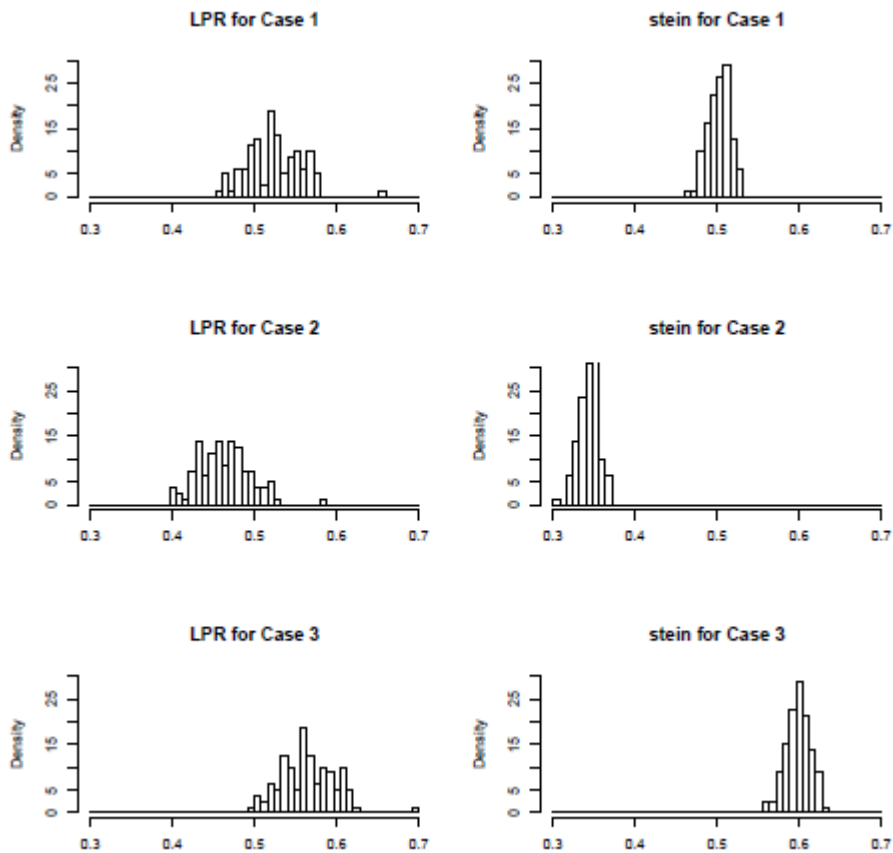


Figure 2: Histograms of the estimators by GSE and by Zhu and Stein (2002) evaluated by 100 simulations for Cases 1, 2 and 3.

References

1. Adler, R.J. (1981). *The Geometry of Random Fields*. Wiley, New York.
2. Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. 2nd edition. Springer, New York.
3. Chilés, J.-P. and Del_ner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. 2nd ed. Wiley, New York.
4. Constantine, A. G. and Hall, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *J. Roy. Statist. Soc. Ser. B* 56, 97-113.
5. Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Revised ed. Wiley, New York.
6. Davis, S. and Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *J. Roy. Statist. Soc. Ser. B* 61, 3-37.

7. Matheron, G. (1973). The intrinsic random functions and their applications. *Adv. Appl. Probab.* 5, 439-468.
8. Mandelbrot, B.B. and Van Ness, J.W. (1968). Fractional Brownian motion, fractal noises and applications. *SIAM. Rev.* 10, 422-437.
9. Robinson, P.M. (1995). Gaussian semiparametric estimation of long range dependence. *Ann. Statist.* 23, 1630-1661.
10. Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, New York.
11. Stein, M. L. (2002). Fast and exact simulation of fractional Brownian surfaces. *J. Comp. Graphical Stat.* 11, 587-599.
12. Solo, V. (1992). Intrinsic stationary random functions and the paradox of $1=f$ noise. *SIAM J. Appl. Math.* 52, 270-291.
13. Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *J. Time Ser. Anal.* 20,1 87-127.
14. Yaglom, A.M. (1957). Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory Probab. Appl.* 2 273-320.
15. Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* 92, 921-936.
16. Zhu, Z. and Stein, M.L. (2002). Parametric estimation for fractional Brownian surfaces. *Statist. Sinica* 12, 863-883.



Prioritisation of Sustainable Development Goals and efforts of SESRIC to support statistical modernisation in OIC member countries



Nebil Dabur, Atilla Karaman
SESRIC

Abstract

The United Nations General Assembly endorsed the Sustainable Development Goals (SDGs) in September 2015 for which all countries have pledged to achieve the 17 goals and 169 targets by 2030. Unlike the Millennium Development Goals period, the challenges are valid for not only developing but also developed countries in implementing the SDGs which in effect shakes the definition of “developing” and “developed” country. The implementation of SDGs will understandably be more complex for economically and technologically disadvantaged countries of the Organisation of Islamic Cooperation (OIC) that had previously faced challenges in achieving the MDGs. In this respect, proper prioritisation, planning and careful consideration of the multidimensional interactions among the SDG targets will be critical in the achievement of SDGs. Tasked with following up the SDG Indicator Framework and identifying the SDGs priorities of the OIC countries, the Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC) designed and circulated the “Tendency Survey on SDG Priorities of OIC Member Countries”. This paper discusses the results obtained from the aforementioned Survey including the SDG prioritisation, expected SDG achievement levels, limiting factors on SDGs, national commitments to SDGs, structure of agencies responsible for SDGs, SDG data availability, cooperation with international agencies on SDGs, and statistical needs and capacities of OIC member countries on SDGs. This paper also covers the efforts of SESRIC in facilitating (i) the flow of know-how and experience sharing on issues related to official statistics and SDGs through its flagship OIC Statistical Capacity Building Programme (StatCaB) since 2007; and (ii) engaging with the global statistical community to support the endeavours of OIC member countries in modernising their national statistical systems in meeting the requirements of the global SDG monitoring and reporting framework.

Keywords

SESRIC; statistical modernisation; OIC member countries

1. Introduction

The twenty-first century is full of contradictions in many aspects. On the one hand, the technologies we developed have facilitated how we live and do

business; on the other hand, these technologies have created new problems we have to solve. While the benefits of these technologies are many, humanity had to sacrifice a lot, including the environment we live in, which may soon be detrimental to our very survival on this planet.

The United Nations Conference on the Human Environment in Stockholm in 1972 and the Earth Summit in Rio in 1992 were the first examples to focus on the actions to be taken to alleviate environmental problems and contribute to the development of the Global South. In year 2000, the endorsement of the United Nations Millennium Declaration put on the shoulders of decision-makers in developing nations the task of achieving the eight international development goals (MDGs) that aimed to improve the well-being and welfare of their countries.

In 2015, marking another special period in international development history, the United Nations embarked on the Sustainable Development Goals (SDGs) for which all countries have pledged to achieve the 17 goals and 169 targets by 2030. Unlike the MDGs, there are challenges for both the developing and developed countries in implementing the SDGs which in effect shakes the definition of "developing" and "developed" country.

The implementation of SDGs will understandably be more complex for economically and technologically disadvantaged countries of the Organisation of Islamic Cooperation (OIC) that faced challenges in achieving the eight MDGs. In this respect, proper planning and careful consideration of the multidimensional interactions among the SDG targets will be critical in the accomplishment of SDGs. This exercise should undoubtedly involve national, regional, and international stakeholders, and requires pertinent prioritization of the SDGs and targets.

Unlike the MDGs period, the member countries and the relevant OIC fora acted timely to include the SDGs into their agenda. Being an important forum of the OIC, the Standing Committee for Economic and Commercial Cooperation of the OIC (COMCEC) has been discussing the SDGs since 2014. In line with the relevant resolutions of the COMCEC Sessions that have been held in 2015 and 2016, SESRIC has been tasked with identifying the SDGs priorities of the OIC countries which will contribute to the operational planning of the activities to be conducted concerning the SDGs until 2030.

This paper discusses the results obtained from the aforementioned Survey including the SDG prioritisation, expected SDG achievement levels, limiting factors on SDGs, national commitments to SDGs, structure of agencies responsible for SDGs, SDG data availability, cooperation with international agencies on SDGs, and statistical needs and capacities of OIC countries on SDGs. This paper also covers the efforts of SESRIC in facilitating (i) the flow of know-how and experience sharing on issues related to official statistics and SDGs through its flagship OIC Statistical Capacity Building Programme

(StatCaB) since 2007; and (ii) engaging with the global statistical community to support the endeavours of OIC member countries in modernising their national statistical systems in meeting the requirements of the global SDG monitoring and reporting framework.

2. Methodology

In order to identify the SDG priorities of the OIC countries in accordance with the Resolution #117 [1] of the 31st Session of COMCEC, SESRIC designed and circulated a questionnaire after the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs) had finalised the initial SDG tier system on 29 July 2016 [2]. Made available in English, Arabic, and French, the questionnaire [3] comprised of the following four sections:

- i. INTRODUCTION: Briefing on the purpose of the questionnaire, fields for contact details of the head of institution responding to the questionnaire, SDG focal point in the responding institution and respondent completing the questionnaire;
- ii. PART A: Prioritisation of SDGs, Expected Achievement Levels, and Limiting Factors on SDGs;
- iii. PART B: National Commitment to SDGs; Relevant Agencies, Their Human Resource Capacities; Cooperation with International Agencies; and Training Needs and Capacities on SDGs; and
- iv. ANNEX: List of SDGs, targets, and indicators under each SDG.

Our sampling frame consisted of all National Statistical Offices (NSOs) of OIC countries as the NSOs are one of the main stakeholders of SESRIC that are instrumental in reaching out to the other relevant government entities that may be responsible for SDG planning, coordination, implementation, monitoring and reporting inside the OIC countries. Additionally, SESRIC sent the questionnaire to the embassies of OIC countries in Ankara, Turkey and in other countries with accreditation status with the OIC General Secretariat.

Between August 2016 and September 2018, responses received in XLS format from the respondent OIC countries. Full responses were received from 17 OIC countries while partial responses were received from 19 OIC countries. To prioritise the 17 SDGs and 169 targets, the countries were given the option to assign either High (numerical score of 4), Medium (numerical score of 3), Low (numerical score of 2), None (numerical score of 1), or Irrelevant (numerical score of 0). The range of priority assignments of OIC countries was between 23 (SDG 14) and 28 (SDG 1). We considered the number of countries that assigned "High" priority and the qualified majority principle to determine the prioritisation of the SDGs and targets at the OIC level. Where there was a tie in the number of "High" assigning countries above the qualifies majority threshold in the related SDG and/or targets, we used simple averages of the

numerical scores assigned by respondents to break the tie and identify the priority of the related SDG and/or target accordingly.

The methodology for the identification of the SDG data availability previously carried out by the UNESCAP [5] has been adopted and data availability of the SDG indicators have been examined by considering two types indicator analyses:

- i. Status of a situation at one point in time; and
- ii. Describing the change in the status of situation as measured by an indicator which requires a minimum of two data points.

In this regard, the SDG data availability analysis in this paper was conducted based on the following four criteria also adopted in the UNESCAP methodology:

- i. Trend analysis possible (Trend OK): If a particular indicator has two or more data points available for 50 per cent (or more) of the OIC countries between 2000-2018;
- ii. Only status analysis possible (Status OK): If a particular indicator has only one data point available for 50 per cent (or more) of the OIC countries between 2000 and 2018;
- iii. Limited status analysis possible (Status LIMITED): If a particular indicator has at least one data point available but for less than 50 per cent of the OIC countries between 2000 and 2018; and
- iv. No analysis possible (No Data): If no data points are available for any of the OIC countries between 2000 and 2018.

More details about the methodology can be also seen at [4].

3. Result

Based on our analysis of SDG prioritisation, the respondent OIC countries prioritised the following eight SDGs as "High":

1. SDG 1: No Poverty;
2. SDG 3: Good Health and Well-being;
3. SDG 2: Zero Hunger;
4. SDG 4: Quality Education;
5. SDG 5: Gender Equality;
6. SDG 8: Decent Work and Economic Growth;
7. SDG 9: Industry, Innovation & Infrastructure; and
8. (SDG 13): Climate Action.

As to the expected SDG achievement levels, only 1 respondent country (Iraq) stated that it would achieve SDGs 9 and 12-17 by 2020. On the other hand, 6 respondent countries envisage that they would achieve SDG 4 (Chad, Guinea, Iraq, Jordan, Sudan, and Yemen); 5 of them have the same hope for the achievement of SDG 2 (Bangladesh, Chad, Jordan, Palestine, and Sudan) by 2030. In addition, another group of 4 countries (Guinea, Indonesia, Jordan,

and Sudan) and (Chad, Jordan, Palestine, and Sudan) also hope to achieve SDG 16 and SDG 6 respectively by 2030. 3 respondent countries (Iraq, Jordan, and Sudan) mentioned that they would achieve SDG 3 by 2030. From all the respondent countries while Sudan stated that they expect to achieve 13 SDGs (1-6, 8, and 11-16) by 2030, Jordan emerges as the only respondent country having stated its hope for the achievement of 17 SDGs by 2030.

Concerning the limiting factors in the achievement of SDGs, an average of 12 respondents stated that “shortage of financial resources” is the most salient limiting factor which is followed by lack of data sources to monitor and evaluate and Lack of technological/IT means (8 respondents on average); inadequate human resources capacity and Lack of methodological knowledge (7 respondents on average); lack of political support and lack of coordination among relevant agencies/stakeholders (5 respondents on average); and lack of laws, regulations, policies (4 respondents on average).

The current commitment of SDG implementation has only been provided by 28 respondent countries. Of those, 14 respondents stated that they are currently committed to the implementation of all 17 SDGs. Due to their landlocked status, 4 respondents stated that they are committed to implement 16 SDGs, excluding SDG 14 “Life under water”. Remaining 10 respondents provided a current commitment of SDG implementation ranging between 1 and 15 SDGs. Based on the responses received, SDG 4 is the top goal that has a current commitment for SDG implementation by 27 countries followed by SDGs 2, 3, 5 and 6 by 26 countries; SDGs 1, 8 and 13 by 25 countries; SDG 9 by 24 countries; SDGs 7 and 10 by 23 countries; SDGs 16 and 17 by 22 countries; SDGs 11 and 15 by 21 countries; SDG 12 by 18 countries; and SDG 14 by 17 countries.

29 out of 36 of the respondent countries stated they have SDG coordinating agencies. The analysis shows that 17 OIC countries assigned their Ministries of Development / Economy / Environment / Foreign Affairs / Planning as their SDG coordinating body while in 2 countries stated that Prime Ministry or the Council of Ministers is directly responsible for the SDG coordination. 4 respondents stated that they have a separate SDG coordination under a General Secretariat mechanism. 6 countries stated their NSOs are responsible for the SDG coordination. SDG monitoring agencies have been observed to be existing in 27 respondent countries. Once more, 11 OIC countries assigned their Ministries of Development / Economy / Environment / Foreign Affairs / Planning as their SDG monitoring bodies corresponding to their coordination role. In 4 countries, Prime Ministry or the Council of Ministers is directly responsible for the SDG monitoring. While 3 countries stated they use separate SDG monitoring under a General Secretariat mechanism (again similar to their coordination role), in 9 OIC countries, NSOs are also directly responsible for monitoring the SDGs. It is also

found out that 29 respondents could name their SDG reporting agencies. 12 respondents indicated their Ministries of Development / Economy / Environment / Foreign Affairs / Planning as their SDG reporting bodies. NSOs are the second mainly assigned SDG reporting agency in 10 respondent countries followed by 4 countries where National SDG Committee undertakes the SDG reporting role. 3 respondent countries also stated that the Prime Ministry and/or Council of Ministers is shouldering the SDG reporting role.

Inhomogeneous responses provided by the countries made it difficult to analyse the overall SDG data availability situation at the OIC level. To remedy this shortcoming, the data currently hosted by the UN Statistics Division (UNSD) in its Global SDG Indicators Database [6] have been considered to depict the data availability situation of the OIC countries concerning the eight prioritised SDGs. Across 116 global SDG indicators allocated for the eight prioritised SDGs and based on the data made available in [6] in March 2019, trend analysis is observed to be possible for 47 indicators (40.5% of the indicators). In this connection, SDG 3 and 9 are the only two SDGs on which a comprehensive coverage is possible for at least 50% of the indicators. On the other hand, SDG 1, 2, and 8 have a ranging level of 35% to 45% of indicators with which a “trend analysis” is possible. It is also observed that due to lack of data, it is not possible to do an analysis on 30 indicators (25.9% of the indicators). The situation is especially not promising for SDG 5 and 13 where there is only 1 indicator fit for a trend analysis.

In connection to the current commitments of the OIC countries for the implementation of the SDGs, 15 respondents acknowledged that they have cooperation with and/or receive consultancy from regional/international organizations from UN agencies, UN Regional Economic Councils, World Bank, and other relevant international and regional agencies concerning all 17 SDGs.

Closely parallel to the results of current commitment for SDG implementation and cooperation linkages with the international agencies, 16 respondents stated that they are in need of capacity building activities for all 17 SDGs. At the goal level, SDGs 1, 7, 8, and 11 are with the highest number of respondents (27) asking for capacity building and were followed by SDGs 4, 10, and 12 (26 countries); SDGs 2, 3, 5, 6, 13, 14, 16, and 17 (25 countries); SDG 15 (23 countries); and SDG 9 (22 countries). Except Jordan and Pakistan that stated it can provide capacity building to other OIC countries on all SDGs, remaining 16 respondents stated they could provide capacity building on SDGs to other OIC countries on different SDGs.

4. Discussion and Conclusion

The response rate for the questionnaire designed and circulated by SESRIC shows that only a few dozen of OIC countries has the necessary infrastructure to plan and coordinate the SDG implementation in their respective countries.

In this regard, the exercise to identify the SDG prioritisation of OIC countries may be repeated at a later stage after 2020 as the procedures to realign the national development mechanisms and entities with the SDGs take longer than expected.

It is also obvious from the analysis that the prioritised eight SDGs focus mainly on the fundamental areas of development; namely, poverty, health, hunger, education, labour and economic growth, and infrastructure. Gender equality and climate action, on the other hand, are deemed also as important as the fundamental areas.

Expected SDG achievement levels by 2030 unfortunately has not indicated a full success by the majority of the respondent OIC countries. This fact is highly connected with the most salient limiting factor in front of achieving the SDGs – shortage of financial resources – to implement the related SDG projects.

Against these, it has not been a surprise to see low levels of commitments to the implementation of all 17 SDGs. Rather, respondents were selective in current SDG commitments focusing on education, hunger, health, gender, water, poverty, labour, and climate change.

While SDG coordinating, monitoring, and reporting agencies are existent in most respondent countries, SDG data availability is also a major obstacle for the coordinating agencies. For the eight prioritised SDGs, out of 116 global indicators, only 47 of them are fit for a trend analysis. The situation is serious in indicators under SDG 5 and 13 where there is only 1 indicator for each for trend analysis.

To build the needed capacities to remedy the data gaps and produce the pertinent statistics instrumental in planning and implementing the SDG related activities and projects, cooperation with regional and international organisations plays a critical role. In this context, SESRIC carries out a flagship statistical skills development initiative titled “OIC Statistical Capacity Building Programme (StatCaB)” to strengthen the National Statistical Systems (NSSs) in the OIC countries with a view to producing better national statistics and thus helping policy-makers introducing better national policies and strategies. In order to identify the statistical needs and capacities of the official statistics producing institutions of the OIC countries, SESRIC circulates biennial questionnaires and matches these needs and capacities through organising statistical activities. In line with resolution #121 of the 34th Session of the COMCEC, SESRIC has recently included the SDG indicators into its 2020-2021 StatCaB Biennial Questionnaire; and for the year 2019, the matching was done by considering the prioritised SDGs indicated by the OIC countries based on the “Tendency Survey on SDG Priorities of OIC Member Countries” conducted by SESRIC. Since its inception in 2007, SESRIC has organised over 370 statistical activities with the participation of thousands of experts and high-level officials

from the NSOs and other constituents of the NSSs of OIC countries. Four of those activities focused on 'Sustainable Development Statistics' which covered work on indicators and frameworks to monitor the economic, social and environmental dimensions of sustainable development. At the broader level, SESRIC continues its cooperation with other regional and international organisations concerning SDGs. As the Secretariat of the OIC Statistical Commission, SESRIC has invited relevant international organisations to its sessions since 2013 to inform the OIC countries on the developments concerning the global SDG measurement framework. Additionally, SESRIC has enacted Memoranda of Understanding or exchanged similar documents to collaborate with these agencies on projects of common interest to better cater the needs of OIC countries in SDG measurement, monitoring, and reporting.

To conclude, our findings in this paper verify the findings of the SESRIC Report in 2016 [7]. The Report states that the realization of sustainable development agenda in OIC countries and elsewhere in the developing world depends largely on their ability to address issues and challenges related with (i) political will and policy dialogue, (ii) institutional capacity and governance, (iii) data collection and monitoring, and (iv) peace and security, and financial resources. Therefore, a strong ownership and a conducive environment strengthened through stable economic and peaceful conditions will pave the way towards increased institutional capacities for SDG planning, coordination, implementation, monitoring, and reporting.

References

1. COMCEC (2015), "Resolutions of the 31st Ministerial Session of the COMCEC", online available at <http://www.comcec.org/en/wp-content/uploads/2016/05/31IS-RES.pdf>
2. UNSD (2019), "Tier Classification for Global SDG Indicators", online available at <https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/>
3. SESRIC (2018), "Tendency Survey on SDG Priorities of OIC Member Countries", online available at <http://www.sesric.org/activities-announcements-detail.php?id=386>
4. SESRIC (2018), "Results of the Tendency Survey on SDG Priorities of OIC Member Countries (October 2018 Edition)", online available at <http://www.sesric.org/publications-earchive.php?year=2018>
5. UNESCAP (2017), "Measuring SDG progress in Asia and the Pacific: Is there enough data?", online available at <http://www.unescap.org/sites/default/files/publications/ESCAP-SYB2017.pdf>
6. UNSD (2019), "Global SDG Indicators Database", online available at <https://unstats.un.org/sdgs/indicators/database/>

7. SESRIC (2016), "Moving from MDGs to SDGs: Prospect and Challenges for OIC Member Countries", online available at <http://www.sesric.org/files/article/568.pdf>



Morocco's experience in South/South cooperation and statistical capacity building



Belkacem Abdous

National Institute of Statistics and Applied Economics, Rabat, Morocco

Abstract

During the last decades, Morocco has developed very active and intense South-South statistical cooperation activities. Indeed, the High Commission for Planning provided statistical training, technical assistance and capacity building support for several National Statistical Offices of African and Arab countries. In this work, we will briefly present some of these fruitful triangular and bilateral cooperation initiatives and show how they build on the specific local context to adopt and apply the international standards and good statistical practices. Besides, we will cast light on the manifest will of Morocco to disseminate and open its data bases to users and international organizations. We will share and discuss some innovative approaches that have been put forward to modernize its statistical system in support of the SDGs.

Keywords

South-South cooperation, Triangular and bilateral cooperation, Statistical capacity building, Modernization of national statistical systems

1. Introduction

South–South cooperation is a concept that emerged during the 1940s, it finds its roots in the Asian and African countries struggle for independence and in the Non-Aligned Movement (1955). Generally speaking, the word “South” refers to countries outside Western Europe and North America. For more details, see Bergamaschi, Moore and Tickner (2017).

There is no precise and internationally accepted definition of South-South Cooperation. In a document entitled “ South-South Cooperation: A Pathway for Development”, the Partners in Population and Development (2009), an Intergovernmental Alliance of developing countries, provides commonly used concepts and definitions of South–South cooperation. Detailed milestones in South–South cooperation are provided as well. That said, according to the Nairobi outcome document of the High-level United Nations Conference on South-South Cooperation (General Assembly Resolution 64/222 of 21 December 2009), South–South cooperation principles might be described as follows:

“South-South cooperation is a common endeavour of peoples and countries of the South, born out of shared experiences and sympathies, based on their common objectives and solidarity, and guided by, inter alia, the principles of respect for national sovereignty and ownership, free from any conditionalities.

South-South cooperation should not be seen as official development assistance. It is a partnership among equals based on solidarity. In that regard, we acknowledge the need to enhance the development effectiveness of South-South cooperation by continuing to increase its mutual accountability and transparency, as well as coordinating its initiatives with other development projects and programmes on the ground, in accordance with national development plans and priorities. We also recognize that the impact of South-South cooperation should be assessed with a view to improving, as appropriate, its quality in a results-oriented manner.”

There are mainly three dimensions in South-South Cooperation concepts: political, economic and technical. Its principles encompass all sectors of international relations. In this work, we will put emphasis on the very specific South-South Statistical Cooperation and review Morocco's experience in terms of cooperation with African countries.

2. South-South Statistical Cooperation

The recent emergence of several development initiatives and roadmaps, such as the Millennium Development Goals (MDGs), the Poverty Reduction Strategy (PRSs), the International Comparison Program for Africa (ICP-Africa), the Agenda 2063 for African development, the Sustainable Development Goals (SDGs, Agenda 2030), etc., have cast light on the crucial need for African countries to produce their own, better quality, relevant and timely statistical information. The availability of more and better quality statistical data is of course vital for monitoring such programs and for supporting the implementation of evidence-based public policies and development needs, while putting a lot of pressure on National Statistical Systems. A consequence of this unprecedented demand is the intensification of bilateral and multilateral aids and technical support to Africa in statistics. Indeed, during several decades, a plethora of initiatives and programs have been put forward to help and strengthen statistical capacity building in “south” countries in general and in Africa in particular. The main interventions/involvement concern: Statistical Systems, National strategy for Development of statistics (NSDS), Statistical Capacity, surveys and censuses, national accounts, registers and manifold trainings. These supports are provided by some key players such as:

- The Partnership in Statistics for Development in the 21st Century (PARIS21)
- UN Economic Commission for Africa (UNECA)
- AFRISTAT
- African Development Bank (AfDB)
- United Nations Statistics Division (UNSD)
- World Bank Development Data Group (DECDG)

Of course, many other agencies and international organizations provide technical assistance and support in statistics fields. It is worth noting that despite all these aids, the statistical systems capacity in Africa varies dramatically from country to country and is still weak in some countries, see Wingfield-Digby (2007). For a critical analysis, a review of key stages in the development of data and statistics in Africa together with a review of the roles of the involved institutions and agencies, we refer to the following papers and reports: Bergamaschi, Moore and Tickner (2017), Della Faille and La France-Moreau (2013), Lehohla (2008), Partners in Population and Development (2009), Round, (2012).

3. Morocco's experience in the South-South Statistical Cooperation

First, let us recall that the first statistical structures of the Kingdom of Morocco were set up in 1942 with the creation of a central statistics service. In order to have reliable statistical information and to coordinate and harmonize the different statistical activities, a national coordinating committee for statistical studies was created in 1959. Also, to meet the growing demand and needs of statisticians, a training center for statistician engineers (*National Institute of Statistics and Applied Economics (INSEA)*), was created in Rabat in 1961. Later on, the national statistics office has been reorganized under the supervision of the High Commission of Planning in a central and regional statistics directorates that oversee the collection, compilation, extraction and release of official statistics relating to the demographic, social, environmental, economic and general activities and conditions of Morocco. In addition to the statistical information produced by the High Commission of Planning, many other ministries and sectoral administrations produce statistics on a day-to-day or regular basis.

Given the increasing external and domestic demands in terms of data, studies, analyzes and simple or composite statistical indicators, the National Statistical System has put the use of digital technology at the heart of its business model. Indeed, for instance, during the last two decades, the High Commission for Planning (HCP) adopted several models, software and various technological applications such as the Automatic Document Reading techniques, the Computer Assisted Data Collection system, the integration of Satellite Imagery and the Geographic Information System, etc.

Started in January 2019, an undergoing digitization process of all the statistical production lines is taking place within the framework of a collaborative work platform, with active participation and valorization of all human and institutional resources. This platform will disseminate and open HCP's data bases to users and international organizations. In summary, the national statistics system of Morocco has a good experience and long history in official statistics with a continuous mutation and modernization process.

As a matter of fact, in the context of the regional integration in Africa and following the recent accession of Morocco to the African Union, South-South Cooperation constitutes a priority for foreign policy. It is included in the Moroccan Preamble to the Constitution of 2011. Morocco is deeply engaged in several initiatives promoting South-South Cooperation and triangular cooperation with important investments.

Regarding official statistics, Morocco has developed strong relationships and collaborations with many National Statistics Offices of Arab and African countries. These cooperation activities are usually organized as study visits (8-10 visits per year) and training courses. The provided technical assistance covers many fields such as national account, household surveys, statistical indices, poverty measurement, census, etc. It builds on the specific cultural and local contexts while complying to the international standards and good statistical practices.

Another dimension of Morocco's South-South Statistical experience takes the form of "triangular cooperation" that involves The High Commission for Planning and its partners. Some examples of triangular cooperation are:

- The French National Institute of Statistics and Economic Studies (INSEE):
 - Maghreb activities to promote expertise and experience exchanges,
 - Assisting the African countries to use ERETES, the main cooperation tool of Eurostat and the French Co-operation in the area of National Accounts.
- Arab Institute for Training and Research in Statistic (INSEE): mainly statistical training for Arab countries
- MEDSTAT: Imbeds the European Union international statistical cooperation with Mediterranean countries in the context of the European Neighborhood Policy (ENP).

The MEDSTAT program offers to south Mediterranean countries exchange of expertise and technical support in several official statistics fields. In its early stages, south National Statistics Offices were considered as beneficiaries of European assistance. Then, the nature of this cooperation shifted toward a new paradigm / approach that gives more responsibilities to south countries in terms of identification of their priorities, implementation and management of

the program, etc. The High Commission for Planning developed a good experience in this context and played a key role in the MEDSTAT programs which have gone through 4 important phases.

- MEDSTAT-I (1996-2003): during this first phase Morocco was only a beneficiary of the support offered by the program
- MEDSTAT-II (2006-2009): this phase focused on the improvement of statistical services quality. Morocco was deeply involved in these activities.
- MEDSTAT-III (2010-2013): a third phase which pushed forward and promoted evidence-based policymaking. Moroccan experts contributed by 36% to the total working days of short term experts. Besides, the High Commission for Planning was an active member of the MEDSTAT-III Consortium in charge of the program management.
- Transition phase (2013-2016): Morocco led 3 working groups, namely, Energy, Transport, Trade and Balance of Payments.
- MEDSTAT IV (2016-2019): Morocco shared the lead coordination of Energy working group with Tunisia and Egypt, and led the External Trade and Balance of Payments working group as well.

Finally, let us mention the technical assistances provided by Morocco to Palestine in the fields of external trade statistics to build online dynamic tabulation software for external trade statistics data, and to Libya to produce freight trade statistics that comply with the international standards. Several francophone African countries have benefited from technical assistance of Morocco's statistical expertise as well: Benin, Cameroun, Comoros, Côte d'Ivoire, Gabon, Mali, Senegal, Togo, etc.

References

1. Dimitri Della Faille, Valérie La France-Moreau (2013). Current and Upcoming Challenges in South-South Cooperation in the field of Social Statistics. 2013. hal-02046911 <https://hal.archives-ouvertes.fr/hal-02046911>
2. Isaline Bergamaschi, Phoebe Moore and Arlene B. Tickner (2017). South-South Cooperation Beyond the Myths, International Political Economy Series, DOI 10.1057/978-1-137-53969-4_13.
3. General Assembly Resolution 64/222 of 21 December 2009, (https://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/64/222)
4. MEDSTAT Program, http://ec.europa.eu/eurostat/statistics-explained/index.php/MEDSTAT_programme
5. Lehohla, P. (2008). 'Statistical Development in Africa in the Context of the Global Statistical System'. Paper prepared for the 39th Session of United Nations Statistical Commission <http://unstats.un.org/unsd/statcom/doc08/BG-AfricaStatDev.pdf>

6. PARIS21: The Partnership in Statistics for Development in the 21st Century. <http://paris21.org/> . It was founded jointly by the UN, EC, OECD, IMF, and the World Bank in November 1999.
7. Partners in Population and Development (2009) "South-South Cooperation: A Pathway for Development" (http://partners-popdev.org/docs/PPD_South-South_Book.pdf)
8. Round, J.I. (2012) Aid and Investment in Statistics for Africa. WIDER Working Paper 2012/093. Helsinki: UNU-WIDER.
9. Wingfield-Digby, P. (2007). Towards Reforming National Statistical Agencies and Systems: a Survey of Best Practice Countries with Effective National Statistical Systems in Africa. The African Capacity Building Foundation, Harare.
<http://www.pwdigby.co.uk/pdf/ACBF%20Best%20Practice%20Study%20-%20BPP%20No.1%20STATNET,%20Sept%2020071.pdf>



Modernization of the Tunisian Statistical System and its impact on statistical production for SDGs



Saidi Hedi

Arab Institute for Training and Research in Statistics

Abstract

The new political reality in Tunisia has introduced radical changes across different fields, resulting in many new challenges facing the National Statistical System (NSS), which were mainly related to the increase of the user's demands for statistical information, especially after the adoption of the new law on access to information. The need to consolidate an image of the credibility and independence of the statistical system, to improve data quality, to put in place more dynamic systems of information dissemination and to preserve confidentiality in accordance with international standards, mandates that the statistical system be renewed. To overcome the challenges and modernize the Tunisian statistical system, most notably the pivotal structure embodied in the National Institute of Statistics (NIS), with a view to meet the challenges and new requirements of SDGs, a twinning project was launched with European statistical institutions for the period 2016-2018. The main objectives of this program are:

- The adoption of a new statistical law;
- The implementation of the Tunisian Statistical Charter (TSC);
- The revision of the organization chart of the NIS-Tunisia;
- The adoption of the SCN2008;
- The development of regional statistics;
- The establishment of a quality management system;
- The development of a communication strategy.

We will introduce the current situation of the Tunisian Statistical System, as well as its strengths and weaknesses, the results of the twinning program and their impacts on statistical production for SDGs and will focus on work progress to initiate the implementation of the 2030 Agenda.

Keywords

The National Statistical System; Statistical modernization; quality management; SDGs.

1. Introduction

Tunisia has had a satisfactory statistical production system for a very long time, making it one of the most developed countries in Africa and the Arab world. The National Statistical System (NSS) is highly centralized as all

programming and planning, design, processing, dissemination, statistical information management, and coordination work takes place centrally. The role of the regional level is reserved solely for data collection and data entry. Statistical production is based largely on traditional sources such as surveys and censuses, both within the National Institute of Statistics (NIS), the main producer in the country, and in other statistical structures (SSP). Despite the advances made in the field of statistics, both in terms of production and dissemination, the system still suffers from some shortcomings and problems, thus necessitating a modernization of the production structures, including the INS. This modernization is also dictated by the challenges faced by the INS, particularly in relation to the country's commitments regarding the various development agendas. It is in this context that the country embarked on a heavy twinning project with the European Union (EU) in 2016, involving European statistical production institutions to support and assist the NSI in priority areas for modernization of the Tunisian statistical production apparatus and overcome, therefore, the various deficiencies observed.

We will try in this paper to focus on the functioning of the NSS, its strengths, its weaknesses and the main axes apprehended within the framework of the twinning. We will also focus on the achievements of this twinning and other work to modernize the Tunisian statistical system while focusing on the NIS.

2. Methodology

I. Challenges of the NSS

Recognizing the inadequacies and shortcomings of the NSS, particularly at the NIS level, as well as the challenges to meet a growing demand for statistics, the NIS conducted a series of internal and external evaluations to identify strengths as well as the real problems and dysfunctions that hinder the smooth functioning of the statistical production process in the country. In this context, a peer assessment was carried out by the EU based on a request from NIS Tunisia, which included two main components:

- The administrative and technical capacity of the statistical system to respond to different requirements, focusing mainly on the effectiveness of the NSS, including the NIS, the quality of statistical production and technical coordination between the different actors;
- The legislative framework governing the statistical activity through in-depth examination of the 1999 Statistical Law, in particular as regards compliance with the fundamental principles of official statistics as well as international standards and best practices.

The main recommendations of this evaluation are as follows:

- Revision of the legislative framework governing statistical activities to ensure good governance of the national statistical system;

- Strengthening the NSS coordination activities to increase the accountability of various actors;
- Supporting the NIS to exercise its responsibility for technical coordination and quality management to fully play its role of assisting other statistical producers in the entire process of statistical production;
- Revision of the procedure for drawing up the national statistical program;
- Strengthening the technical capacity of the NIS to assume its role as the central executive body of the NSS and its technical coordination responsibility.
- The twinning program with the EU has taken into consideration only a part of these recommendations judged by the experts and the various evaluations as priorities. NIS and other SSP have continued work on the other components based on other tools and mechanisms for technical and financial assistance.

II. Legislative and institutional framework of the Tunisian statistical system

The NSS consists of four actors: the NIS, the National Council of Statistics (NCS), the SSP and the training institutions. It is governed by the 1999 law and the organizational decrees of the NSS as well as certain circulars and memos organizing the statistical activity in the country. The Tunisian legislative framework has enabled the various statistical production structures to better manage and govern the statistical activity in the country, particularly concerning the clarification of powers and the partial adherence to the fundamental principles of official statistics. However, despite the strengths of this law, the assessments have led to some shortcomings and difficulties, including two important aspects. The first aspect concerns adherence to the fundamental principles of official statistics and the African Charter on Statistics as well as EU good practices, mainly in relation to the professional independence of statisticians and production structures. The second aspect concerns the NSS governance, since it has not been able to guarantee a clear division of responsibility for the coordination and management of the system between the main actors, notably the NCS and the NIS. The current organization has created major dysfunctions that limit the effectiveness of the NSS and the quality of its productions. These dysfunctions have negatively impacted the effectiveness of the NIS and the strategic role of technical coordination by limiting the necessary initiatives to develop the SSP. As for the NCS, the dysfunctions noted in terms of coordination are explained by the large number of SSPs (48 SSPs), which makes it very difficult to draw up a National Statistics Program. Access to micro data for research and in-depth

analysis is also a very serious limitation for the NIS, and is even threatening its relationship with users, including the government.

To overcome all of these difficulties and dysfunctions, and with the aim of reinforcing, in particular, adherence to the fundamental principles of official statistics (professional and scientific independence, quality, dissemination, etc.), a commission has been created within the NCS bringing together all the actors of the system (producers and users, including civil society) and a bill has been proposed to revise the current 1999 law. This draft law is inspired by the framework law developed by the UN and the African Charter on Statistics. The support of the twinning project was mainly in the organization of visits to France and Italy to see closely the functioning of these systems, as well as the presentation of strengths and weaknesses of the statistical laws of certain countries in terms of statistical governance and compliance with the European Code of Good Practice and others. The main novelties of this new law are:

- Adherence to the fundamental principles of official statistics, in particular the professional independence of statisticians and institutions responsible for statistical production;
- Reviewing the composition of the NSS by reducing the number of SSPs and the scope of official statistics;
- Provision of micro-data for research purposes while ensuring commitment to statistical confidentiality;
- Formalization of a procedure for drawing up the statistical program for better coordination of the NSS.

Other measures have been taken in this context, including the start of the implementation of the Tunisian Statistical Charter at NIS level, which aims to adhere to the fundamental principles of official statistics and the African Charter on Statistics. The work focused on developing a set of indicators of evaluation and self-evaluation indicators and a roadmap for its implementation. A highly participatory approach is followed to better concretize the principles of the charter and encourage adherence of public statistical structures to them.

III. Statistical production: Developing and strengthening economic statistics

Statistical production in Tunisia has experienced a spectacular development in recent years encouraged by the opening of the NIS on its environment, in particular the main users of statistical information including government, researchers and civil society ... etc. In this context, it is important to note that the demographic and social component is relatively covered despite the shortcomings noted, particularly in terms of the quality and dissemination of certain demographic indicators. Nevertheless, economic statistics continue to experience significant gaps in the quantity and quality of

statistical production in the country. This part of the paper will focus on the strengths and weaknesses of economic statistics and the new developments.

Economic statistics include both annual national accounts, income statistics and short-term and foreign trade statistics from different sources and from several producers. The production of these statistics within the NIS has significantly improved in recent years, particularly with regards to short-term statistics, placing Tunisia at the top of African countries and similar countries. However, the international environment has changed with a revision of international classifications and the adoption of the 2008 SNA and the Balance of Payments Manual No. 6. The political changes that took place after the revolution in 2011 also had a great influence on the need for economic statistics on the one hand and for the structure of the economy on the other hand because the 1997 base became completely obsolete.

To meet these requirements and to overcome the shortcomings observed, the NIS has emphasized, in the context of the twinning project with the European institutions, the activities relating to the change of the base year to better reflect the economic reality of the country, and to introduce the recommendations of the 2008 SNA progressively according to available statistics and to produce other statistics on this occasion if we need that. The NIS has made a considerable effort in this context to improve the map of available indicators, particularly in relation to the SDGs at first, and to improve the quality of production of these indicators in a second step. A significant number of activities have been programmed and taken into consideration as part of the project to meet these expectations. These activities are summarized as follows:

- Choosing a base year and transforming to the new system of national accounts of the United Nations 2008 SNA, using a highly participatory approach that takes into account the availability of statistics, as well as the exploration of other sources outside the NIS;
- Development of a methodology for regional distribution of GDP and the calculation of regional GDP according to the production perspective, in response to the increasing demand of regional actors for planning and programming purposes;
- Measuring the contribution of non-profit institutions and the informal sector, focusing on the proposal of a method for estimating the share of the informal sector in the economy and expertise of the non-profit sector;
- Development of flow and wealth financial accounts by evaluating available sources and methods used for flow-through financial accounts;
- Development of derivative accounts and environmental statistics with a focus on water satellite accounts, methodology for drawing up

environmental protection accounts, and exchange of environmental data between different stakeholders.

The hard work and research undertaken within the NIS by the different think tanks have resulted in concrete activities and achievements as well as projects for which the NIS is called upon to invest more according to a roadmap already identified by the experts. Indeed, the transfer to a new base year for a new generation of national accounts considering the main recommendations of the 2008 SNA requires the mobilization of a large body of data from different sources such as surveys, administrative data, new surveys, etc. Some information calls for preparatory work and budgetary programming, as well as a strong commitment from all stakeholders, because this project is national in nature and far exceeds the competence of the NIS. Despite all of these constraints, the NIS, in partnership with the main partners, can achieve most of the set objectives, including:

- A new base year of accounts has been defined (2015-2016) and work has progressed to develop accounts according to this new base year and well-identified roadmap;
- The recommendations of the United Nations 2008 SNA have been thoroughly reviewed and concrete steps have been taken to ensure the best possible transition to National Accounts in compliance with this system;
- A demand decomposition approach has been implemented and work is well in advance;
- The calculation of the regional GDPs for the chosen year (2013) has been finalized and validated by the experts of the NIS at first, and then revalidated by other experts, mainly academics who work and master these aspects. The main conclusion from all the work on this file is that the first sector results are consistent with what is observed on the ground;
- The water satellite accounts for 2010 and 2015 have been developed to measure imbalances between limited potential and exponential demand;
- The NIS has produced a publication on environment statistics which aims to improve the quality and scope of data with dissemination of key statistics in a single medium.

The work of relining the national accounts and considering the recommendations of the SNA 2008 is an absolute priority not only for the NIS, but also for all economic and social actors, to better reflect the economic reality and the weight of the sectors in the national economy. Highly sensitive issues, such as the informal economy and the regionalization of GDP and VA, have been treated and rigorously apprehended for the first time in Tunisia. The NIS will continue some activities for a period of two to three years

depending on the completion of certain surveys identified by experts and professionals in business statistics.

IV. Reorganization of the NIS

The current organization of the NIS has shown its limitations in governance of the process of statistical production and coordination with other statistical information producers. This limitation is most notable at the regional level where the NIS only assumes the role of data collection from surveys and some administrative sources. To solve these problems and their disruptiveness to the smooth running of the production activities, a new organization of the NIS at the central and regional level has been implemented to guarantee a quality and continuous statistical production and to meet the requirements of the development statistics agendas. This new organization is in line with the Generic Statistical Business Process Model (GSBPM) and has made it possible to organize the regional directorates into three areas: collection, production and dissemination. At the central level, new functions, such as methodology, quality and training have been created. Other activities have also been carried out including:

- Launch of ISO 9001 certification process for the salary-employment survey and organization of a hard-core training within the NIS on the implementation of an ISO project';
- A new integrated information system architecture has been proposed, taking into account the NIS environment in accordance with the GSBPM. The roadmap for the implementation of this new integrated information system has been established for a pilot area (external trade statistics).

This first reorganization of the technical services of the NIS will certainly have an added value on the progress of the statistical work, but is still insufficient in relation to the enormous challenges that the establishment will have to meet the important needs for the SDGs and the African Union Agenda 2063. The NIS is expected to continue these efforts of openness on its environment and to better manage its relation with the users.

3. Result

The main results of this work are summarized as follows:

- Proposal of a new law revising the current 1999 law with the aim of reinforcing adherence to the fundamental principles of official statistics and in line with European and international good practice,
- Start up the change of the base year to better reflect the economic reality of the country, and introduction of the recommendations of the 2008 SNA progressively,

- Development of new methodologies for regional GDPs and informal sector,
- Improvement of quality of environment statistics and scope of data with dissemination,
- Proposal a new organisation of the NIS based on the international standards,
- Improvement of the quality of communication and relation with users.

4. Discussion and Conclusion

In conclusion, it is very important to note that the general objective is "to promote the establishment of a coherent, efficient and permanent public statistical information system", and the work carried out in the framework of the twinning has enabled the consolidation of the NIS activities in line with European and international good practice. Indeed, a plan for implementing the Tunisian Statistical Charter, in line with European recommendations and the African Charter on Statistics, has been proposed and validated by the NIS. The NIS staff have been trained in international statistical standards such as GSBPM and SDMX. An architecture of a new integrated information system has been proposed and tested in a pilot area related to foreign trade statistics. The NIS staff have been trained in quality according to European standards and a roadmap for ISO certification has been defined for a pilot survey. National accounts have become compliant with the international standard of SNA 2008. The work to change the base year (2015 and 2016) has been initiated and requires the development of current sources, the mobilization of new sources and the updating of the nomenclatures of activities and products. The calculation of regional GDPs has been finalized for the 2013 reference year.

The list of statistical domains where the NIS has come close to EU standards and international standards are important including the Tunisian Statistical Charter, national and regional accounts and reorganization according to the GSBPM standard, ISO 9001 certification process for Wage-job survey and SDMX standard integration.

If the results foreseen in the Twinning Contract are achieved, several activities remain to be achieved. This is mainly the transition to the 2008 SNA, the base year change and the activities for which roadmaps have been established and must now be implemented. This involves, for example, setting up a quality unit and a training unit to certify a statistical process and set up an integrated information system.

References

1. Arrêté du Ministre du Développement et de la Coopération Internationale du 2 juin 2010, fixant les modalités de transmission des informations disponibles auprès des administrations et des structures publiques à l'Institut National de la Statistique, à des fins exclusivement statistiques.
2. Décret n° 2005-1643 du 30 mai 2005, fixant l'organigramme de l'Institut National de la Statistique, modifié et complété par le Décret n° 2005-2857 du 24 octobre 2005
3. Décret-Loi n° 2011-41 du 26 mai 2011, relatif à l'accès aux documents administratifs des organismes publics,
4. Global Evaluation of the National Statistical System of Tunisia, National Institute of Statistics Tunisia, 2014
5. Loi du 13 avril 1999, relative au Système National de la Statistique (appelée plus loin « Loi statistique ») Le Décret n°99-2797 du 13 décembre 1999, modifié en dernier ressort par le Décret n° 2004-2659 du 29 novembre 2004, fixe la composition, l'organisation et les modalités de fonctionnement du Conseil National de la Statistique.
6. Loi organique n° 2004-63 du 27 juillet 2004 sur la protection des données à caractère personnel,
7. Modernization of the Tunisian statistical system, final report, 2018



**Modernization and monitoring SDGs in area of
conflicts or in fragile conditions:
Case study, "The Palestinian Central Bureau of
Statistics"**



Ola Awad

Palestinian Central Bureau of Statistics (PCBS), Ramallah, Palestine

Abstract

On the 1st of January 2016, the world officially began the implementation of the 2030 Agenda for Sustainable Development, this agenda that took the slogan of "leave no one behind" is most needed for the conflict and fragiled countries. The national agencies in-cooperation with their partners in countries affected by fragility and conflict need to prioritize and sequence the efforts of SDGs specially SDG 16 in ways that take account of these countries realities as well as ensuring the most suitable outputs for it. It is agreed that the National Statistical Offices take a significant role in Measuring and Monitoring the Sustainable Development Goals; particularly it is expected to serve as national focal points for statistics for SDGs to deal with this process at local level with national partners as well as within the framework of international cooperation with different countries or organizations. Modernization of SDGs monitoring and measuring process is much needed in the conflict and fragile countries; it will affect the whole system to be modernized in the fields of managing the available capacities and resources in light of the national priorities and needs within the framework of the SDGs, as well as the high need for the national owned process to build trust among all different actors and will help to implement the SDG17 as well. Moreover, it will present the different set ups of inter-relations among partners and the understanding of building partnership. The main principle for monitoring the SDGs in conflict and fragile countries is to select the core set of indicators to track the progress of the SDGs in cooperation with the international community and that fit the context of these countries, in addition this set will be complemented with national owned indicators that reflect own needs and priorities. One of the main issues is how to review and harmonize the existing mechanism among all partners at national, regional and international levels for the purpose of reducing the burden through sharing of relevant methodologies and standards, this will be mainly achieved by reaching consensus on the common indicators for these countries as well as developing agreed methodologies for the national ones. To meet such requirements, the review will touch base the whole statistical system that is led by the official statistical officies in the conflict and fragile countries in wide coordination with other national partners and in cooperation with developed countries and expertise of the international agencies. This paper seeks to to address

measurements taken by PCBS to overcome challenges and turn weaknesses into strengths in regards to modernization of statistical system in fragile environments. PCBS hopes that the paper shall be beneficial for NSOs operating under similar situations.

Keywords

SDGs; Partnership; Area of Conflicts; Modernization; Monitoring

1. Introduction

Countries are faced by conflicts and fragile conditions almost everywhere; However the types and sizes of conflicts were different during the last few decades, before that we were talking about a limited number of countries, but now the number of countries with conflict is increasing sharply, conflict creates special contexts in the countries of extreme socio- economic- environmental conditions represented mainly by poverty, high unemployment rates, weak economies, in addition to weak institutions, lack of resources and deteriorated infrastructure; which lead to the ranking of statistics as a least priority, decreased statistical awareness and increased statistical illiteracy. In reality strong statistics is considered as written history by providing accurate figures despite being under conflict leading to evidence based interventions and assure monitoring for vital life areas which will help not only in making life better but also in saving lives.

After the international recognition of SDGs under the slogan of “leave no one behind”, there's a differentiation from previous plans and agendas especially MDGs for taking the dimension of conflict in the goal 16 for the ways of dealing with such countries, and goal 17 in enhancing partnership to help in achieving development and bridging gap between different countries; which makes statistics and providing the needs of SDGs as a necessity and opportunity to create change under fragile and complex conditions, and to boost human rights.

The requirement of SDGs represents a huge challenge for every statistical system, and the challenge is duplicated for the countries which suffer from conflicts and fragile conditions, and the way for achieving such huge needs is to modernise and strengthen statistical systems and statistical institutions to create an efficient monitoring system that lead to evidence based decisions, and intervention to reach development for every person.

Modernization is defined as the way for developing, changing and improving methodologies, tools, using new technologies, knowledge and statistics science to “Increase the efficiency of statistical processes to produce outputs that better meet user’s needs¹”, and the Keys for achieving

¹ MacMillan. A.L (2015). Modernizing official statistics.

modernized National Statistical System (NSS) is mainly described from PCBS case:

- ✓ **Enhancing institutionalised national statistical system (NSS)**, PCBS developed a statistical national strategy 2018-2022; which links statistical programs to national agendas and priorities of SDGs and improving the statistical system environment, investing in administrative records, and IT infrastructure.
- ✓ **Adapting and applying the Cape Town Global Action Plan** especially the objectives; "Strengthen national statistical systems and the coordination role of national statistical offices", "Modernize governance and institutional frameworks to allow national statistical systems to meet the demands and opportunities of constantly evolving data ecosystems", and "Modernize statistical standards, particularly those aimed to facilitate data integration and automation of data exchange across different stages of the statistical production process".
- ✓ **Stakeholder's Engagement as a Tool of Modernization of the Statistical System**, the importance of stakeholders' engagement comes from their roles and responsibilities in developing the NSS, in relation to different dimensions:
 - Sustainability: Assure an institutionalized culture of statistics, increase Frequency, coverage, and timeliness².
 - Financing: Reducing cost of data collection based on fieldwork surveys, resilience to the exogenous agenda of donors.
 - Human resources: Through interaction, sharing experience, reduce labor intensive in data collection and response burden.
 - Technology: Harnessing the power of technology to enhance a solid statistical infrastructure
 - Stakeholder's scope: Inclusion of new players such as the media, academia, and private sector.
- ✓ **Platforms as an important mean for data dissemination and monitoring**, according to Dubai declaration, the key factors of platforms where determined as: First with data Interoperability, which "addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data"³, and it is considered as one characteristics of good quality data. Open Data, which assure the transparency in all process of producing and dissemination, methodologies, metadata and access.

² UNECE. (2016). Using Administrative and Secondary Sources for Official Statistics.

³ Data Interoperability Standards Consortium

In addition; PCBS took the responsibility to modernize the national statistical system as a whole and create strong monitoring system for SDGs, by working on different levels, institutional level, national, and international levels, the following milestones represent the journey of change:

I. Institutional Level (Strong statistical system)

PCBS is responsible for official statistics and to provide statistical requirements of SDGs agenda, it was important and critical to take a revolutionary way to continuously strength and modernise statistical system specially under rapidly changing conditions in conflict and crises, clear and steady steps were taken as follows:

- Special general directorate was created “Registers and Statistical Monitoring Directorate” with main task is to fulfill the statistical requirements of SDGs, developing national monitoring system, and working on improving the use of new data sources.
- Investing in IT infrastructure and moving to electronic data collection starting with full electronic compilation of the Population, Housing, and Establishments census, 2017.
- Geographical information system (GIS), statistical data are linked to GIS for the importance of geographical dimension in statistics to assure the coverage of remote and vulnerable areas in addition to analytical aspects.
- Adapting the Generic Statistical Business Process Model (GSBPM)⁴– “This model describes the core business processes undertaken by statistical organizations to produce statistical outputs. It can help statistical organisations to find common processes across organisation to reduce inefficiency and redundancy”.
- Creating platforms, which deals with (Production, Process, and Dissemination) with cooperation with international institutions.
- Changing capacity building methods programs and area of interest to move to open data, big data, and data science.
- Creating a communication strategy which is aimed to enhance and improve the way of communicating statistics to create an atmosphere for statistical literacy and in a way that lead to better use of data by users.
- Investing in dissemination tools to create the culture of statistics and reach different groups for example PCBS conducted a theatre play which addressed gender discrimination issues, two rounds of school competitions were organized to raise statistical awareness for school, students, videos depending on real testimonies to wide spread statistics awareness.

⁴ UNECE. (2018). Report of the United Nations Economic Commission for Europe High-level Group for the Modernisation of Official Statistics.

II. National level (Institutionalisation)

On the national level, PCBS worked hard to enhance the capacity of the national statistical system, and enhance the stakeholders' engagement in production, and dissemination of statistics and not only the classical role related to use of statistics as follows:

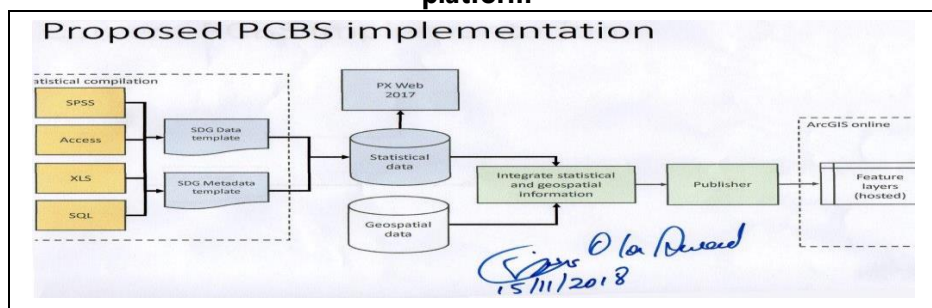
- Creating a high-level committee for SDGs headed by Prime Minister Office responsible for aligning SDGs with national priorities reflected in the National Policy Agenda 2017-2022.
- Creating a national technical team headed by PCBS responsible to fulfill the requirements of national priorities of SDGs.
- PCBS Advisory Council for Official Statistics headed by the Prime Minister, which increases the awareness, support, and common understanding of statistics for politicians and decision makers which lead to securing government support and financing with statistical independency.

III. International level (Partnership)

Being Part of the international dialogue and strong partnership with different international institutions represented a learning process and evolutionary changing process to reach new eras and boost modernisation:

- PCBS is a member of the High-level Group for Partnership, Coordination and Capacity-Building for statistics for the 2030 Agenda for Sustainable Development, which put Palestine as key players in seeking change and development.
- Best practices exchange between national statistical institutions, by twining procedure, a good example for that is with ISTAT which helped in creating the mapping matrix of SDGs, and creation of the "Palestine SDGs Data Dissemination Platform".
- Open data watch report, PCBS was ranked as 56 on the world in meeting data openness and 4th in the Middle East Region, which is done by working on enabling and improving means and coverage
- UNSD-DFID project platform, allows the dissemination of data for better monitoring of SDGs as reflected in figure 1:

Figure 1: PCBS propose implementation process for UNSD-DFID project platform



- ESRI ArcGIS Open data Platform

Palestine is one of the initial participating countries in this platform in addition to Morocco, Qatar, UK, Tanzania, Colombia, Canada, Brazil, Kenya, Ireland, Mexico, Philippines, S. Africa, and Senegal, this modern technology platform enables the overlay and integration of multiple national and global data sets to better understanding of the relationships, interlinkages, and to address inconsistencies. It also engages communities of interest of SDGs with country owned policy initiatives by geographical dimension.

Figure 2: Palestine interface for ESRI ArcGIS Open data Platform



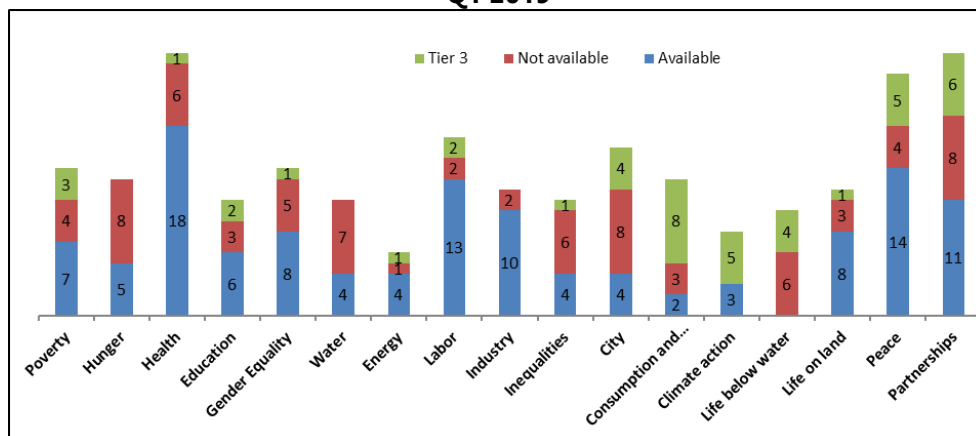
In addition to the geographical dimension, the map story element was developed to empower the global users' community to organize all knowledge about the world spatially. Map story is an infrastructure for enabling "MapStorytelling" as means of communicating important issues to global and national users with friendly manner. PCBS started to produce map stories which covered different areas such as early marriage, labor market inequalities and poverty taking the Palestinian conflict context into the analysis process such as refugee camps, Palestinian areas fully controlled by the Israeli occupation.

PCBS achievements on SDGs:

Despite hard conditions related to Palestinian situation and lack of resources, PCBS took the opportunity of SDGs to achieve improvements by:

- Defining national priorities with the government.
- Creating SDGs mapping matrix, which assess where is Palestine from SDGs, and figure 3 explains indicators availability by goal as update at the end of first quarter, 2019.

Figure 3: Situation of Palestine from SDGs indicators by goal and availability, Q1 2019



- Localization of SDGs metadata, metadata reflected in a way that is related to the local context specially using the Arabic language.
- Creating an SDGs web data base "Palestine SDGs Data Dissemination Platform", which created in cooperation with ISTAT, as interactive data base and platform and the most important element that it is established in Arabic and English to reach national and global users.

Challenges:

- Current data sources are not sufficient to cope with accelerating needs and demands for statistics to measure society under conflicts and meet SDGs requirements.
- Lack of resources bring to the surface the necessity of prioritisation and enhances cooperation, especially south to south.
- Dealing with unstable, rapidly changing environment, which lead to reassess needs and resources frequently.
- Working under pressure, where users specially decision makers ask for data in short time for humanitarian issues that is mostly not linked to long term development needs data.
- Weak institutions and lack of institutionalised data collection systems increase the burden and cost of data collection.
- Maintain cooperation of data providers specially households during the peak of conflicts or crises.
- Data confidentiality and statistical independency still the highest priority under collecting data under conflicts and unstable environment.
- Availability of appropriate IT infrastructure and lack of local experts to meet accelerating IT revolution and data revolution.

- Rapidly changing context imposes rapidly changing Indicators, and this requires improved methodologies, new means and ways of data sources and data collection.

2. Discussion and Conclusion

Modernization is considered as a continuously updated process with no end specially under rapidly changing and fragile conditions, those issues must be taken into consideration to succeed in this long journey:

- Setting a special strategy and action plan with clear mechanisms
- Starting with SDGs assessments, for current situations analysis, priorities, and technical and financial needs.
- Enhancing stakeholders' engagements on both the national and the international level which will shorten the way and decrease the cost.
- Creating platforms that meet the keys of openness and interoperability as tool of monitoring and ensure transparency.
- More responsibility on NSIs towards development, Ownership of the Statistical System: Roles and Responsibilities must be well defined and monitored.
- Investing in capacity building and concentrating on new fields such as big data, data science, and IT.
- Any initiative must be country owned and reflecting the national special context.

To meet such requirements, the modernization must include the whole statistical system that is led by the official statistical offices in the conflict and fragile countries in wide coordination with other national partners and in cooperation with developed countries and expertise of the international agencies.

Stereotyping in producing statistics has changed with different sources, future requirements and shifting towards data science with increasing all ways and means of communication, In addition, what's really counts that being a country under conflict does not mean to be behind developments process, but to turn challenges into opportunities towards change and equal development.

References

1. MacMillan. A.L (2015). Modernizing official statistics.
2. UNECE. (2018). Report of the United Nations Economic Commission for Europe High-level Group for the Modernisation of Official Statistics.
3. UNECE. (2016). Using Administrative and Secondary Sources for Official Statistics
4. UNSD, Esri. (2017) Research Exercise to establish a Federated Information System for the SDGs Communiqué.



Modernisation of statistical systems - Experiences of GCC-Stat



Sabir Al Harbi, Nancy McBeth
GCC-Stat, Muscat, Oman

Abstract

GCC-Stat - the Statistical Centre for the Cooperation Council for the Arab Countries of the Gulf has a key role in supporting member countries in both the modernization of their statistical systems and in providing the indicators necessary for the 2030 Agenda. The establishment of the centre in 2011 to provide a common official pool of statistics and data for member states; built on the strong history of statistical cooperation in the region. This paper will discuss the role of GCC-Stat in supporting modernization, with particular reference to the provision of SDGs indicators. The paper will describe how active cooperation between member states, and a focused regional statistical office, underpinned by common strategic frameworks and international best practices have been key to the statistical modernization agenda in the region. In reviewing some of the achievements, the paper will describe some of the lessons and broader implications.

Keywords

SDGs, Regional Cooperation, Statistical Strategy, Capacity Building, Data Revolution, Harmonized Projects

1. Introduction

The six countries United Arab Emirates, Kingdom of Bahrain, Kingdom of Saudi Arabia, Sultanate of Oman, State of Qatar and the State of Kuwait formed the Cooperation Council for the Arab Countries of the Gulf (GCC) in 1981 in order to achieve a high level of institutional coordination in economic, social, political, defence and security fields. The six members are all major producers of oil and gas, share many historical, economic and cultural links, and actively cooperate in many fields including statistics.

In 2015, GCC-Stat – the Statistical Centre for the Cooperation Council for the Arab Countries of the Gulf, together with the National Statistical Institutes in the Gulf region began to implement the 2015-2020 GCC regional statistical strategy. This strategy reflected a need to improve the range of available statistical information, at both the GCC and country level, and to ensure that statistics produced in the GCC were internationally comparable. The strategy also emphasises the need to improve efficiencies of statistical operations,

through modernization of statistical processes and methodologies. (GCC-Stat, 2015)

At the heart of the 2015-2020 strategy are a number of regional statistical projects, with the aim of producing harmonised regional statistics based on international best practices. Provision of statistical information in support of Sustainable Development Goals, is a key element.

This paper discusses the role that GCC-Stat has played in supporting the modernization of the GCC statistical system, with particular focus on provision of SDG indicators. Before discussing the status of SDG reporting in the region, the paper begins by discussing some of the drivers for statistical modernization in the GCC.

2. Drivers for Statistical Modernisation in the GCC

The first statistical offices in GCC countries were established in the early 1960's and produced a range of economic and social indicators, derived from surveys and a range of administrative records. This information was key for much of the planning of GCC countries throughout the next few decades.

As in many other countries, the demands for statistical information have increased significantly, in response to new local needs and international reporting requirements. A leading driver has been the increasing diversification of the economy, as countries move away from a reliance on Oil and Gas. At the same time, there is an increasing range of data and information available in the region, from both the public and private sector (McBeth, Al Harbi, Al Muzahmi, 2018),

While many of the National Statistical Institutions (NSIs) still conduct surveys, all are transforming. This includes using new data sources (including Big data), new ways of using existing sources (including administrative data) and transforming dissemination directions and systems. In many cases, these changes are rapid, consistent with the direction set out in the UN Data Revolution report. (United Nations, 2014)

In 2016, countries across the Arab world issued the Doha Declaration on a Data Revolution in the Arab World. This declaration recognised the great potential of the Data Revolution in the region, including opening up of new data sources, developing institutional and governance frameworks that provide open access to new data sources, and the importance of partnerships to transfer knowledge and share new data. (MDPS, 2016). Subsequently the First GCC Statistical Forum was held in Riyadh, Saudi Arabia in early 2017. This forum of producers, users and academics from across the GCC, emphasized the needed to strengthen cooperation and dialogue mechanisms between users and producers, explore public private partnerships in statistics to increase the frequency of data and develop new data sources, and to take full advantage of modern statistical tools and technologies (GASat, 2017).

3. Statistical Development Goals in the GCC

The measurement of development has moved from the sole focus on GDP, to indicators such as the Human Development Index (HDI), Millennium Development Goals (MDGs) and now SDGs. These tools have been important in benchmarking progress with other countries, given the rapid progress in GCC countries and NSIs have played a key role.

However, it is clear that SDGs are raising the data expectations bar in both the range of required indicators, and the range of disaggregations, to support the SDG spirit of “leave no one behind. (PARIS21, 2019). As PARIS21 note, many SDGs require new methodologies, new data to be collected or indeed new techniques to provide data at the required disaggregations.

Meeting these SDG requirements in the GCC, calls for a substantial effort in strategic leadership, collection and capacity building at both a country and regional level.

SDGs are recognised as a strategically important project at both the GCC and country levels. Within the GCC, the GCC Secretariat has established a regional SDG Task Force, comprised of representatives of planning agencies, relevant Secretariat personnel and GCC-Stat. The taskforce facilitates regional cooperation, through identifying common priorities for SDGs at the GCC level, mapping achievements against the national priorities for SDGs, exchanging experiences in the preparation of Voluntary National Reviews (VNRs), as well as reviewing the regional SDG report, prepared by GCC-Stat. In this way, the regional taskforce plays a key regional coordinating role in identifying regional priorities, as well as monitoring and understanding regional progress in meeting the SDG targets and goals. As a priority for the GCC statistical system, SDGs are the main focus of the GCC Statistical project – Development, Progress and Sustainability Indicators, and are increasingly shaping requirements in other projects. The GCC SDG indicators project focuses on producing regional level SDG indicators, such as periodical reports such as “Progress Report about the performance Towards the Achievement of Sustainable Development Goals – 2030 in the GCC Countries” (GCC-Stat, 2018). and agreeing on harmonised regional requirements. The project also provides the framework for regional capacity building in SDGs (discussed below).

Consistent with their coordination roles in their respective national statistical systems, the NSIs in the region also have a key role in the production, dissemination and analysis of SDG indicators at the country level. Most countries, including for example Bahrain (Kingdom of Bahrain, 2018) and UAE (FCSA) have high-level SDG committees, chaired at the Ministerial level. The NSI has a lead role in these committees, e.g. in the UAE, the Federal Statistics and Competitiveness Agency (FCSA) is the Deputy chair of the committee, as well as providing the Secretariat. NSIs have a key role in the production of the SDG

Voluntary National Reviews (VNR) presented to the UN High Level Political Forum. In 2018, UAE, Bahrain, KSA and Qatar presented their reviews. Oman and Kuwait will submit their reviews in 2019.

The Handbook for the Preparation of Voluntary National Reviews (United Nations, 2019) recommends that countries prepare a work plan for the preparation of the VNR. Some countries, such as Qatar have built on this and prepared a Road Map for all the activities associated with producing statistics and indicators (MDPS, 2017). Other local initiatives in GCC countries include National SDG Portals and National SDG Committees.

4. Regional Statistical Modernisation Initiatives to enhance availability of SDG indicators in GCC

Statistical Modernisation Initiatives in the GCC, therefore take place within the international, regional and local environment, including, but not exclusively, SDGs and the GCC Statistical strategy. This section of the paper highlights some of the regional initiatives led or facilitated by GCC-Stat; which through statistical modernization in the region, aim to enhance the availability of SDG indicators.

As noted above, the 2015-2020 GCC regional statistical strategy drives the overall direction of statistical activities, including modernization initiatives conducted by GCC-Stat. GCC-Stat supports member countries through a range of capacity development activities. These include tailored training to specific countries and joint regional workshops. Regional workshops specific to SDGs have included Population, Social and Environment indicators in March 2017, an upcoming workshop on Economic SDGs indicators in November 2019, as well as the regular meetings of the GCC Development Indicators Standing committee. In addition, GCC-Stat has provided targeted assistance to member countries in the form of field visits by its experts, as well as associated office based support by e-mail and telephone calls.

Monitoring and reporting on SDGs is still a major challenge in the region, due to the shortage of detailed data, including at the required disaggregations. While some aggregated data is available, many countries still struggle to provide the required disaggregations. These disaggregations often require additional data, including from more specialized surveys as well as from greater use of administrative data.

A number of the other GCC statistical priority projects, such as Environment Statistics, will increase the range of statistics. For example, the Environment Statistics projects will enhance the availability of indicators related to water and sanitation management, life under water, wild life reserves, etc. Regional and country specific capacity building initiatives for these projects therefore also contribute to increasing the range and quality of data for SDGs.

In support of the broader goals of producing internationally comparable statistics, GCC-Stat takes an active role in appropriate international and regional forums related to SDGs. For example, GCC-Stat has actively participated as an observer at almost all meetings of the Inter Agency Experts Group for SDGs (IAEG-SDGs). In the broader region, GCC-Stat has supported the Technical Committee for Population Indicators in SDGs formed by UNESCWA, UNFPA and the League of Arab States, and as noted above, is a member of the GCC SDG Committee. These forums are an opportunity to contribute to both the methodology of indicator development, but also help understand local needs.

The IAEG-SDG has achieved considerable progress in defining and describing the standardized indicators, that are at the heart of the global monitoring and reporting system. An important part of these standard definitions is the use of international classifications. Within the GCC strategy, this is supported by the “Statistical Standards, Classifications, Methodology and Data Quality” statistical priority project. The project aims to implement contemporary statistical standards, recommendations, international standard classifications and common methodologies across GCC statistics, with particular focus on the statistical priority projects set out in the strategy. (GCC-Stat, 2015). An important aspect is providing training in international standard classifications. Examples of recent training includes implementation of international and GCC classifications in administrative registers related to cultural statistics, and implementation of the GCC occupation classification in labour and related statistics. Recent training specific to member countries, has included the review of industrial classification and their implementation in Kuwaiti and Saudi registers, and the development of economic classification mapping in the Bahrain national accounts system.

This key element of statistical modernization therefore has considerable benefits to those SDG indicators that utilize data from the other statistical priority projects, as well as supporting the application of international standard calculation methodologies for SDGs.

As noted above, while SDG indicators will require a greater range of surveys, administrative records are a key source. GCC-Stat has established a GCC wide project to enhance the range of administrative data based indicators. The project has identified a standard set of potentially administrative based indicators to meet SDG and GCC policy requirements. The next stage is to work with each country to identify potential sources for the data elements required for each of the indicators.

The GCC-Stat work on modernization is not just confined to supporting countries to provide standardized data inputs or standardized indicators. GCC-Stat has established a project to deliver an Integrated End-to End solution. This project will provide the necessary IT solution (hardware and software) to

support automatic information exchange between GCC-Stat and GCC member countries, and provide an integrated data dissemination platform. This dissemination platform will strengthen data visualization and data presentation through for example an enhanced data portal. This IT solution, scheduled to be fully operational in mid-2020, will have a key role in enriching accessibility of GCC and country statistics, including SDGs.

In the meantime, GCC-Stat remains committed to the dissemination of as wide a range as possible of SDGs and other official statistics about the GCC. Consistent with that commitment, GCC-Stat has adopted the Open Data principles in the dissemination of GCC statistics.

5. Discussion

The GCC statistical system, comprising GCC-Stat, the six GCC countries and the user communities, is progressing on a modernization path. As the paper has shown, GCC-Stat plays a key role in many of the regional modernization initiatives. Some such as the Integrated End-to End solution are designed to provide a technology based solution to some of the issues associated with data management and dissemination. Other such as the adoption of Open Data principles and protocols, are designed to enhance the use and reuse of statistical data. Common to all are strategic frameworks which enable countries to work together to achieve improved efficiency and effectiveness. At the GCC level, the 2015-2020 Statistical Strategy provides the framework for not just improved efficiency, through modernization of statistical processes and methodologies, but also improvements in the range of available and internationally comparable statistics.

The GCC strategy has emphasized the importance of not just implementation of harmonised priority projects, but also enhancements in data availability and capacity building. The linkage of these three components is key. Greater supply of GCC level data requires more harmonised data. Capacity building is key to supporting the NSIs deliver on the harmonised projects.

While the GCC strategy pre-dates the international SDG framework, the strategy has served countries and GCC-Stat well in the initial implementation of SDGs. However, as the OECD have noted, building statistical capacity is a long-term process (OECD, 2017). In our experience, while the numbers of courses and attendees provide a useful metric, they do not reflect the consequential improvements in the range and quality of statistics, which take time to implement. For example, four countries have provided VNRs and two will submit their reviews in 2019.

Yet there remain gaps. The latest GCC report on SDG indicators only represented a range of indicators, and didn't relate to all the goals. This reflected the availability of data, in part due to local priorities. The need to prioritise reflects

both the need to ensure local relevance as well as consistency with the global standards, but also the broader budget constraints. In describing the potential of the Data Revolution to transform statistical offices in low and middle income countries, commentators have noted the large investments that will be needed in statistics to support both the monitoring and reporting on SDGs (e.g. MacFeely and Barnat 2017, OECD 2017, Paris21 2019). As MacFeely notes, estimates for additional support for the 77 lower-income countries range from \$1US billion to \$1.25US billion per annum; equating to between 13 -18 \$USD million per country per year just on SDG monitoring and reporting. GCC countries are higher income countries, yet they too must prioritise government spending. While detailed budgets are not available, it is unlikely that statistics budgets, even including the budgets in other ministries, will come close to the types of additional investment sought for SDG monitoring in lower-income countries.

An additional challenge is the need for local relevance, while meeting the international requirements for disaggregation noted earlier. While the multiple disaggregations requirements for international reporting meet many local requirements, they do not meet all local requirements. For example, the specific population distribution in the GCC, with high proportion of non-citizens, mean other detailed disaggregations are required. A related issue is that Goals, Targets and Indicators have all been prepared and published in English. As the Monitoring framework becomes available in other languages – e.g. Arabic – the language of the GCC; it is likely that there will be additional local requirements.

These factors means that there is a strong need for regional cooperation in the modernization of statistical systems in the region. The 2015-2020 GCC Statistical Strategy has provided the framework for this cooperation, including SDGs; and builds on the existing high level of institutional cooperation and cooperation.

6. Conclusion

This paper has discussed the role that GCC-Stat has played in supporting the modernization of the GCC statistical system, with particular focus on provision of SDG indicators. As the paper has shown, GCC-Stat facilitated modernization activities have included statistical methodological leadership in SDGs, as well as other statistical priority projects, which provide data for SDGs. As part of a broader project to enhance the production of administrative based indicators, GCC-Stat is leading work to identify administrative data that can be used to provide a greater range of indicators, including SDGs. In addition, GCC-Stat has recently initiated a project to provide an Integrated End to End solution for data management and dissemination.

The paper has highlighted the importance of regional, in this case South-South, cooperation, as a way of working together to modernize statistical

systems and enhance capability. While there remains much work to do, the range of initiatives completed under the GCC Statistical Strategy provides a solid base for enhancing the range of statistical information available to support the monitoring and reporting on SDGs.

References

1. FCSA, UAE National Committee on SDGs <http://fcsa.gov.ae/en-us/Pages/SDGs/The-National-Committee-on-SDGS.aspx>
2. GStat (2017) Riyadh Announces 14 Recommendations at the Conclusion of the First GCC Statistical Forum, <https://www.stats.gov.sa/en/news/173>
3. GCC-Stat (2015) GCC Statistical Strategic Planning and Road Maps for Development 2015-2020 <https://www.gccstat.org/en/center/about>
4. GCC-Stat (2018) Progress Report about the performance Towards the Achievement of Sustainable Development Goals – 2030 in the GCC Countries <https://www.gccstat.org/images/gccstat/docman/publications/209-development.pdf>
5. Kingdom of Bahrain, (2018), Voluntary National Review Report on the SDGS, 2018 <https://www.bahrain.bh/wps/wcm/connect/72881a32-957e-43eb-af3b-f245bff64ac9/SDGs+EN.PDF?MOD=AJPERES>
6. MacFeely. S, Barnat. N, (2017) "Statistical Capacity Building for Sustainable Development: Developing the fundamental pillars necessary for modern national statistical systems" in *Statistical Journal of Official Statistics*, vol 33, no. 4 <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji160331> MDPS (2016) https://www.mdps.gov.qa/ar/media1/events/Documents/ArabicStatistics16/Doha_Declaration_E.pdf
7. MDPS (2017) Modernisation of Official Statistics in Support of Sustainable Development Goals in Qatar: The Road Map https://www.mdps.gov.qa/en/media1/events/Documents/sdgworkshop/Road_map_on_statistics_for_the_SDGs16November2017EN.pdf
8. OECD (2017) "The role of national statistical systems in the data revolution", in *Development Cooperation Report 2017: Data for Development*, OECD Publishing, Paris, <https://doi.org/10.1787/dcr-2017-8-en>.
9. PARIS21 (2019) "Mobilising Data for the SDGs", Paris21 Discussion Paper, No 15, Paris. <http://paris21.org/paris21-discussion-and-strategy-papers>

10. United Nations (2014) "A World that Counts: Mobilising the Data Revolution for Sustainable Development"
<http://www.undatarevolution.org/report/>
11. United Nations (2019) "High Level Political Forum on Sustainable Development Handbook for the Preparation of Voluntary National Reviews, The 2019 Edition"
https://sustainabledevelopment.un.org/content/documents/20872VNR_hanbook_2019_Edition_v2.pdf



The influence of telematics device on driving behaviour of commercial vehicles across long and short haul drivers



Firdaus Abhar Ali, Mohd Azman Mohd Ismail, Vaijyenthi Gurdalay Singh,
Shafiq Naim Shahrudin
Atilze Digital Sdn Bhd, Petaling Jaya, Selangor, Malaysia

Abstract

This paper reviews the effect of Advanced Driver Assistance System (ADAS) on driver's behaviour who drive commercial vehicles. The study separates the drivers into two groups of drivers, the long haul group and short haul group. The long haul group is defined as a group of drivers that drives more than 5,000 km / month and short haul group is defined as a group of driver who drives less than 5000 km / month (CEPI, 2010). It is known that for the drivers that drives long distance, the exposure to negligent driving behaviour such as sudden braking and tendency of speeding are higher compared to the short distanced drivers due to reduced focus and fatigue. ADAS was used as an instrument to measure various data points that makes up the driving behaviour. In addition, ADAS also helps in improving the driving style by alerting the drivers whenever a dangerous driving behaviour is detected thus helping the driver to correct the behaviour. The effect of ADAS may have on the exposure to negligent driving of the drivers are studied further in this paper. The driver's behaviour is measured using the Malaysia Driver Score (MDS) which was published by MIROS. The established algorithm uses parameters derived from ADAS telematics device and it is an indication of the overall driving quality, where lower scores refers to higher chances in involvement with negligent driving and accidents. A total of 35 commercial vehicles with ADAS, with time shift-based drivers, participated in the study. The data from the ADAS was collected for the duration of 1-month and the difference of long-haul and short-haul group's score in MDS format are compared using independent T-Test to ascertain differences in the driving behaviour. The result proves that there is almost no difference in the MDS score between the long haul and short haul group in their driving behaviour. The usage of ADAS is known to help in reducing the road accident related risk levels of both of the groups being studied.

Keywords

Driving; ADAS; Safety; Self-Regulatory Practices; Driver's Score

1. Introduction

It is widely known that Advanced Driver Assistance System (ADAS) is proven to significantly reduce the number of road accidents in general. A recent study done by MIROS also establishes that driver behaviour of passenger vehicles fitted with ADAS shows greater improvement than those without the device, in terms of lower number of logged incidents. This paper however will focus on commercial vehicles particularly for two different groups; long-haul and short-haul drivers. There are arguments that long haul drivers contribute more to road accidents numbers compared to short haul drivers as discussed by Mishra, B., Sinha Mishra, N. D., Sukhla, S., & Sinha, A. (2010) in their paper. According to Mishra et. al (2010), road accidents are highly associated with the distance travelled. Similar argument was also discussed by P.Philip, J.Taillard, C.Guilleminault, Salva Quera, M.A., B.Bioulac, M.Ohayon (2019), where the finding draws a relation between long distance travel and sleep-related accidents.

Road accidents may result from human factors, environment and/or design of roads and vehicles factors. However, human factor often plays the greatest role in causing road accidents, especially those involving commercial vehicles (Abang Abdullah and Von, 2011). Human factor can be measured using the driver score model which will help to determine risk profile of drivers - whether the driver falls in the good score or a bad score band according to their driving behaviour. It encompasses three predictive driving behaviour parameters which contributes significantly to road crashes (based on MIROS's study). Research objective of this paper is to prove differences in driver score between long haul driver and short haul driver group. Scope of study for this analysis are three logistics companies based in Klang Valley which trips covers both long haul and short haul travels across Peninsular Malaysia.

The findings of this study will be significant in understanding the changes in driver behaviour & associated risk factors due to the usage of ADAS in commercial vehicle fleets. Relevant government authorities and regulators may look at the prospect of implementing driver score as a new alternative to measure driver's risk of all categories and types of vehicles.

2. Methodology

This chapter discusses the analysis method used to derive driver score for the two groups of drivers (long haul & short haul). For the purpose of this study, independent t-test was used to ascertain the differences in driver score means for the two groups of drivers. The source and background of the data set will also be discussed in this section followed by the driver score calculation.

This research study uses secondary data as a method of analysis. Data set was obtained for one month period for trips and events for each vehicle. This said data was obtained from RUPTELA's telematic platform which stores MOBILEYE's ADAS events data. In total there were 33 drivers who did trips for the whole month of February 2019 from three logistics companies. These drivers were categorized into two groups; long haul and short haul drivers.

Table 2.1
Description of Variables

Variable	Description
FCW	Forward Collision Warning
LDW	Lane Departure Warning
SPD	Speeding Warning
Distance	Distance in km

2.1 Driver Score

MDS model calculation was used in determining the driver's behaviour risk in this study. The calculation consists of two parts, one being the score weightage and another is the calculation using number of events recorded. The weightage percentage for the score was determined by road crash statistics provided by Malaysian police department for a period of 2 years, from 2013-2015.

Table 2.2 below shows the weightage percentage for each parameter which will be used to calculate the driver score.

Table 2.2
Score weightage for each event

Variable	Weightage
FCW	19%
LDW	30%
SPD	51%

Based on the weightage set for each type of violation, the score will be calculated following the below formula:

Score = $100 - (\text{FCW Penalty Score} + \text{LDW Penalty Score} + \text{SPD Penalty Score})$ (MIROS, 2018).

where,

FCW Penalty Score	=	FCW/Distance * 19, maximum score is 19
LDW Penalty Score	=	LDW/Distance * 30, maximum score is 30
SPD Penalty Score	=	SPD/Distance * 51, maximum score is 51

2.2 Independent Sample t-test

The independent sample t-test compares the means of two independent groups which in this study is long-haul and short-haul driver group in order to determine whether there is statistical evidence that the associated

population means are significantly different. Below are the hypotheses for independent sample t-test used in this study:

$$H_0 = \text{There is no different in mean driver score between long haul and short haul driver}$$

$$H_1 = \text{There is different in mean driver score between long haul and short haul driver}$$

In the sample data in this study, there are two variables: Driver category and Driver score. The variable Driver category has values of either "0" (Long haul driver) or "1" (Short haul driver), these will be the independent variables in this t-test. The variable Driver score is a numeric variable, and it will function as the dependent variable.

Table 2.3
Description of variables Driver and Driver Score

Variable	Description
Driver	"0" (Long haul driver) and "1" (Short haul driver)
Driver Score	Numeric value in percentage (%)

3. Result

The behaviour of drivers is explained in the table below using simple descriptive statistics.

Table 3.1
Overall descriptive statistics for two groups

Event	Number of Event Triggered
LDW	87709
FCW	6601
SPD	4689

Table 3.1 above shows that the event triggered by the two categories of drivers. Lane Departure Warning (LDW) was the highest event triggered with 89% of the all event triggered, followed by Forward Collision Warning (FCW) with 7% and the lowest event triggered was Speed (SPD) which accounted for 5%.

Table 3.2
Overall descriptive statistics for two groups

	Variable	Overall	Long haul	Short haul
Distance	Average	6240.44 km	9968.15 km	3134.01 km
	Max	16347.56 km	16347.56 km	5034.46 km
	Min	17.93 km	5836.85 km	17.93 km
Driver Score	Average	82.22%	85.17%	79.76%
	Max	99.09%	99.09%	93.51%
	Min	18.12%	64.8%	18.12%

Table 3.2 shows that the average distance traveled for those two groups of driver was 6240.44 km. There was a noticeable gap in between the minimum distance travelled accounting for only 17.93 km while the maximum distance travelled was a whopping 16347.56 km. Driving score average did not vary much as the differences between both the groups was only 5.41 %. But in contrast to a study by (Mishra, B., Sinha Mishra, N. D., Sukhla, S., & Sinha, A., 2010) which stated long distance travel were found to be associated with high percentage of Road Traffic Accident (RTA), driving score for short haul group attributed to the lowest score with only 18.12% which translates into risky driving.

To explain and support the argument for the differences in driving score between long and short haul drivers, independent t-test was used. Independent t-test was analyzed using r software and below is the result:

$$\begin{aligned} & \text{Two Sample t-test} \\ & t = 1.0412, df = 22.861, p\text{-value} = 0.3087 \end{aligned}$$

From the analysis, we can see that the independent sample t-test analysis showed that p-value is 0.3087 and we accept H null. Hence, we can conclude that there is no difference in driver score mean between long haul and short haul drivers. This shows regardless of short or long trips, drivers risk factors has been brought to a similar level due to the usage of ADAS. It is assumed that short-haul group risk factors remain lower than the long-haul group.

4. Discussion and Conclusion

In this study we investigated the usage of ADAS in commercial vehicle that travels short & long haul. The purpose of the study is to observe, and validate the assumptions that long distance driver has much higher tendency in dangerous driving behaviour. In addition, the effect of ADAS in improving the driving behaviour on those drivers, have been observed.

The Malaysia Driver Score (MDS) is a scoring model to assess and score a driver's driving characteristics through the use of In-Vehicle Telematics Device (IVTD), which is in our case, ADAS was used. MDS can be an effective tool to cultivate, nurture and promote safe driving behaviour among Malaysian drivers by providing objective and real-time driving behaviour assessment. Drivers can obtain their overall driving score and improve their driving by examining the risky driving behaviours as identified by MDS. By improving the overall driving behaviour, it is expected that the number of crashes, especially those due to driver errors can be further reduced.

From the study, it is seen that there are no noticeable difference of MDS score between long haul and short haul group. ADAS is known to help to reduce the risk of the two haul group in dangerous driving, thus reducing the

chance of negligent driving. The ADAS alarm & notification system has helped to notify the driver when dangerous or negligent driving behaviour is observed, and the drivers can use it to retroactively correct their driving style within the trip. This finding shows that the long-haul driving group, especially, has significant benefit in using ADAS as it lowers their risk factors to the same level as short haul drivers.

References

1. Austin, P. C., Steyerberg, E. W. (2015). The number of Subject per Variable Require in Linear Regression Analyses. *Journal of Clinical Epidemiology*, 68(6), 627-638. Bihani, A. (2014). A new Approach to Monte Carlo Simulation of Operations. *Laser*, 20, 0-20 .
2. Abang.Abdullah., and H.L. (2011). Factors of Fatigue and Bus Accident. 14, 317 - 321.
3. The need for flexible and optimal solutions on European roads (pp. 2-4, Rep.). (2010). Brussels, Belgium: Confederation of European Paper Industries (CEPI).
4. Malaysian Institute of Road Safety Research (MIROS). (2018). Development of Malaysia Driver Score - MIROS Pilot Study (pp. 1-24, Rep.).
5. Masayoshi Tanishita & Bert van We, (2017). Impact of vehicle speeds and changes in mean speeds on per vehicle-kilometer traffic accident rates in Japan.
6. P.Philip, J.Taillard, C.Guilleminault, Salva Quera, M.A., B.Bioulac, M.Ohayon (2019). Long Distance Driving and Self-Induced Sleep Deprivation among Automobile Drivers.
7. Mishra, B., Sinha Mishra, N. D., Sukhla, S., & Sinha, A. (2010). Epidemiological study of road traffic accident cases from Western Nepal. *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine* , 35 (1), 115–121. <https://doi:10.4103/0970-0218.62568>
8. Williamson, A., Bohle, P., Quinlan, M., & Kennedy, D. (2009). Short Trips and Long Days: Safety and Health in Short-Haul Trucking. *ILR Review*, 62(3 , 415–429. <https://doi.org/10.1177/001979390906200309>



Inspecting ecological communities structure via FDA



Tonio Di Battista¹, Francesca Fortuna¹, Fabrizio Maturo²

¹University of Chieti-Pescara, Italy

²National University of Ireland, Galway

Abstract

Despite there is a general recognition that species diversity indicates the status of the ecosystem or ecological communities, and thus the quality of the living environment, no consensus measure exists because biodiversity is a very complex concept that is intrinsically multidimensional and multivariate. For this reason, both at the academic and institutional level, there is a lively discussion about how to properly measure and monitor biodiversity. Hence, this study proposes a new methodological approach for inspecting ecological communities 'structure via functional data analysis; specifically, a functional approach to diversity profiles and some related functional tools is suggested. This research underlines and shows how each functional tool may highlight specific aspects of community's structure and, in addition, an inferential approach is proposed for building confidence intervals on functional instruments' mean. The goal of this research is to provide ecologists, policymakers, and scholars with additional methods for detecting ecological communities characteristics. Effectively, the combined use of these instruments provides a useful method for identifying areas of high environmental risk.

Keywords

Diversity profile; FDA; biomonitoring; convex functions; bootstrap confidence bands

1. Introduction

Biodiversity plays a crucial role in environmental monitoring because it is one of the most important indicator of environmental health (Burger et al., 2013) as it decreases in relation to ecosystem stressors.

Biodiversity is a multidimensional concept accounting for both species richness, that is the number of different species in a community; and species evenness, that is the relative abundance of each species in an area. The main issue for statisticians is to provide a suitable measure which takes into account the multivariate nature of biodiversity. Indeed, despite the enormous number of indices which have been developed to assess biodiversity (see Gove et al. (1994) for a wide review on this topic), a universally accepted measure has not been established (Di Battista et al., 2016, 2017). A possible solution to this

issue is given by diversity profiles. They are non-negative and convex curves, which express diversity as a function of the relative abundance vector. They provide a comfortable representation of diversity because they consider its multivariate nature; return a graphical representation of diversity; and allow to compare different communities when the profiles do not intersect. Thus, the behaviour of these curves gives important information about biodiversity in a community.

In the literature, different diversity profiles have been proposed; the main ones are the β diversity profile (Patil and Taillie, 1979), the intrinsic diversity profile (Patil and Taillie, 1979), and the Hill's number Hill, 1973), among the others. The formulation of a diversity profile can be generalized as follows:

$$\{\Delta_x: x \in R\} \quad (1)$$

where Δ_x are various diversity measures obtained by varying x in the domain R , which can be finite or infinite. The curve which joins the (x, Δ_x) pairs for $x \in R$ is termed a diversity profile, and depicts in a single picture simultaneous values of diversity measures with varying sensitivities to the rare and abundant species as a function of the parameter x . Hence, a diversity profile measures diversity through a curve rather than a scalar as in the case of diversity indexes. This emphasizes the importance of using such an approach in environmental studies as it does not collapse the information of a multidimensional set (the biological community) into a single number (Gattone and Di Battista, 2009).

Due to these characteristics, Gattone and Di Battista (2009) proposed to analyze them through the functional data analysis (FDA) approach (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). The latter allows to obtain several advantages in an ecological context. Indeed, we can analyze the shape of the profile through functional tools (Di Battista et al., 2016, Maturo e Di Battista, 2018) and evaluate the behaviour of the profile throughout the reference domain. The functional approach is particularly helpful when an inferential approach for biodiversity is required. Indeed, making inference on diversity profiles starting from the abundance vector with standard multivariate procedures, involves many unresolved issues. The solutions proposed in the literature mainly concern the use of independent replications of a sampling design (Barabesi and Fattorini, 1998) and the use of jackknife to build confidence intervals for the diversity estimator (Fattorini and Marcheselli, 1999). However, in practice, replications of paths for a given sampling design could be quite expensive and time consuming (Di Battista and Gattone, 2004), and the jackknife requires that the elements of the frequencies vector are all different from each other. Moreover, in some cases, the jackknife procedure may fail to return a convex diversity profile. On the other hand, the FDA approach analyses the profile as a function; thus, for each sample unit, a single observation is observed, overcoming problems of simultaneous multivariate inference (Di Battista and Fortuna, 2017).

Starting from these considerations, simultaneous confidence bands for the mean diversity profile function are obtained using a bootstrap procedure. Since diversity profiles are constrained functions, our proposal consists in pre-processing diversity profiles via a differential equation method (Ramsay, 1998) and bootstrapping the unconstrained functions. This allows us to respect the characteristics of the diversity profile and to work in an appropriate functional space, that is the L^2 space.

The paper is organized as follows: Section 2 provides a brief review of the diversity profile evaluation in a functional context focusing on the estimation of constrained curves. The section continues with the construction of bootstrap simultaneous confidence bands for the functional mean estimator. Section 4 deals with an application to a real dataset concerning fish biodiversity in Lazio rivers (Italy) and Section 5 concludes the paper.

2. Methodology: FDA approach for evaluating diversity profiles

Diversity profile are functions of the abundance vector, whose knowledge requires a census of the population under study, which is unfeasible in most cases (Barabesi and Fattorini, 1998).

Let us suppose that an ecological population is composed of N units and is partitioned into s species $j=1,2,\dots,s$. Let $\mathbf{N}=(N_1,\dots,N_s)^T$ be the species abundance vector whose generic element N_j represents the number of individuals belonging to the j -th species, and let $\mathbf{p}=(p_1,\dots,p_s)^T$ be the relative abundance vector with $p_j=N_j/\sum_{j=1}^s N_j$ such that $0\leq p_j\leq 1$ and $\sum_{j=1}^s p_j=1$. The abundances must be estimated by means of a sample survey, following a model-based or a design-based approach. The latter is widely applied in an ecological context, because it considers the values of a variable of interest as fixed quantities and the selection probabilities, introduced with the design, are used in defining the properties of the estimators, without making any assumptions about the population (Thompson, 1992).

Let us suppose that abundance data have been collected from a biological community. Then, for each i -th sample unit (habitat, environmental site, etc.), $i=1,2,\dots,n$, a diversity profile Δ_{ix} can be obtained. Since diversity profiles are presented as curves, they may be represented in a functional framework as follows (Gattone and Di Battista, 2009):

$$\Delta_{ix}(x) = f_i(x) + e_i(x) \quad x \in R, \quad i = 1, 2, \dots, n \quad (2)$$

where $f_i(x)$ is an arbitrary smooth function; and $e_i(x)$ denotes an unknown independent zero-mean error term. Usually, the functional form of the diversity profile, $f_i(x)$, can be reconstructed from the observed raw sampled data points $\{\Delta_x: x \in R\}$ using basis function expansion and smoothing (Ramsay and Silverman, 2005). However, diversity profiles are a special case of functional data in that they are non-negative, convex and decreasing curves.

Thus, some restrictions should be placed on $f_i(x)$ to ensure the characteristics of diversity profiles. In the following, we adopt $f(x)$ to indicate the constrained functional approximation of a generic diversity profile. It may be represented as follows (Ramsay, 1998):

$$f(x) = \beta_0 + \beta_1 D^{-(m-1)} \exp(\int w(u)du) \tag{3}$$

where β_0 and β_1 are arbitrary constants, D^{-1} means taking the indefinite integral, and $w(x)$ is a Lebesgue square integrable function. Specifically, $f(x)$ is the solution of the differential equation $D^m f(x) = w(x)D^{m-1}f(x)$. It is straightforward to verify that for $m = 1$ the solution is a positive function, for $m = 2$ the solution is a monotone function and for $m = 3$ the solution is a convex function. Since the function $w(x)$ is unconstrained, it can be expanded in terms of K known basis functions as follows:

$$w(x) = \sum_{k=1}^K c_{ki} \phi_k(x) \tag{4}$$

where c_{ki} is the k -th coefficient which defines the linear combination and $\phi_k(x)$ is the k -th basis function. The main result of the transformation in Eq. (3) is that we can overcome the problem of finding the constrained functional $f(x)$ by estimating the unconstrained function $w(x)$ that, in what follows, we shall term the unconstrained diversity profile. We point out that the function $w(x)$ has the same information on diversity of $f(x)$ and, whenever it is required, we can invert the transformation and go back to the constrained diversity profile functions by putting the coefficients $w(x)$ in Eq. (3).

Once the functional approximation of the profile has been obtained, different functional tools can be computed to evaluate the biodiversity of a community (Di Battista et al, 2016). Then, the assessment of uncertainty of an obtained estimator is a fundamental step in all statistical analysis. In function estimation problems, simultaneous confidence bands provide a unified set of graphical and analytical tools to harness such tasks as data exploration, model specification or validation, assessment of variability in estimation, prediction, and inference (Di Battista and Fortuna, 2017). In FDA and particularly in an ecological context, the bootstrap methodology turn out to be often the only practical alternative to derive the sampling distribution of a functional statistic (Cuevas et al., 2006). Effectively, the normality assumption could be too strong for diversity profiles. For this reason, we extend the bootstrap methodology to the framework of biodiversity assessment to build confidence intervals for the mean diversity profile.

Let $w_1(x), w_2(x), \dots, w_n(x)$ be the sample of unconstrained functions derived from the original diversity profile observed in the i -th sites, $i=1,2,\dots,n$, and $T = T(w_1(x), \dots, w_n(x))$ be the sample statistics under consideration. The bootstrap resamples $w_1^*(x), \dots, w_n^*(x)$ are calculated following the following scheme (Febrero-Bande and de la Fuente, 2012):

$$w_i^*(x) = w_i(x) + \epsilon(x) \tag{5}$$

where $\epsilon(x)$ is normally distributed with mean 0 and covariance matrix $\lambda \Sigma_x$, λ is the smoothing parameter and Σ_x is the covariance matrix of $w_1(x), \dots, w_n(x)$. We indicate with $T^{*b} = T(w_1^{*b}(x), \dots, w_n^{*b}(x))$ the bootstrap sample statistic of the generic b -th replication, with $b=1, \dots, B$. The ordering of the bootstrap statistic is obtained by introducing a suitable distance, $d(T(x), T^{*b}(x))$, using e.g.:

- L^2 -metric: $\left(\int (T(x) - T^{*b})^2 dx \right)^{1/2}$
- The supremum L^∞ -metric: $\|T(x) - T^{*b}(x)\| = \sup_x |T(x) - T^{*b}(x)|$

Thus, it is easy to define the $(1 - \alpha)\%$ bootstrap confidence band such that $d(T(x), T^{*b}(x)) \leq d_\alpha$ with d_α the $(1 - \alpha)\%$ quantile of the distances between the bootstrap re-samples and the sample estimate. As shown by Cuevas et al. (2006), the performances of the bootstrap approximations via extensive simulations and real data applications provides asymptotic validity results without the strong assumption of normality.

3. Result: Application to the Rivers of Lazio

As an example, the method illustrated in Section 2 is applied to a real dataset concerning fish biodiversity of Lazio's current waters. The data set is available at the website <http://dati.lazio.it/catalog/it/dataset/bioittica> and consists of fish abundances of 54 species detected in 33 rivers of Lazio (Fig. 1) in 2015.

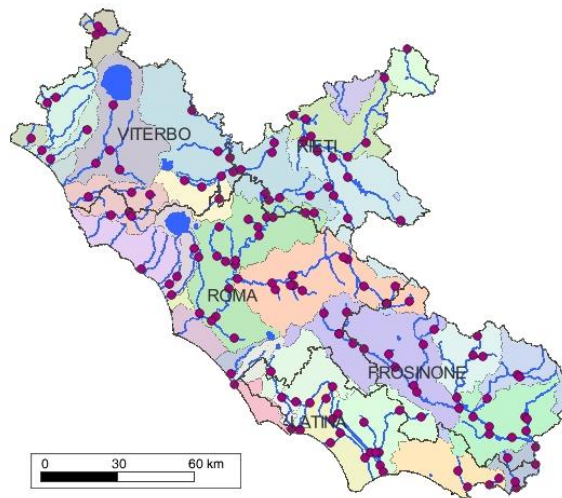


Figure 1: Map of Lazio's watercourses (<http://www.arpalazio.gov.it/>)

To evaluate fish biodiversity in this area, a diversity profile approach is adopted, choosing the Hill's number (Hill, 1973):

$$H_q = \left(\sum_{j=1}^s p_j^q \right)^{1/(1-q)} \tag{6}$$

Then, the functional approximation of the profiles is computed following Eq. (3) and it is showed in Fig. 2.

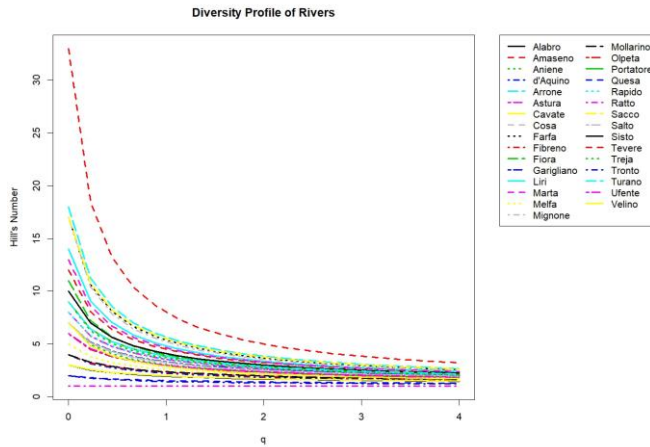


Figure 2: Functional diversity Hill's numbers of Lazio's rivers

It is evident that the "Tevere" river is the most diverse because it is above all other profiles and does not intersect with other curves. On the contrary, the less diverse river is the "Ratto", with only one species, followed by the "Quesa" and "Tronto" with two species. Regarding the other rivers, most of them cannot be ranked because many profiles intersect. To provide a summary measure of the biodiversity in the considered region, the mean diversity profile (solid black line in Figure 3) is computed. Figure 3 shows that, on average, the stations present 10 different species. The same figure shows the bootstrap simultaneous confidence bands for the mean estimator of level $1-\alpha=95\%$ (dotted red lines in Fig. 3). We note that, in the first part of the domain, there is greater variability; this is due to the characteristics of the profile. Indeed, in the second part of the domain, the profile tends to be very flat and the sample curves become constant (Di Battista et al., 2016), hence the sample variance tends to zero.

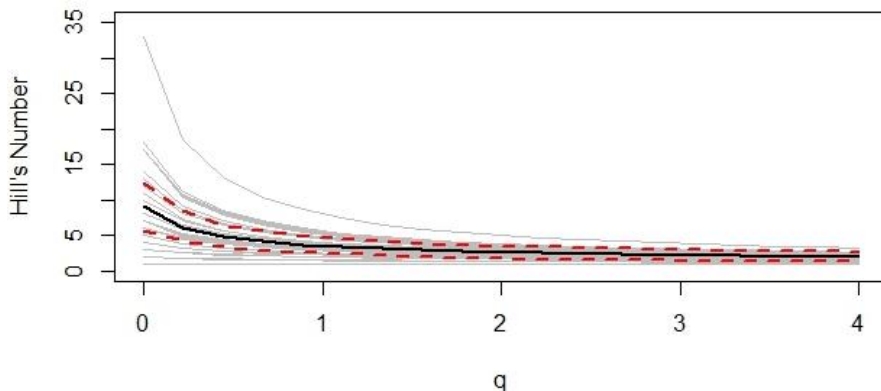


Figure 3: Functional diversity Hill's numbers of Lazio's rivers (solid grey lines) with their functional mean (solid black line) and bootstrap confidence bands (dotted red lines) at level $1-\alpha=95\%$

4. Discussion and Conclusion

This paper focuses on the multivariate nature of biodiversity and aims to provide a new methodology for overcoming the issues of the classical indicators in a functional framework. Specifically, we have proposed a functional approach to diversity profiles taking into account the constrained nature of these data. Then, an inferential approach to diversity profile mean estimator is considered. We emphasize the usefulness of the FDA approach to overcome some drawbacks typical of an inferential approach for the diversity profile based on the abundance vector. The main advantage derives from the fact that the profile is a function, that is, a single variable observed on a sample unit, rather than a multivariate vector. Moreover, FDA consents an in-deep evaluation of the profile curves behaviour through the reference domain, showing different aspects of diversity as the emphasis shifts from rare to common categories. The final goal of this research is to provide Ecologists, policymakers, and scholars with additional tools for evaluate biodiversity and detect areas with high environmental risk.

References

1. Barabesi, L. and Fattorini, L. Design-Based Approaches for Inference on Diversity, 189-195. Dordrecht: Springer Netherlands, 1998
2. J. Burger, M. Gochfeld, C. Powers, J. Clarke, K. Brown, D. Kosson, L. Niles, A. Dey, C. Jeitner, and T. Pitt_eld. Determining environmental impacts for sensitive species: Using iconic species as bioindicators for management and policy. *Journal of Environmental Protection*, 4:87-95, 2013.
3. A. Cuevas, M. Febrero-Bande, R. Fraiman. On the use of bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis* 51(2):1063-1074, 2006
4. Di Battista, T. and Fortuna, F. Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). *Statistical Analysis and Data Mining: The ASA Data Sci Journal*, 10, 21-28, 2017.
5. T. Di Battista, F. Fortuna, and F. Maturo. Environmental monitoring through functional biodiversity tools. *Ecological Indicators*, 60:237-247, 2016.
6. T. Di Battista, S.A. Gattone. Multivariate bootstrap confidence regions for abundance vector using data depth. *Environmental and Ecological Statistics*, 11, 355-365, 2004.
7. Fattorini, L. and Marcheselli, M. Inference on intrinsic diversity profiles of biological populations. *Environmetrics*, 10, 589-599, 1999.
8. M. Febrero-Bande and M. de la Fuente. Statistical computing in functional data analysis: The r package *fda.usc*. *Journal of Statistical Software, Articles*, 51(4):1-28, 2012.

9. F. Ferraty and P. Vieu. Nonparametric functional data analysis. Springer, New York, 2006.
10. S.A. Gattone and T. Di Battista. A functional approach to diversity profiles. *Journal of the Royal Statistical Society*, 58:267-284, 2009.
11. J. Gove, G. Patil, D. Swindel, and C. Taillie. Ecological diversity and forest management. In G. Patil and C. Rao, editors, *Handbook of Statistics*, vol.12, Environmental Statistics, pages 409-462. Elsevier, Amsterdam, 1994.
12. M. Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54:427-432, 1973.
13. Maturo, F. and Di Battista, T. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. *Ecological Indicators*, 84, 70-81, 2018.
14. G. Patil and C. Taillie. An overview of diversity. In J. Grassle, G. Patil, W. Smith, and C. Taillie, editors, *Ecological Diversity in Theory and Practice*, pages 23-48. International Co-operative Publishing House, Fairland, MD, 1979.
15. J. Ramsay. Estimating smooth monotone functions. *Journal of the Royal Statistical Society*, 60:365-375, 1998.
16. J. Ramsay and B. Silverman. *Functional Data Analysis*, 2nd edn. Springer, New York, 2005.
17. Thompson, S. (1992) *Sampling*. New York: John Wiley & Sons.



Recent advances in ecological networks: Regularized grouped Dirichlet-multinomial regression



R. Ayesha Ali¹, C. Crea²

¹Department of Mathematics and Statistics, University of Guelph, Guelph, Canada.

²Geosyntec Consultants, Guelph, Canada.

Abstract

Having a solid understanding of why species interact in ecological settings can help policymakers in the development of conservation or restoration management plans. Quantitative ecologists are now able to collect more detailed species traits, either from the field or through published literature, that can supplement observed counts of species interactions in ecological networks. While there is great interest in using such data to further enhance our understanding of the mechanisms driving the observed interactions, there is a lack of statistical models that can accommodate the complexities presented by such data. Here we discuss some of these modelling challenges and present new advances in the context of plant-pollinator networks. In particular, we present regularization for grouped Dirichlet-multinomial regression.

Keywords

Grouped Dirichlet multinomial regression; interaction networks; ecological modelling; adaptive lasso

1. Introduction

Mutualistic networks arise out of several ecological processes including pollination, seed dispersal, and host-parasite relationships, to name a few. Studies on these networks have revealed common structural patterns, such as the nested organization of pairwise interactions and the skewed distribution of links per species (Montoya et al. 2006; Bascompte and Jordano 2007; Vázquez et al. 2009a). It is believed that these structural patterns are driven by both evolutionary and ecological processes. Neutrality states that species' interactions are totally random whereas linkage rule theory suggests that the functional traits between species must match for species to interact with each other.

For example, an insect's tongue length must be long enough to reach a plant's reproductive parts for pollination to take place. The characteristics pollinators seek in plant species may be better anticipated if species interactions are modelled by the functional traits that drive them. Grouped Dirichlet-multinomial (DM) regression provides a framework to quantify the contributions of species traits and/or linkage rules to pollination.

Dirichlet-multinomial regression (Guimarães and Lindrooth, 2007) is a multinomial logistic regression model developed in econometrics that accounts for over-dispersion. Under a consumer-behaviour framework, we assume that a pollinator faced with a number of choices (plant species in the network) assigns a level of utility to each plant species and then selects the one with the maximum utility (McFadden, 1974; Guimarães and Lindrooth, 2007). These utilities may be modelled as a function of plant species attributes/traits, pollinator species characteristics/traits, or as interactions between the two (e.g. linkage rules). Therefore, the probability of a specific plant-pollinator pair interacting can also be modelled as a function of these traits. Since modelling is done at the species level, this model corresponds to a grouped DM regression (Mosimann, 1962).

Dispersion may be constant for all pollinator species or may be a function of pollinator-specific covariates such that dispersion varies across pollinators. Estimates of the regression coefficients quantify the relative contribution of each trait/linkage rule to the interaction probabilities, while the dispersion parameter accounts for heterogeneity in the data that is not explained by the covariates.

However, it is only within the past few years that field ecologists have collected detailed species data to better understand the mechanisms that drive pollination. Investigators may be faced with a large number of potential covariates to go into the model. Variable selection techniques can reduce the size and complexity of the model, paring down the features to a set of ones predictive of plant-pollinator interaction while both avoiding overfit and increasing interpretability. Here we present a new regularized grouped DM regression model using standard and adaptive lasso methods. Tuning parameters are selected using an information criterion while optimization is achieved via the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle, 2009). All the proposed methods are evaluated via simulated and empirical data sets and all implementations of the standard and regularized grouped DM regression model are publicly available as routines in R (R Core Team, 2017).

Section 2 provides a brief description of the DM parameterizations for both the unpenalized and penalized grouped DM regression model. Section 3 investigates the performance of lasso-type methods for the regularized DM regression model via simulation and Section 4 shows the results of the analysis of an empirical plant-pollinator network using our proposed lasso approach. Discussion and conclusions are provided in Section 5.

2. Methodology

DM regression for grouped data assumes that individuals within a group share common characteristics and are faced with the same choice set; hence,

the true level of variation is at the group level (Guimarães and Lindrooth, 2007). Let X be a $G \times J \times K$ design array containing covariate information, with entries x_{gjk} . We use \mathbf{x}_{gj} to represent the K -length vector of covariates associated with group g and choice j . The multinomial probabilities \mathbf{P}_g are modelled as a function of covariates through a logit formulation. Under the random utility framework (McFadden, 1974) with utility function given by the right hand side of (1) plus independent errors that follow an extreme value distribution, the logit formulation follows directly.

$$\log\left(\frac{p_{gj}}{1-p_{gj}}\right) = \boldsymbol{\beta}^T \mathbf{x}_{gj} + \eta_{gj}, \tag{1}$$

for $g = 1, \dots, G$ and $j = 1, \dots, J$, where $\boldsymbol{\beta}^T$ is a K -length vector of unknown regression coefficients associated with the covariates in \mathbf{x}_{gj} and η_{gj} is a random group effect that accounts for unobservable heterogeneity among individuals within a group.

We further make the assumption that the $e^{\eta_{gj}}$'s follow independent gamma distributions with both shape and scale parameters $\delta_g \lambda_{gj}$, $\delta_g > 0$. Under these assumptions, it can be shown that the probabilities \mathbf{p}_q follow a Dirichlet distribution with parameters $\alpha_g = (\delta_g^{-1} \lambda_{g1}, \dots, \delta_g^{-1} \lambda_{gJ})$, where δ_g is used to quantify the overdispersion. Maximization of the log-likelihood provides estimates of $\boldsymbol{\beta}$ and δ_g . In the special case that $\eta_{gj} = 0$ for all g and j , there is no random group effect and the DM model reduces to a group conditional logit model. If, further, there are only group-specific covariates, then the model reduces to a standard multinomial logit model.

Regularized DM regression

Consider the constant dispersion model in which $\delta_g = \delta \forall g$. To perform variable selection, we add a penalty term to the log-likelihood and proceed by minimizing the penalized (negative) log-likelihood function:

$$\ell_p(\boldsymbol{\beta}^*, \tilde{\lambda}) = -\ell(\boldsymbol{\beta}^*) + \tilde{\lambda} \mathcal{J}(\boldsymbol{\beta}^*),$$

where $\boldsymbol{\beta}^*$ is an M -length parameter vector for which the first $K = M - 1$ elements contain the regression coefficients while the last element is the dispersion parameter δ , $-\ell(\boldsymbol{\beta}^*)$ is the negative DM likelihood, $\mathcal{J}(\boldsymbol{\beta}^*)$ is the penalty term, and $\tilde{\lambda} > 0$ is the tuning parameter. Since our main interest is in a sparse solution where some elements of $\boldsymbol{\beta}^*$ are shrunk exactly to zero, we let $\mathcal{J}(\boldsymbol{\beta}^*)$ be the L1-norm, which is the standard lasso (Tibshirani, 1996). Within a lasso framework, the tuning parameter $\tilde{\lambda}$ determines the strength of the regularization such that smaller values of $\tilde{\lambda}$ correspond to less shrinkage and larger values of $\tilde{\lambda}$ lead to sparser solutions.

The penalized likelihood for the DM lasso becomes:

$$\ell_p(\beta^*, \tilde{\lambda}) = -\ell(\beta^*) + \tilde{\lambda} \sum_{m=1}^K |\beta_m^*|, \quad (2)$$

where $\sum_{m=1}^K |\beta_m^*|$ is the L1-norm. Note that although we estimate the dispersion δ , we do not penalize for this parameter (which is estimated in the M^{th} element of β^*).

We also implemented the adaptive lasso (Zou, 2006), which scales the L1-norm term by an adaptive data-driven weight vector, $\hat{w}_m = |\hat{\beta}^{ML}|^{-\tilde{\gamma}}$, where $\tilde{\gamma} > 0$ is a tuning parameter that adjusts the weights. The adaptive lasso penalizes irrelevant predictors more than relevant predictors, thereby leading to consistent model selection and optimal prediction. Minimization of (2) finds the penalized MLEs with certain parameters shrunk to exactly zero thus achieving variable selection and parameter estimation simultaneously. For minimization, we implemented FISTA (Beck and Teboulle, 2009) with the ADADELTA (Zeiler, 2013) learning rate (i.e., stepsize) method. The optimal value of $\tilde{\lambda}$ was tuned via BIC over an equally-spaced log grid of 100 $\tilde{\lambda}$ values, starting from $\tilde{\lambda}_{max} = \max_M \frac{|S_M|}{\hat{w}_m}$ to $\tilde{\lambda}_{min} = 0.01$, where S_M is the gradient vector evaluated at $\beta_m^* = 0$, $m \neq M$. For the adaptive lasso, we let $\tilde{\gamma} = 1, 2, 3$ before running FISTA to find the optimal $\tilde{\lambda}$. R code to fit these models is available on GitHub (Crea, 2016).

3. Simulation Study

Here we consider two DM regression dispersion structures: none ($\delta_g = 0$) and constant ($\delta_g = 6$). Networks were generated in R, per Crea et al. (2016), based on a gamma-Poisson parameterization of the DM model. We considered three network sizes: small (25×10), medium (50×20) and large (100×30). Entries of the covariate array X were generated based on complementarity linkage rules following Santamaría and Rodríguez-Gironés (2007). See Crea et al. (2016) for more details. We generated $K=20$ covariates in total, of which only the first 4 were relevant to pollination. We let $\beta^* = (-0.5, 1, -1, 2, 0, \dots, 0)$ and evaluated the performance of the regularized group DM model in the fixed parameter dimension. Results were averaged over 100 replicates.

To evaluate variable selection, we calculated the percent of models among the 100 replicates that selected the true model, an underfit model and an overfit model, respectively. We also calculate the average number of zero coefficients correctly estimated to be nonzero (i.e., true positives) and the average number of nonzero coefficients incorrectly estimated to be zero (i.e., false negatives) over the 100 replicates. Finally, we calculated the average mean squared error to assess parameter estimation.

Table 1 shows the results from fitting a regularized regression model to the simulated data. When data were generated with no dispersion, our

method never returned an underfit model or a model with false negatives, and all MMSE values were relatively low. For a given network size, the adaptive lasso with $\tilde{\gamma} = 3$ had the highest percentage of correct model fits, the highest true positives, and the lowest MMSE. Performance improved as network size increased. The standard lasso always returned an overfit model whereas the adaptive lasso consistently returned better fit models as $\tilde{\gamma}$ increased. Similar trends were observed when data were generated with constant dispersion, but with decreased performance.

4. Analysis of Terceira Island Network

Plant-pollinator interactions across fifty 10m×1m transects were surveyed from June to September in 2013 and 2014. Sampling protocols are described in Picanço et al. (2017). The network consists of $G=54$ insect species, $J=48$ plant species and a total of 2,134 observed interactions (flower visits). There were 9 unidentified insect species that were removed from the network because no trait information was available, leaving a total of 2,018 observed flower visits for analysis.

Table 1: Results on model consistency and average mean square error (MMSE) of adaptive lasso when 4 out of $K=20$ covariates are relevant, based on 100 replicates per network size. $\tilde{\gamma} = 0$ is equivalent to the lasso.*

Model	Size	$\tilde{\gamma}$	Under	Correct	Over	TP	FN	MMSE
No Dispersion	Small	0	0	0	100	1.52	0	0.78
		1	0	14	86	13.54	0	0.33
		2	0	57	43	15.27	0	0.26
		3	0	75	25	15.62	0	0.24
	Medium	0	0	0	100	0.47	0	0.15
		1	0	17	83	13.81	0	0.06
		2	0	84	16	15.78	0	0.04
		3	0	93	7	15.92	0	0.04
	Large	0	0	0	100	1.66	0	0.05
		1	0	22	78	14.44	0	0.02
		2	0	87	13	15.86	0	0.01
		3	0	96	4	15.96	0	0.01
Constant Dispersion ($\delta = 6$)	Small	0	0	0	100	3.39	0	4.35
		1	0	4	96	11.89	0.04	2.44
		2	3	21	76	13.87	0.13	2.08
		3	5	34	61	14.70	0.16	1.84
	Medium	0	0	0	100	1.24	0	0.76
		1	0	7	93	13.19	0	0.35

Model	Size	$\tilde{\gamma}$	Under	Correct	Over	TP	FN	MMSE
		2	0	49	51	15.17	0	0.27
		3	0	76	24	15.72	0	0.23
	Large	0	0	0	100	0.47	0	0.24
		1	0	11	89	12.81	0	0.11
		2	0	69	31	15.66	0	0.08
		3	0	90	10	15.90	0	0.07

*Under = % underfit models; Correct = % correct models; Over = % of overfit models; TP = avg. number true positives; FN = avg. number of false negatives; MMSE - mean of mean squared errors ($\times 100$).

Plant and insect traits were compiled using existing published and unpublished datasets from the region. Plant traits included: (1) life span: short (annual/biennial) or long (perennial); (2) flower type: single flower or inflorescence; (3) corolla colour: red/blue, white, or yellow; (4) flower size: small, medium, or large; (5) floral symmetry: zygomorphic or actinomorphic; (6) plant origin: native/endemic or introduced; (7) plant morphology: monoecious or dioecious; (8) corolla shape: regular or irregular; and (9) type of plant: herbaceous or woody. Pollinator traits included: (1) minimum and maximum body length (mm); (2) insect behaviour: social or solitary; (3) insect trophic: herbivore or non-herbivore; and (4) insect origin: native/endemic or introduced.

Table 2 shows the unpenalized MLEs and the refit of the final model selected by the adaptive lasso with $\tilde{\gamma}=3$, using BIC to select the final model. Although the BIC value associated with the regularized model was smaller, the difference is small enough (< 10) that the fit of the two models are comparable. However, the regularized model has fewer covariates since flower size, plant type and perennial status have been removed from the model, as has the variable corolla colour yellow.

The plant traits retained by the adaptive lasso may be partially explained by pollination syndromes, which are evolved suites of floral traits (e.g., colour, shape, size, etc.) among flowers pollinated by a particular functional group (e.g., bees, beetles, flies, etc.). For the Terceira Island data set, flies are the dominant guild (Picanço et al., 2017) and the floral traits associated with fly pollination, namely, white corollas (Arnold et al., 2009), symmetric flowers (actinomorphic), and regular (wheel-shaped) corollas, had positive non-zero coefficients. Hence, there is a higher estimated log-odds of pollinator species interacting with plant species that possess these traits. Plant morphology (plants with inflorescences) and flower type (dioecious plants) also had non-zero coefficients. This result is not surprising since pollinators are attracted to inflorescences, which tend to have a greater nectar/pollen reward, and

dioecious plants are generally preferred over monoecious plants (Zito et al., 2016).

Table 2: Results of analysis of Terceira Island network. Unpenalized maximum likelihood estimates and the refit estimators of final model chosen by adaptive lasso with $\tilde{\gamma}=3$, using BIC for parameter tuning.

Predictor	MLE	BIC (refit)
Monoecious	-0.585	-0.605
Corolla Colour White	0.504	0.387
Corolla Colour Yellow	0.150	0
Flower Size Medium	0.127	0
Flower Size Large	0.236	0
Corolla Shape Regular	0.389	0.530
Actinomorphic	0.433	0.282
Inflorescence	0.581	0.518
Plant Type Woody	0.049	0
Perennial	0.038	0
Introduced Plant Species	-0.333	-0.416
Dispersion	-2.996	-2.574
BIC	3258.45	3252.36

However, introduced plant species were associated with a lower log-odds of being visited over native or endemic plant species. From an ecological adaptation perspective, insects may be driven to native/endemic plants because they are familiar or well known. In fact, although there were more than two times the number of introduced (exotic) species than native/endemic species, the proportions of total number of visits to these latter two groups of plants were 54% and 46%, respectively. This result provides some evidence in favour of the notion of ecological adaptation.

5. Discussion and Conclusion

Our regularization of the grouped DM regression model was motivated from an ecological context. Ecologists and evolutionary biologists seek to understand how species traits and linkage rules influence the interactions in mutualistic networks. Recent ecological studies involve collecting counts of interactions along with detailed trait data, and it is for this reason that model selection is necessary for analyzing ecological networks. This new regularized grouped DM regression can simultaneously select the correct covariates in the model and estimate regression parameters with low bias.

The simulation study demonstrated that the adaptive lasso performs better for larger networks, which often contain more data. The final model is conservative in that it tends to select the correct covariates, but also tends to include unnecessary covariates. However, increasing the tuning parameter $\tilde{\gamma}$

can help mitigate model overfit. In the empirical analysis, the BIC associated with the penalized and unpenalized models were similar, but the penalized model was much simpler and could be explained from an ecological perspective.

Potential improvements to model selection consistency could be attained by using an adjusted BIC. For example, Hui et al. (2015) suggest the extended regularization information criteria (ERIC), a modification of the BIC, where the model complexity term is also a function of $\tilde{\gamma}$. Although we studied network sizes motivated by ecological network sizes, extending the framework to the high dimensional setting where $K \gg M$ likely only requires another modification of the BIC, tailored specifically for the high dimensional case. Regardless, regularized grouped DM regression is a useful tool for ecologists and can provide insights into the functional traits driving species' interactions.

References

1. Arnold, S. E., Savolainen, V. and Chittka, L. (2009). Flower colours along an alpine altitude gradient, seen through the eyes of fly and bee pollinators, *Arthropod-Plant Interactions* 3(1): 27-43.
2. Bascompte, J. and Jordano, P. (2007). Plant-Animal Mutualistic Networks: The Architecture of Biodiversity, *Annual Review of Ecology Evolution and Systematics* 38(1): 567-593.
3. Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* 2(1): 183-202.
4. Crea, C. (2017). Variable Selection for Grouped DM regression, https://github.-com/ccrea/ VarSelect_DM_Regression.
5. Crea, C., Ali, R. A. and Rader, R. (2016). A new model for ecological networks using species-level traits, *Methods in Ecology and Evolution* 7(2): 232-241.
6. Guimarães, P. and Lindrooth, R. C. (2007). Controlling for overdispersion in grouped conditional logit models: A computationally simple application of Dirichlet-multinomial regression, *The Econometrics Journal* 10(2): 439-452.
7. Hui, F. K., Warton, D. I. and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using ERIC, *Journal of the American Statistical Association* 110(509): 262-269.
8. McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka (ed.), *Frontiers in econometrics*, Academic Press, New York, pp. 105-142.
9. Montoya, J. M., Pimm, S. L. and Solé, R. V. (2006). Ecological networks and their fragility, *Nature* 442(7100): 259-264.

10. Mosimann, J. E. (1962). On the Compound Multinomial Distribution, the Multivariate β -Distribution, and Correlations Among Proportions, *Biometrika* 49(1/2): 65-82.
11. Picanço, A., Rigal, F., Matthews, T., Cardoso, P. and Borges, P. (2017). Impact of land-use change on flower-visiting insect communities on an oceanic island. *Insect Conservation and Diversity*. 10: 211-223.
12. R Core Team (2017). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria. URL: <https://www.R-project.org/>
13. Santamaría, L. and Rodríguez-Gironés, M. A. (2007). Linkage rules for plant-pollinator networks: trait complementarity or exploitation barriers? *PLoS Biology* 5(2): e31.
14. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267-288.
15. Vázquez, D. P., Blüthgen, N., Cagnolo, L. and Chacoff, N. P. (2009a). Uniting pattern and process in plant-animal mutualistic networks: a review, *Annals of Botany* 103(9): 1445-1457.
16. Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method, **arXiv preprint arXiv:1212.5701** .
17. Zito, P., Scrima, A., Sajeve, M., Carimi, F. and Dötterl, S. (2016). Dimorphism in inflorescence scent of dioecious wild grapevine, *Biochemical Systematics and Ecology* 66: 58-62.
18. Zou, H. (2006). The adaptive lasso and its oracle properties, **Journal of the American Statistical Association** 101(476): 1418-1429.



Change detection and harmonisation of atmospheric large spatiotemporal series



Alessandro Fassò¹, Hsin-Cheng Huangy², Igor Valli¹, Fabio Madonnaz³

¹Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy

²Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

³CNR-IMAA, C.da S. Loja, Tito Scalco, PZ, Italy.

Abstract

The purpose of this paper is to discuss the change detection and harmonisation process for temperature profiles from the Integrated Global Radiosonde Archive (IGRA) which consists of global radiosonde observations dating back to 1905. Harmonisation methods developed for radiosonde have a long history, see for example Haimberger et al. (2012), Thorne et al. (2011) and Sherwood et al. (2008). In this paper, we propose a locally stationary 4D geostatistical model to compute fitted values. The residuals are then used as input to a fused LASSO approach capable of identifying changes.

1. Introduction

The purpose of this paper is to discuss the change detection and harmonisation process for temperature profiles from the Integrated Global Radiosonde Archive (IGRA).

IGRA consists of radiosonde and pilot balloon observations at over 2,700 globally distributed stations. The earliest data date back to 1905, and recent data become available in near real time. Observations are available at standard and variable pressure levels, fixed- and variable-height wind levels, and the surface and tropopause. Variables include pressure, temperature, geopotential height, relative humidity, dewpoint depression, wind direction and speed, and elapsed time since launch.

Harmonisation methods developed for radiosonde have a long history, see for example Haimberger et al. (2012), Thorne et al. (2011) and Sherwood et al. (2008).

In this paper, the focus is on undocumented changes in single stations. In particular permanent step changes, temporary step changes and impulses or outliers are considered.

The idea is to use a simple locally stationary 4D geostatistical model to compute fitted values. The residuals are then used as input to a fused LASSO model capable of identifying all the above changes.

2. Data and notation

We consider a 40 year temporal domain $t \in D_T = [1, \dots, T]$ with 12 hour time step starting from 1 Jan 1978 00:00UTC and ending 31 Dec 2017. The spatial domain is denoted by $D_s = \{s_1, \dots, s_n\}$ representing 655 "good" stations of the IGRA network. Altitude of measurements is expressed in pressure level $925 \geq h \geq 50 \text{ hPa}$.

Let $y(s, t, h)$ denote the response process (e.g. temperature in Kelvin) and let the data be denoted by:

$$y(s_j, t_j, h_{j,t,i}) \tag{1}$$

where

1. $s_j = (lat_j, lon_i) \in D_s$ are the coordinates of the $j - th$ station;
2. $t_j \in D_T$ is time at 12h scale;
3. $h_{j,t,i} = 1, \dots, q_{t,j} \leq q_{max}$ is altitude expressed in hPa $h \in [925, 50]$.

Let $y_{t,j} = y_t(s_j)$ be the $q_{t,j}$ dimensional vector of all data at time t and station s_j and let $y_t = (y'_{t,1}, \dots, y'_{t,n})'$ be the $q_{t,o}$ dimensional vector of all data at time t , with $q_{t,o} = \sum_{j=1}^n q_{t,j}$. For the sake of simplicity, we will assume that altitudes are time invariant. That is $q_{t,j} = q, h_{j,t,i} = h_{j,i}$ and nq dimensional vector. This is the case of the so-called mandatory levels.

Eventually, let $Y_{t;t+a} = (y'_t, \dots, y'_{t+a})'$ and let $Y = Y_{1:T}$ be the full dataset vector.

3. A 4D Gaussian Process

Consider the Gaussian process (GP) with 4D continuous index, given by

$$y(s, t, h) = z(s, t, h) + \varepsilon(s, t, h) \tag{2}$$

where $s = (lat; lon) \in Spherical\ shell, h \in [925, 50]$ and let $u = (s, t, h)$.

The data described in Equation (1) are observations at possibly non-regular time points $t \in D_T$, and spatial points $D_s = \{s_1, \dots, s_n\}$. Hence, using the $nq - dim$ vector y_t , the above equation is written as

$$y_t = z_t + \varepsilon_t$$

where $E(y_t) = \mu$ and covariance function given by

$$Cov(y(u), y(u')) = \sigma_z^2 \exp\left(-\sum_{j=1}^4 \frac{|u_j - u'_j|}{\theta_{t,j}}\right) + I(u = u')\sigma_\varepsilon^2.$$

The corresponding GP parameter set is $\psi = (\mu, \theta_1, \dots, \theta_4, \sigma_z^2, \sigma_\varepsilon^2)$.

Although the above setup is time invariant and defines a stationary GP, in the sequel, we assume local stationarity with respect to both space and time.

4. Recursive (local) estimation

We consider a time interval Δt for local estimation, e.g. 30 days, and define the $\Delta t - periods$ index by $m = m(t) = 1, \dots, M$. If Δt is constant, and assuming $T = \Delta t \times M$, we have $m(t) = int\left(\frac{t-t_0}{\Delta t}\right) = 1, \dots, M$. Moreover, let

$\tau(m)$ be the set of time steps corresponding to the $m - th$ period. If Δt is constant, than $\tau(m) = \Delta t(m - 1) + 1 : \Delta tm$.

Now, the maximum likelihood estimate for $m - th$ period is given by

$$\hat{\psi}_m = \hat{\psi}_{T(m)} = \arg \max_{\psi} \log L_{\psi}(Y_{T(m)})$$

and the fitted values for $y_t(s_i), t \in \tau(m)$ are obtained by

$$y_t(s_j|a) = E_{\psi_{m(t)}}^0(y_t(s)|Y_{t-a:t+a}^{(\sim j)}).$$

In the last equation, $Y_{t-a:t+a}^{(\sim j)}$ are the data in a neighborhood of s_j, s_j excluded.

If $\psi_{m+1} \neq \psi_m$, we may have an artificial discontinuity of \hat{y} around the border of the $m - th$ and $(m + 1) - th$ periods. For change detection, this discontinuity is not a concern as long as it does not propagate to the residuals. In any case, this issue may be substantially mitigated using a smooth update of $\hat{\psi}_m$.

In particular, an approach similar to the recursive estimation of dynamical models of Grillenzoni (1994 and 1997) considers a single step of a Newton-Raphson algorithm at each Δt -period:

$$\hat{\psi}_m = \hat{\psi}_{m-1} + l'_m(l''_m)^{-1} \tag{4}$$

where l' and l'' are Jacobian and Hessian of $\log L_{\psi}(Y_{\tau(m)})$.

5. Fused LASSO change detection

Let us denote the change model for a fixed station s_j and altitude h , by

$$y_t = y_t^0 + \beta_t$$

where y^0 is the zero mean GP of Equation (2) and β_t defines isolated, temporary or permanent changes as discussed in Section 1. Formally, β_t is a deterministic, piece-wise constant function with change points at an unknown number $k \geq 0$, of unknown change points $t_1^*, \dots, t_k^* \in \{1, \dots, T\}$.

Once the fitted values of the previous section are available for the station at s_j , the following (sign changed) residuals are computed:

$$e_t = e(s_j, t, h) = y(s_j, t, h) - \hat{y}(s_j, t, h)$$

for $t = 1, \dots, T, h \in [h_1, \dots, h_q]$. Now, conditionally on the GP parameter estimates of the previous section, we have that

$$E(e_t) = \beta_t.$$

We than test for changes the station in s_j using the fused LASSO approach, Tibshirani et al. (2005). In particular, we assume that

$$e_t = \beta_t + \zeta_t$$

where $\zeta_t \equiv NID(0, \sigma^2)$.

The model (5) has T parameters β , plus $\sigma_{\epsilon}^2 = Var(\epsilon_t)$. Hence, we regularise the estimation using the following penalised criterion:

$$\sum_{t=1}^T (e_t - \beta_t)^2 + \lambda_1 \sum_{t=1}^T |\beta_t| + \lambda_2 \sum_{t=2}^T |\beta_t - \beta_{t-1}|$$

where the first penalty term controls the number of $\beta_t \neq 0$ and the second one controls smoothness of β_t ; hence identifying temporary and permanent changes of Section 1.

In practice the choice of λ_1 and λ_2 for step changes and for impulses may be quite different. Hence, we will focus on the former problem.

In Equation (6); an important case is $\lambda_1 = 0$, known as mean filtering in signal processing (e.g. Ottersten et al., 2016). In this case, we rewrite Equation (6) in matrix form

$$\|e - \beta\|_2^2 + \lambda_2 \|D\beta\|_1$$

where D is the first order difference. Hence reparametrizing $\delta = D\beta$, we have

$$\|e - D^{-1}\delta\|_2^2 + \lambda_2 \|\delta\|_1 \quad (7)$$

Nonetheless, optimising Equation (6) with SLEP package (Liu et al., 2010 and 2015) is computationally more efficient than optimising Equation (7) with lasso function of Matlab package. Moreover preliminary results show that it is also more stable and flexible. Infact λ_1 may be optimised using CV, giving small but non-null $\lambda_1 > 0$.

Harmonisation After β_t are estimated, say $\hat{\beta}_t$, the harmonised measurement are obtained by

$$y_t^* = y_t - \hat{\beta}_t$$

with harmonisation uncertainty

$$\sigma(\hat{\beta}_t)$$

which may be approximately computed using Theorem 1 of Tibshirani et al. (2005).

References

1. Grillenzoni C. (1994) Optimal recursive estimation of dynamic models. *Journal of the American Statistical Association*. 89:427. 777-787.
2. Grillenzoni C. (1997) Recursive generalized M-estimators of system parameters. *Technometrics*. 39:2, 211-224.
3. Haimberger, L., Tavolato, C., Sperka, S., (2012) Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *J. Clim.* 25, 8108.8131.
4. Liu J., Ji S., and Ye J. (2015) SLEP: Sparse Learning with Efficient Projections, Version 4.1. <https://github.com/divelab/SLEP>.
5. Liu J., Yuan L, and Ye J., (2010) An Efficient Algorithm for a Class of Fused Lasso Problems, KDD.
6. Ottersten J., Wahlberg B., Rojas C.R., (2016) Accurate Changing Point Detection for l1 Mean Filtering, *IEEE SIGNAL PROCESSING LETTERS*, 23:2, 297-301.

7. Sherwood, S.C., Meyer, C.L., Allen, R.J., Titchner, H.A., (2008) Robust tropospheric warming revealed by interactively homogenised radiosonde data. *J. Clim.* 21, 5336.5352.
8. Tibshirani R., Saunders M., Rosset A., Heights Y., Zhu J., Arbor A., and Knight K. (2005) Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:1, 91.108.
9. Thorne, P.W., Brohan, P., Titchner, H.A., et al., (2011) A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.* 116, D12116.



Geographic variation, trends and determinants of hypertension in South African adult population, 2008 -2017



Glory Atilola¹, Ngianga-Bakwin Kandala^{1,2}, Guangquan Li¹; Samuel Manda³

¹Northumbria University, Department of Mathematics, Physics and Electrical Engineering, Faculty of Engineering and Environment, Newcastle upon Tyne, NE1 8ST, United Kingdom

²University of the Witwatersrand, Division of Epidemiology and Biostatistics, School of Public Health, Johannesburg, South Africa

³Biostatistics Research Unit, South African Medical Research Council, Pretoria South Africa

Abstract

The burden of hypertension in Sub-Saharan Africa in the past two decades has become a serious cause for regional concern. High blood pressure is the most powerful predictor of stroke and other cardiovascular outcomes in the region. An estimated 11.1% rise in regional burden was reported between 1990 and 2010 along with a projected increase of 216.8 million affected individuals by 2030 from 130.2 million in 2010 across the region especially in South Africa. In this study, we investigate hotspots, temporal trends and determinants of prevalent hypertension in South African adult population between 2008 and 2017. We utilized spatial and spatio-temporal structured geo-additive regression models to map hotspots, and temporal trends of hypertension in South African adult population between 2008 and 2017. Known individual risk factors were controlled for. Analysis was conducted for four consecutive national income dynamics household surveys. Implementation was carried out within Bayesian framework using MCMC simulation methods. An overall absolute decline of 6% point and a relative decline of 24% in the prevalence of hypertension were found at the national level between 2012 and 2017. Hotspots exist across districts in Western Cape and Eastern Cape while districts in Limpopo province had considerably low risk of hypertension. This pattern of geographic variation in risk was consistently observed over the four consecutive time points between 2008 and 2017. Controlling for known individual risk factors explained a substantial amount of geographic variation in risk over time except in RSM in North West, Uthukela and Ugu districts in Kwazulu-Natal where average risk of hypertension remained high. Risk factors of hypertension in South African adult population include age, coloured population group, education, lack of exercise and diabetes. Study findings demonstrate evidence that South Africa is making progress towards the national 2020 target of 25% reduction in prevalent hypertension at the national level. However, considerable variation in risk exists across the districts. Cost-effective policy responses to the emerging trends in cardiovascular disease burden across the region depend on accurate estimates of distribution and determinants of cardiovascular health outcomes at both national and

small area levels. Understanding this dynamic in context will improve identification of high risk groups for effective targeting of public health interventions and resource allocation in resource limited settings.

Keywords

High blood pressure; Geo-additive models; CVD risk factors; Bayesian MCMC; Space-time Modeling

1. Introduction

The burden of hypertension in Low and Middle-income countries in the past two decades has become a serious cause for regional concern. High blood pressure is a leading predictor of stroke and other cardiovascular outcomes due to both chronic and communicable diseases across Sub-Saharan Africa (Agyei-Mensah). Recent studies and reports have demonstrated that these changes are, to a large extent, driven by lifestyle changes such as increase in alcohol consumption and poor dietary choices, as well as decline in physical activity over time due to rapid urbanization and population growth (Aikins et al, 2010).

In South Africa, the burden of hypertension remains an important health system challenge with an estimated prevalence burden of 77.9%, the highest of any country in Sub-Saharan Africa (Lloyd-Sherlock et al, 2014). More so, the evolving public health significance of geographic location continues to shape patterns of health and disease outcomes across the region. Few studies so far have attempted to study the role of geography (small area) as a primary exposure risk in the observed prevalence patterns of hypertension and other chronic diseases in Sub-Saharan Africa (Kandala et al, 2013; Kandala and Stranges 2014; Weimman, 2016).

Two previous studies have examined the spatial variation in hypertension in South African adult population (Kandala et al, 2013; Kandala et al, 2013). While both studies provide important insight into the spatial epidemiology of hypertension in South Africa, the South African demographic and health (DHS) survey data analyzed was conducted in 1998, thereby leaving a significant gap with regard to current situation since the turn of the millennium. In this paper we aim to quantify the burden of hypertension from 2008 to 2017 in South African adult population by mapping geographical variations in risk and to evaluate trend before and after the launch of the National Strategic policy in 2012 to monitor cardiovascular morbidities across the districts (DoHSA, 2013).

2. Methodology

Outcome variable: The primary outcome is the Bernoulli distribution of hypertension (also known as raised blood pressure) in South Africa adult

population. We considered a condition at cut-point of blood pressure (BP) $\geq 140/90$ mmHg or self-reported diagnosis or on medication as captured in the NiDS for the four consecutive surveys.

Exposure variables: A key exposure variable investigated in this study was the effect of geographic location of respondent (at the time of the survey) on the risk of hypertension given that closer districts (neighbors) are more likely to have similar disease patterns. In addition, we controlled for known individual risk factor variables such as the age of respondent (as a continuous covariate), gender, race (Black African/Coloured/Indian-Asian/White), and educational attainment. Others include lifestyle factors such as exercise, alcohol and binary indicators for smoking, diabetes, fever and arthritis.

Statistical Analysis: We considered the class of Bayesian generalized geo-additive mixed regression models in which the probability of hypertension in individual i ($i = 1, \dots, n_{ijt}$) in district j ($j = 1, \dots, S$) at time t ($t = 1, \dots, T$), follows a Bernoulli distribution with mean $\pi_{ijt} = E(y_{ijt}|X, \beta)$. Using appropriate prior specification, we modelled the likelihood of hypertension by replacing the linear predictor with a more flexible structured additive predictor with a logit link specification. The flexibility of this class of models allows us to account for nonlinear effects of continuous covariates, spatial heterogeneity and spatial dependency structure between neighbouring districts as well as temporal dependence in the observed data within a unified framework. Full Bayesian inference was implemented using Markov Chain Monte Carlo simulation method. Model evaluation and comparison were carried out using the Deviance Information Criterion (Spiegelhalter et al. 2002).

Model 1: Spatial model regression framework

$$y_{ij} \sim \text{Bern}(\pi_{ij})$$

$$\eta_{ij} = \text{logit}(\pi_{ij}) = \beta_0 + \mathbf{X}_{ij}\beta + f_1(\text{Age}_{ij}) + f_{\text{spat}}(\text{district})$$

$$f_{\text{spat}}(\text{district}) = f_{\text{unstr}}(\text{district}) + f_{\text{str}}(\text{district})$$

Model 2: Spatio-temporal model regression framework

$$y_{ijt} \sim \text{Bern}(\pi_{ijt})$$

$$\eta_{ijt} = \text{logit}(\pi_{ijt}) = \beta_0 + f_1(\text{Age}_{ijt}) + f_{\text{spat}}(\text{district}_{jt}) + f_2(\text{survey year})$$

$$f_{\text{spat}}(\text{district}_{jt}) = f_{\text{unstr}}(\text{district}_{jt}) + f_{\text{str}}(\text{district}_{jt})$$

Where η_{ij} and η_{ijt} are the structured additive predictor with a logit link function, $f_1(\text{Age}_{ijt})$ and $f_2(\text{survey year})$ are the nonlinear effect of age and year modelled as nonparametric smooth function using Bayesian P-Spline with a second order random walk. In addition, $\mathbf{x}_{ij}\beta$ are the vectors of covariate values with their unknown regression parameters assigned a vague prior distribution. The spatial effect of district $f_{\text{spat}}(\text{district})$ is decomposed into

spatially structured, $f_{str}(district)$ and unstructured effects $f_{unstr}(district)$ of district on the log likelihood of hypertension in individual i at time t and modelled using the Markov random field (Besag et al, 1991).

3. Results

In this section, we present our findings to understand the spatial epidemiology of hypertension in South Africa between the period 2008 and 2017 using evidence from the National income dynamics survey data. National prevalence of hypertension was estimated at 23.7%, 24.9%, 19.7% and 19.0% in 2008, 2012, 2015 and 2017 respectively in the South African adult population.

Evidence of geographic variation in hypertension was found across the 52 districts especially between low risk northern and high risk southern districts in South Africa. Significant hotspots were found across districts in Western Cape and Eastern Cape provinces while districts in Limpopo province had significantly low risk of hypertension between 2012 and 2017 (Fig. 1 & Fig. 2). However, controlling for known individual risk factors explained a substantial amount of geographic variation in risk over time except in RSM in North West, Uthukela and Ugu districts in Kwazulu-Natal where average risk of hypertension remained significantly high (Fig. 3a)

Risk factors of hypertension in South African female adult population include coloured population group, low education attainment, lack of exercise and diabetes (Table 1). A linear trend was found in the nonlinear effects of age on hypertension peaking at 80 years. Evidence of considerable decline in prevalent hypertension was found between 2012 and 2017 (Fig. 3b).

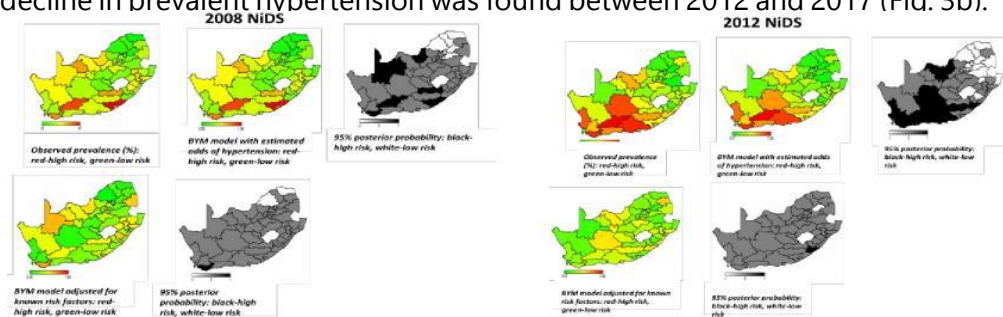


Figure 1: Observed prevalent hypertension(a), BYM model of posterior odds of hypertension(b) along with 95% posterior probability(c) and BYM model adjusted for known individual risk factors(d) along with 95% posterior probability(e) in 2008(left) and 2012(right) among South African Adults. (NB:Red colour indicates high risk districts. Green colour indicates low risk districts. Black colour indicates significantly high risk areas. White colour indicates significantly low risk districts. Grey colour indicates non-significance).

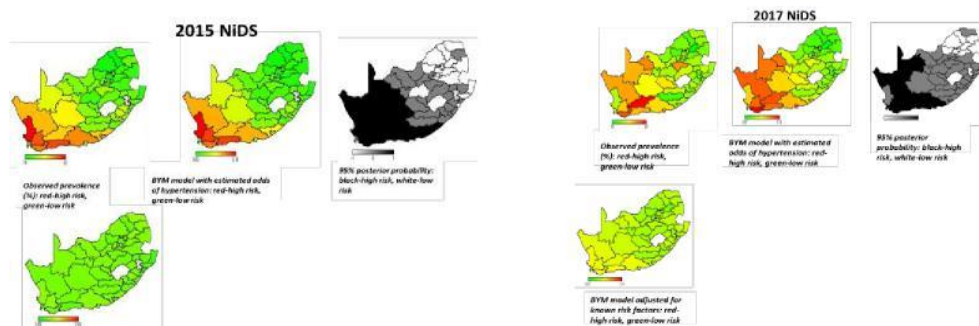


Figure 2: Observed prevalent hypertension(a), BYM model of posterior odds of hypertension(b) along with 95% posterior probability(c) and BYM model adjusted for known individual risk factors(d) in 2015(left) and 2017(right) among South African Adults. (NB: Red colour indicates high risk districts. Green colour indicates low risk districts. Black colour indicates significantly high risk areas. White colour indicates significantly low risk districts. Grey colour indicates non-significance).

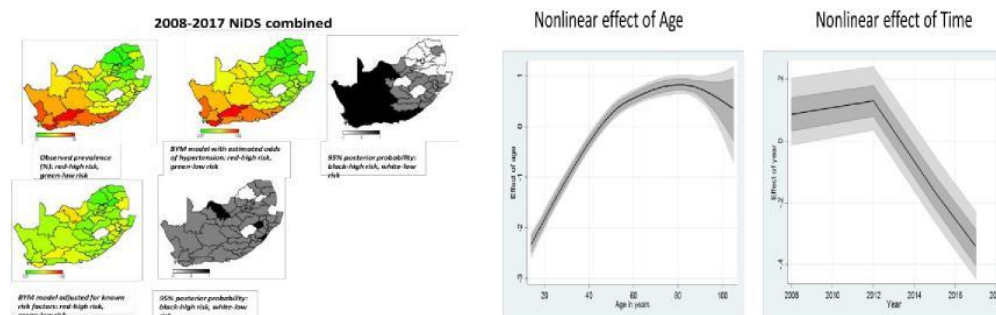


Figure 3a&3b: Observed prevalent hypertension(a), BYM model of posterior odds of hypertension(b) with 95% posterior probability(c) and BYM model adjusted for known individual risk factors (d) along with 95% posterior probability(e) for the four surveys combined among South African Adults(3a) Nonlinear effects of Age and survey year(3b). (NB: Red colour indicates high risk districts. Green colour indicates low risk districts. Black colour indicates significantly high risk areas. White colour indicates significantly low risk districts. Grey colour indicates non-significance).

Table 1. Bayesian Structured Geo-additive binary regression models for Hypertension in South African adult population, 2008-2017

	2008	2012	2015	2017	2008-2017
Predictor	OR(95%CI)	OR(95%CI)	OR(95%CI)	OR(95%CI)	OR(95%CI)
Gender					
male	1.00	1.00	1.00	1.00	1.00
female	0.86(0.73-1.02)	0.85(0.77-0.93)	0.78(0.69-0.88)	0.73(0.61-0.89)	0.79(0.75-0.84)
Geotype					
farm	1.00	1.00	1.00	-	-
traditional	0.86 (0.68-1.01)	1.05(0.88-1.25)	0.81(0.65-1.04)	-	-
urban	0.86 (0.66-1.13)	1.01(0.85-1.19)	0.85(0.67-1.08)	-	-
Race					
african	1.00	1.00	1.00	1.00	1.00
coloured	1.23 (0.93-1.63)	1.26(1.03-1.55)	1.43(1.10-1.86)	1.14(0.78-1.68)	1.28(1.13-1.47)
asian/indian	0.87 (0.39-1.86)	0.66(0.37-1.13)	0.76(0.41-1.39)	0.84(0.47-1.61)	0.79(0.57-1.11)
white	0.43(0.31-0.58)	0.52(0.40-0.69)	0.65(0.43-0.93)	0.78(0.58-1.03)	0.58(0.50-0.67)

	2008	2012	2015	2017	2008-2017
Education					
higher	1.00	1.00	1.00	-	-
secondary	1.49(1.19-1.88)	1.21(1.04-1.40)	1.18(1.01-1.38)	-	-
primary	1.35(1.04-1.78)	1.17(0.99-1.39)	1.22(0.99-1.51)	-	-
none	1.43(1.05-1.93)	1.02(0.85-1.23)	1.19(0.94-1.51)	-	-
Marital status					
never married	1.00	1.00	1.00	-	-
married	0.92(0.76-1.13)	0.93(0.82-1.04)	0.99(0.86-1.15)	-	-
living with part	1.05(0.82-1.38)	0.99(0.83-1.17)	1.12(0.89-1.40)	-	-
widow	0.97(0.74-1.26)	0.98(0.84-1.15)	1.11(0.92-1.34)	-	-
divorce/separa	0.90(0.61-1.34)	0.97(0.75-1.26)	1.36(0.96-1.91)	-	-
Religion					
christian	1.00	1.00	1.00	1.00	1.00
muslim	0.56(0.22-1.44)	1.20(0.60-2.31)	0.56(0.21-1.23)	1.18(0.38-3.24)	0.85(0.56-1.28)
hindu	1.24(0.56-2.78)	1.42(0.75-2.58)	0.71(0.35-1.56)	1.17(0.59-2.35)	1.12(0.80-1.59)
traditional	0.81(0.57-1.16)	0.82(0.70-0.96)	0.96(0.80-1.16)	0.97(0.71-1.33)	0.89(0.80-0.99)
none	1.34(1.04-1.76)	0.98(0.83-1.15)	1.21(0.95-1.51)	1.20(0.89-1.65)	1.16(1.05-1.29)
Province					
gauteng	1.00	1.00	1.00	1.00	1.00
eastern cape	0.93(0.50-1.65)	1.05(0.73-1.44)	1.24(0.91-1.75)	1.07(0.73-1.54)	1.09(0.89-1.32)
northern cape	0.67(0.33-1.26)	0.84(0.58-1.23)	1.22(0.87-1.79)	1.21(0.73-1.93)	0.91(0.73-1.13)
free state	0.94(0.46-1.88)	0.83(0.59-1.18)	1.01(0.74-1.42)	1.07(0.73-1.60)	0.96(0.78-1.15)
kwazulu-natal	0.50(0.28-0.88)	0.94(0.69-1.27)	1.36(1.01-1.88)	0.95(0.70-1.28)	0.90(0.75-1.04)
north west	0.68(0.40-1.20)	1.08(0.79-1.49)	1.05(0.77-1.47)	0.74(0.47-1.10)	0.95(0.79-1.16)
western cape	0.78(0.43-1.35)	1.26(0.87-1.85)	1.60(1.09-2.42)	0.97(0.59-1.48)	1.22(0.99-1.50)
mpumalanga	0.85(0.46-1.61)	0.73(0.52-1.03)	0.89(0.64-1.22)	0.63(0.43-0.93)	0.71(0.58-0.86)
limpopo	0.81(0.42-1.55)	0.70(0.50-0.99)	0.77(0.57-1.04)	0.60(0.39-0.91)	0.76(0.63-0.93)
Employment					
employed	1.00	1.00	1.00	-	-
unemployed	1.20(0.91-1.59)	0.89(0.67-1.18)	0.74(0.44-1.18)	-	-
unemployed	1.00(0.78-1.27)	1.01(0.87-1.15)	0.94(0.78-1.12)	-	-
not active	1.02(0.86-1.21)	0.99(0.89-1.10)	0.99(0.87-1.13)	-	-
Exercise					
never	1.00	1.00	1.00	1.00	1.00
< once a week	1.03(0.76-1.40)	0.89(0.74-1.06)	1.02(0.83-1.26)	0.83(0.60-1.12)	0.91(0.82-1.02)
1-2 times a	1.03(0.80-1.28)	1.04(0.89-1.23)	0.92(0.76-1.11)	0.85(0.65-1.13)	0.95(0.86-1.05)
≥ 3 times a	1.11(0.89-1.94)	1.09(0.94-1.27)	0.91(0.76-1.09)	0.96(0.73-1.27)	1.01(0.92-1.11)
Alcohol					
never	1.00	1.00	1.00	-	-
no longer	1.09(0.86-1.36)	0.98(0.83-1.16)	1.08(0.91-1.26)	-	-
rarely	1.11(0.90-1.36)	1.28(1.13-1.46)	1.29(1.09-1.52)	-	-
1-2 days a week	1.30(0.89-1.94)	1.34(1.06-1.68)	1.76(1.35-2.26)	-	-
3-4 days a week	1.92(1.02-3.73)	0.94(0.59-1.47)	1.29(0.76-2.22)	-	-
5-6 days a week	2.15(0.78-5.70)	1.25(0.68-2.26)	1.10(0.40-2.78)	-	-
daily	1.24(0.58-2.57)	0.51(0.22-1.09)	3.81(1.15-15.64)	-	-
Cigarette					
no	1.00	1.00	1.00	1.00	1.00
yes	0.81(0.61-1.07)	0.90(0.72-1.12)	0.73(0.57-0.93)	0.88(0.64-1.23)	0.85(0.74-0.96)

	2008	2012	2015	2017	2008-2017
Fever					
no	1.00	1.00	1.00	1.00	1.00
yes	1.06(0.87-1.29)	1.09(0.97-1.23)	1.18(1.03-1.35)	1.18(0.97-1.43)	1.12(1.04-1.20)
Persistent Cough					
no	1.00	1.00	1.00	1.00	1.00
yes	1.10(0.87-1.38)	0.92(0.81-1.04)	0.88(0.73-1.05)	0.74(0.55-0.98)	0.93(0.85-1.02)
Chest Pain					
no	1.00	1.00	1.00	1.00	1.00
yes	0.91(0.72-1.12)	0.90(0.77-1.05)	0.83(0.69-1.01)	0.95(0.69-1.31)	0.90(0.81-0.99)
Joint					
no	1.00	1.00	1.00	1.00	1.00
yes	1.14(0.94-1.38)	1.15(1.00-1.30)	1.01(0.86-1.18)	1.22(0.96-1.56)	1.12(1.03-1.22)
Weight Loss					
no	1.00	1.00	1.00	1.00	1.00
yes	0.95(0.63-1.35)	0.84(0.66-1.07)	0.74(0.55-0.99)	0.61(0.35-0.99)	0.82(0.70-0.96)
Diabetes					
no	1.00	1.00	1.00	1.00	1.00
yes	1.37(1.05-1.77)	1.22(1.04-1.41)	1.14(0.88-1.51)	1.17(0.85-1.59)	1.19(1.06-1.33)
Year					
2008	-	-	-	-	1.00
2012	-	-	-	-	1.01(0.88-1.18)
2015	-	-	-	-	0.72(0.62-0.83)
2017	-	-	-	-	0.61(0.52-0.71)

4. Discussion and Conclusion

Using a combination of advanced statistical and GIS methods, this study was able to quantify the spatial variation in hypertension at the sub-national level of district, and evaluate temporal trends in prevalent hypertension. At the same time, determinants of hypertension in South African adult population were identified from 2008 to 2017. District municipalities across Western and Eastern Cape provinces had highest odds of hypertension while majority of districts in Limpopo province showed consistently low levels of hypertension burden across the four surveys. Geographic variation may be due to higher concentration of urban communities in the Western Cape as well as parts of Eastern Cape and North West relative to Limpopo and Mpumalanga provinces. Also, the extent to which prevention and control strategies/policies are effectively implemented across district municipalities from 2014 to 2017, may partly explain the observed trend. However, this is not immediately clear as no district level information was collected. Risk factors of hypertension in South African adult population include age, coloured population group, education, lack of exercise and diabetes. A similar pattern of high prevalence rate of

hypertension among the coloured population group was found in the age-adjusted prevalence and model-based estimates after accounting for the confounding effects of other factors. Findings by Kandala et al, 2013 and Peltzer & Phaswana-Mafuya (2013) were consistent with our findings with respect to geographic variation and risk factors of hypertension in South African adult population.

In conclusion, we consider this study a critical data-driven evidence for program managers, policy makers, and international stakeholders to understand the geography and determinants of hypertension among South African adults over time. This will ensure that context-specific interventions are provided in a cost-effective and efficient manner to optimize programme outcome and impact. Future research effort will evaluate trends and determinants in the male and female adult populations at district level as well as interaction between space and time in the risk of hypertension outcomes at small area level. Study limitations include the use of a binary indicator for smoking status and other known risk factors of hypertension such as salt intake and air pollution which were not accounted for in geographic variation of prevalent hypertension.

References

1. Agyei-Mensah S, and Aikins A.(2010). Epidemiological transition and the double burden of disease in Accra, Ghana. *J Urban Health*. 87(5): 879-897
2. Department of Health, South Africa. (2013). The South African National strategic plan for the prevention of Non-Communicable diseases, 2013-2017.
3. Aikins A, Unwin N, Agyemang C, Allotey P, Campbell C, and Arhinful D.(2010). Tackling Africa's chronic disease burden: from the local to the global. *Globalization and Health*.6:5
4. Maimela E, Alberts M, Modjadji SEP, Choma SSR, Dikotope SA, Ntuli TS, et al. (2016) The Prevalence and Determinants of Chronic Non-Communicable Disease Risk Factors amongst Adults in the Dikgale Health Demographic and Surveillance System (HDSS) Site, Limpopo Province of South Africa. *PLoS ONE* 11(2): e0147926.
5. Lloyd-Sherlock P, Beard J, Minicuci N, Ebrahim S, & Chatterji S. Hypertension among older adults in low-and middle-income countries: prevalence, awareness and control. *International Journal of Epidemiology*. 2014; 43(1): 116-128
6. Kandala N-B., Tigbe W., Manda S., & Stranges S. (2013) Geographic Variation of Hypertension in Sub-Saharan Africa: A Case Study of South Africa, *American Journal of Hypertension*,26(3): 382–391

7. Kandala N-B., & Stranges S (2014). Geographic Variation of Overweight and Obesity among Women in Nigeria: A Case for Nutritional Transition in Sub-Saharan Africa. *PLoS ONE*, 9(6): e101103. <https://doi.org/10.1371/journal.pone.0101103>
8. Weimann A., Dai D., & Oni T.(2016). A cross-sectional and spatial analysis of the prevalence of multimorbidity and its association with socioeconomic disadvantage in South Africa: A comparison between 2008 and 2012. *Social Science & Medicine*. 163(2016): 144-156
9. Kandala NB., Manda OM., Tigbe W, Mwambi H, and Stranges S.(2013). Geographic distribution of cardiovascular comorbidities in South Africa: a national cross-sectional analysis. *Journal of Applied Statistics*, 41:6, 1203-1216
10. Fahrmeir L., Lang S., and Spies F.(2003). Generalized geoaddivitive models for insurance claims data. *Blatter der DGVFM*, 26(1): 7-23
11. Besag, J, York J, and Mollie A. (1991). A Bayesian Image Restoration with two applications in spatial statistics, 43(1): 1-59.
12. Peltzer K., & Phaswana-Mafuya N. (2013). Hypertension and associated factors in older adults in South Africa. *Cardiovascular journal of Africa*, 24(3): 67–71. doi:10.5830/CVJA-2013-002
13. Spiegelhalter J., Best G., Carlin P., and Van der Linde A. (2002). Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society, Series B*; 64(4):583-616



Modelling and mapping prevalence of Female Genital Mutilation/C (FGM/C) among 0-14 years old girls in Kenya, Nigeria and Senegal



Chibuzor Christopher Nnanatu¹, Glory Atilola¹, Paul Komba¹, Lubanzadio Mavatikua¹, Zhuzhi Moore², Dennis Matanda³, Ngianga-Bakwin Kandala¹

¹Northumbria University, Newcastle, UK

²Population Council, Kenya

³Independent Consultant

Abstract

World Health Organisation defines Female Genital Mutilation/cutting (FGM/C) as all forms of injury caused to the external female genitalia for non-medical reasons. FGM/C is a public health and human right issue, which is strongly anchored on customs and traditions, without any established benefit. The practice has both short- and long-term consequences ranging from haemorrhage to complications during child birth. It is estimated that over 200 million women and girls alive today globally, have undergone FGM/C at some point in their lives. FGM/C is rampant in Africa where it is feared that some 3 million girls are at risk of being cut each year. Recent studies showed that FGM/C prevalence among women aged 15-49 in Kenya was estimated at 27.1% in 2008-9. On the other hand, in 2017, FGM/C prevalence among girls aged 0-14 years was estimated at 14.0% and 25.3% in Senegal and Nigeria, respectively. There are several change-provoking interventions geared towards eliminating the practice. Consequently, change has begun but rather sluggish, and this calls for the generation of credible statistical evidence that sufficiently describes where, when and how change is taking place. Robust Bayesian hierarchical space-time models which simultaneously accounted for unobserved effects of space and time, as well as space-time interactions, whilst controlling for other linear and nonlinear covariates were employed. These models were developed and fitted on the available datasets in a coherent mixed models regression framework. Posterior inference was carried out using Markov Chain Monte Carlo (MCMC) techniques, while model fit and complexity assessments utilised Deviance Information Criterion (DIC) approach. The approach adopted in this study allowed us to jointly account for individual-, household-, community-level factors, map and identify patterns and spatial and temporal variations in the practice, thus unmasking FGM/C hotspots and patterns over time, as well as their characteristics across the three countries. Factors found to associate with higher risk of the practice included mother's FGM/C status, support for FGM/C continuation, household wealth index, level of education of mother, region and type of place of residence, marital status and religion. Our findings are important in various ways: First, it is now clear, at least to an extent, where and when changes are

taking place in Kenya, Nigeria and Senegal. Second, the characteristics of the identified hotspots may be exploited by policymakers and programme implementers in the design and evaluation of bespoke programmatic interventions.

Keyword

Bayesian Geo-additive models; Spatial modelling; Space-time interactions; Female circumcision; Social norms

1. Introduction

World Health Organisation defines Female Genital Mutilation/cutting (FGM/C) as all forms of injury caused to the external female genitalia for non-medical reasons [1]. FGM/C is a public health and human right issue, which is deeply rooted in customs and traditions. The practice has both short-term and longterm consequences with immediate consequences including haemorrhage and shock. while long term consequences include increased risk of complications during child birth [2]. It is estimated that over 200 million women and girls alive today globally, have undergone FGM/C at some point in their lives. FGM/C is a common practice in most African countries with some 3 million girls being at risk of cutting each year [4].

Recent studies showed that FGM/C prevalence among women aged 15-49 in Kenya was estimated at 27.1% in 2008-9. On the other hand, in 2017, FGM/C prevalence among girls aged 0-14 years was estimated at 14.0% and 25.3% in Senegal and Nigeria, respectively [3]. There are several programmatic interventions in the affected countries geared towards eliminating the practice. Consequently, decline in prevalence has been reported albeit sluggishly.

This study aims to

- 1) Identify and map FGM/C hotspots in Nigeria, Kenya and Senegal.
- 2) Identify the key individual-level and community-level factors and see how these compare across the three countries.

2. Methodology

Data Sources

Data on FGM/C prevalence in Nigeria were drawn from six nationally-representative surveys from Nigeria Demographic and Health Surveys (DHS) and Nigeria Multiple Indicators Cluster Surveys (MICS) comprising of 2003DHS, 2007MICS, 2008DHS, 2011MICS, 2013DHS, and 2016-17MICS. Data from FGM/C prevalence 0-14 years old in Kenya were drawn KDHS 1998, KDHS 2003, KDHS 2008, KDHS 2014. Finally, data on prevalence among Senegalese girls were drawn from 2005 SDHS, 2010-11SDHS, 2015 SDHS, and 2017SDHS

Response variable

The main response variable is *whether respondent's daughter(s) underwent FGM/C?* This was coded as a binary variable where a value of 1 denotes that daughter was cut and 0 denotes that daughter was not cut.

Exposure variables

Covariates included in the community-level spatial model were indicators of social norms with the following surrogate variables: (1) *mother's FGM/C status* and (2) *mother's support for FGM/C continuation*. Individual-level covariates comprised the girl and her mother's background characteristics, their geographical location -region and state of residence, type of place of residence (urban vs rural), socio-demographic variables such as age of mother, ethnicity, wealth index, marital status, employment status, and level of education.

Statistical Analysis

Bayesian geo-additive logistic regression model

Let y_i denote a realisation from the random variable $Y_i = 1, \dots, n_i$. For our purpose, we define

$$y_i = \begin{cases} 1, & \text{if girl is cut} \\ 0, & \text{if girl is uncut.} \end{cases} \tag{1}$$

Then, Y_i , 0-14-year-old Nigerian girl FGM/C status, is Bernoulli random variable with parameters, π_i , that is, $Y \sim Ben(\pi_i)$. We used a class of mixed models called structural additive regression (STAR) models [5, 7, 10- 13], to estimate the effects of different covariates on the observed data. Unlike the standard regression model, which assumes strictly linear relationship between the covariates and the response variable, STAR models allow us to simultaneously control for both linear and non-linear, continuous and categorical covariates in a coherence regression framework, such that the link function η_i ,

$$\begin{aligned} \eta_i = \text{logit}(\pi_i) &= \text{log} \frac{\pi_i}{1 - \pi_i} \\ &= \beta_0 + \mathbf{z}'_i \boldsymbol{\beta} + f_1(x_{i1}) + \dots + f_p(x_{ip}) + f_{str}(s_i) + f_{unstr}(s_i) + f_{st}(s, t) \end{aligned} \tag{3}$$

for $s_i = 1, \dots, S, t = 1, \dots, T$. where $f_1(\cdot), \dots, f_p(\cdot)$ are the functions (may be smooth) of non-linear continuous covariates, x_i s such as age, time effects, etc. β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ are unknown coefficients of other class of covariates, \mathbf{z}_i s. Also, s_i is the geographically referenced location of girl i , $f_{str}(\cdot)$ and $f_{unstr}(\cdot)$ denote the structured (correlated) and the unstructured

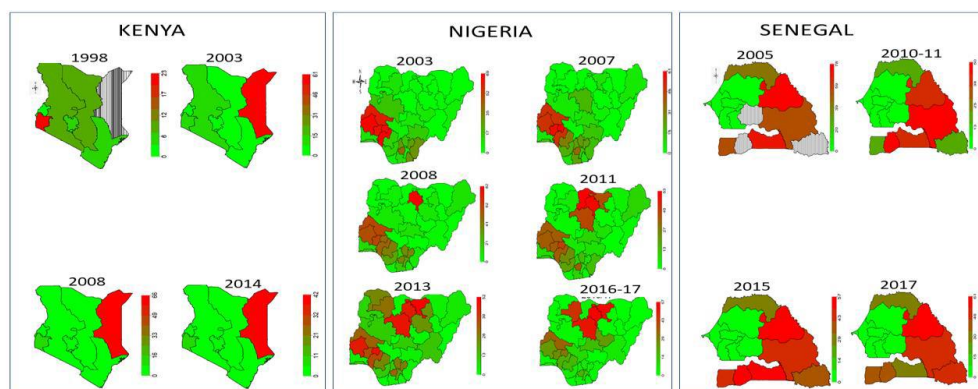
(uncorrelated) spatial effects, time main effect *year* as a smooth function, $f_y(\text{year})$ and an interaction term, $f_{st}(s, t)$.

Statistical analysis and inference were carried out in Bayesian framework via Markov Chain Monte Carlo (MCMC) techniques and implemented in R statistical programming software through its R interface to BayesX known as R2BayesX [8]. Model fit and complexity were tested using Deviance Information Criteria (DIC) proposed by [9], the smaller the better.

3. Results

Figure 1 shows the crude (observed) FGM/C prevalence across the survey years in Kenya, Nigeria and Senegal, indicating a clear picture of geographical variations in the practice. Red colour indicates highest prevalence regions or states, while green colour indicates lowest prevalence regions or states

Figure 1: Evolution of 0-14-year-old girl's FGM/C prevalence in Kenya, Nigeria and Senegal



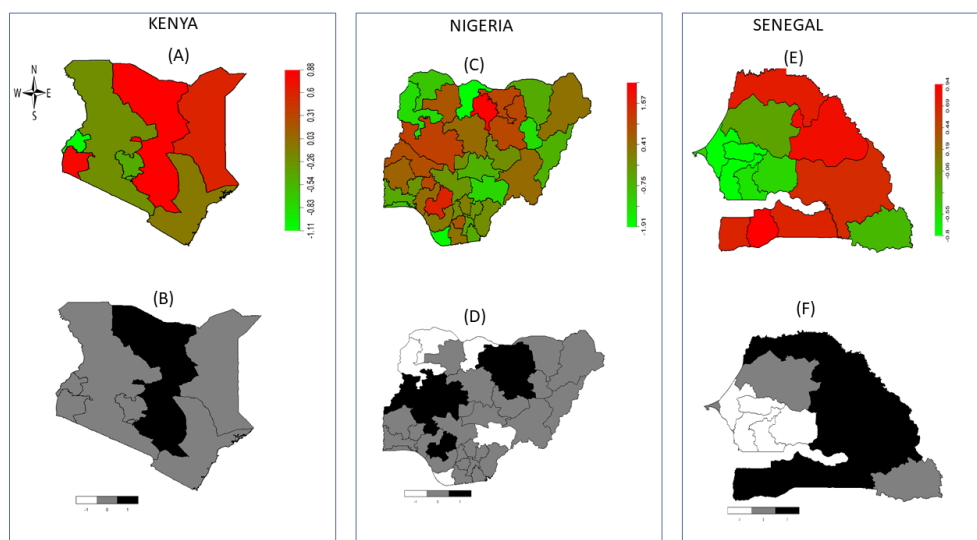
In Table 1, we present the posterior odds ratio (POR) from the fully adjusted model which accounted for other confounders including temporal, spatial and spatio-temporal effects. The results shows that in Kenya, girls who lived in urban region were more likely to be cut than their counterparts. However, in Nigeria and Senegal, rural girls had higher likelihood of being cut. The likelihood of FGM/C was not significantly influenced by household wealth index in Kenya and Senegal, however, girls from lowest wealth quintile households were more likely to be cut. A girl who professed Muslim faith had higher likelihood of FGM/C than her Christian counterpart in Kenya. Across the three countries, we found that a girl's likelihood of being cut increased if her mother was circumcised and if her mother her poor level of educational attainment.

In Figure 2, we show identified and mapped hotspots (red) across the three countries where the observed FGM/C prevalence were largely due to unobserved effects of geographical locations of the respondents.

Table 1. Fully adjusted posterior odds ratios (POR) and associated 95% credible intervals (CI) from Bayesian Geo-additive hierarchical logistic regression models, using the pooled datasets for Kenya, Nigeria and Senegal.

Predictor	Level	KENYA POR (95% CI)	NIGERIA POR (95% CI)	SENEGAL POR (95% CI)
Place of residence	Rural	1.00	1.00	1.00
	Urban	1.24 (0.96, 1.60)	0.96 (0.81, 1.16)	0.53 (0.44, 0.63)
	Middle	1.00	1.00	1.00
	Lowest	1.08 (0.85, 1.34)	1.2 (0.9, 1.55)	0.89 (0.74, 1.09)
	Second	0.89 (0.70, 1.14)	1.21 (0.95, 1.53)	0.95 (0.81, 1.14)
	Higher	0.82 (0.62, 1.12)	1.06 (0.88, 1.3)	1.09 (0.88, 1.36)
	Highest	0.59 (0.38, 0.86)	1.06 (0.87, 1.31)	0.89 (0.64, 1.28)
Marital status	Currently married	1.00	1.00	-
	Never married	0.67 (0.33, 1.33)	0.62 (0.29, 1.34)	-
	Formerly married	0.69 (0.52, 0.90)	1.45 (1.08, 1.98)	-
Religion	Christian	1.00	-	1.00
	Muslim	2.85 (1.99, 3.97)	-	1.14 (0.80, 1.63)
	Others	0.61 (0.36, 1.08)	-	0.56 (0.26, 1.16)
Mother cut	No	1.00	1.00	1.00
	Yes	28.30 (18.81, 43.48)	11.51 (9.66, 14.06)	26.51 (21.49, 33.61)
Circumcision should continue or be stopped.	Discontinued	1.00	1.00	1.00
	Continued	2.03 (1.51, 2.71)	14.31 (12.45, 16.34)	4.54 (4.09, 5.15)
	Depends/ Don't know	0.93 (0.43, 1.85)	3.17 (2.5, 3.93)	1.20 (0.94, 1.57)
Woman's education	Higher	1.00	1.00	1.00
	No education	3.93 (2.10, 7.16)	1.68 (1.19, 2.35)	2.48 (1.19, 5.09)
	Primary	2.48 (1.39, 4.44)	1.36 (1.06, 1.74)	1.90 (0.92, 4.04)
	Secondary	1.70 (0.92, 3.12)	1.37 (1.1, 1.77)	2.93 (1.37, 6.18)

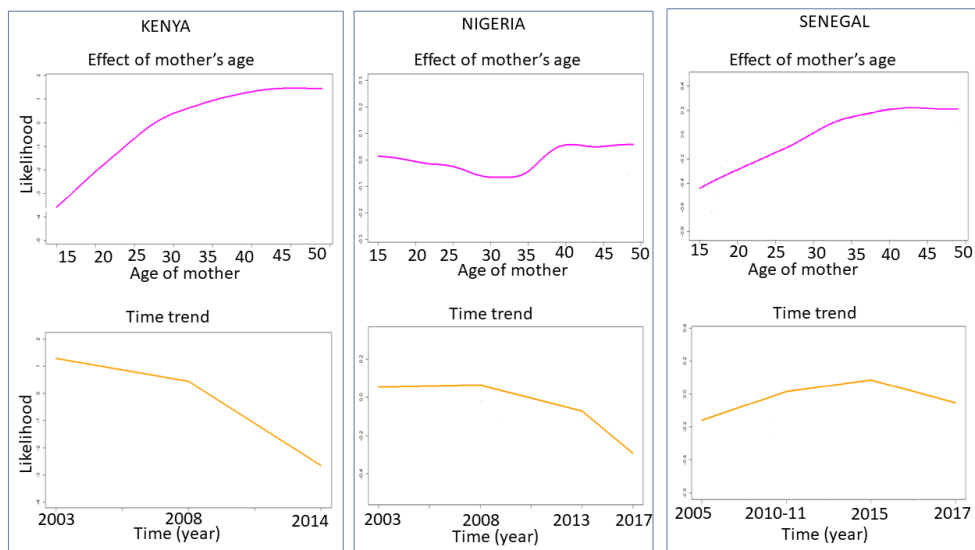
Figure 2: Total unobserved effects maps of spatial locations on the observed FGM/C prevalence in Kenya (A), Nigeria (C) and Senegal (E), with the corresponding posterior probability maps (B), (D) and (F).



Note. Red colour indicates FGM/C highest risk areas, while green colour indicates FGM/C low risk spatial locations. Black colour indicates significantly high risk area; white colour indicates significantly low risk area; grey indicates nonsignificant area.

Figure 3 (top) shows the non-linear effects of mother's age on her daughter's likelihood of undergoing FGM/C. It was found that in both Kenya and Senegal, a girl's likelihood of experiencing FGM/C increased with her mother's age unlike in Nigeria where there was no significant effect of mother's age on her daughter's likelihood of being cut. Figure 3 (bottom) also shows the temporal variation of FGM/C prevalence from the fully adjusted models across the three countries depicting clear picture of temporal variations and changes.

Figure 3: Effects of mother's age on the observed FGM/C prevalence among 0-14 years old girls in Kenya, Nigeria and Senegal, with the corresponding time trends.



4. Discussion and Conclusion

In this paper, we have used advanced statistical approaches to model and map the prevalence of female genital mutilation/cutting (FGM/C) among 0-14 years old girls in Kenya, Nigeria and Senegal using household-based data from demographic and health survey (DHS) and multiple indicators cluster survey (MICS). The approach allowed us to simultaneously account for the individual- and community-level factors that are key to the observed FGM/C prevalence in the selected countries. For each country, we analysed a single dataset obtained by pooling together all the available data from DHS and MICS surveys in order to investigate trend and identify patterns. Significantly higher likelihood of FGM/C associated with a girl who lived in urban region (Kenya); who lived in rural areas (Nigeria and Senegal); who professed Muslim faith (Kenya); whose mother was poorly educated; whose mother was circumcised or supported FGM/C continuation; and a girl who comes from the lowest wealth quintile household (Nigeria).

Finally, we found that rapid change towards abandonment of the practice has taken place in Kenya followed by Nigeria. However, we found no significant decline in the practice in Senegal across the survey years.

The findings in this study, were able to suggest at least to an extent, where and when changes are taking place in Kenya, Nigeria and Senegal. The characteristics of the identified hotspots may be exploited by policymakers and programme implementers in the design and evaluation of bespoke programmatic interventions aimed at achieving a sustainable stable abandonment.

References

1. WHO Eliminating Female Genital Mutilation: An Interagency Statement. (WHO, UNFPA, UNICEF, UNIFEM, UNHCHR, UNHCR, UNECA, UNESCO, UNDP, UNAIDS); World Health Organization: Geneva, 2008.
2. Muteshi JK, Miller S, Belizan JM (2016). The ongoing violence against women: Female Genital Mutilation/Cutting. *Reproductive Health*, 18;13:44.
3. Kandala, N-B, Komba, K, Nnanatu,C.C, Atilola,, et al. (2019). Modelling and Mapping od state disparities associated with FGM/C among girls aged 0-14.
4. UNICEF (2016). Female genital mutilation/cutting: A global a concern.
5. Utazi C.E., Afuecheta, A.O, and Nnanatu, C.C. (2018). Bayesian latent process spatiotemporal regression model for areal count data. *Spatial and Spatio-temporal Epidemiology*, 25, 25(37).
6. Kandala, N.-B.; Nwakeze, N.; Ngianga, S.; Kandala, I. I. (2009), Spatial distribution of female genital mutilation in Nigeria. *American Journal of Tropical Medicine and Hygiene* 81 (5), 784-792.
7. Achia, T. NO (2014). Spatially modelling and mapping of female genital mutilation in Kenya. *BMC public health*, 14:276
8. Umlauf, N., Adler, D. and Kneib, T (2015). Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software*, 63(21).
9. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002). Bayesian mea- sures of model complexity and fit. *J R Stat Soc Ser B (Stat Methodol)* ;64(4):583–639. doi: 10.1111/1467-9868.00353
10. Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18), 2555-2567
11. Brezger A, Lang. Generalized Structured Additive Regression Based on Bayesian P-Splines (2006)." *Computational Statistics & Data Analysis*, 50, 947-991.
12. Kneib T, Fahrmeir L. \Structured Additive Regression for Multicategorical Space-Time Data (2006): A Mixed Model Approach." *Biometrics*, 62, 109{118.
13. Kamman EE, Wand MP (2003). Geoaddivite Models." *Journal of the Royal Statistical Society C*, 52, 1-18.
14. Besag, J, York J, and Mollie (1991), A. Bayesian Image Restoration with two applications in spatial statistics, 43(1): 1-59



Spatial heterogeneity of childhood anaemia in four sub-Saharan African countries



Danielle Jade Roberts, Temesgen Zewotir
University of KwaZulu-Natal, Westville, South Africa.

Abstract

Childhood anaemia is a significant public health problem faced by many developing countries, particularly in Africa. It contributes to adverse health problems in children by affecting their cognitive and physical development, as well as their immune function which can lead to increased susceptibility to infections. The causes of anaemia are multifactorial and interrelate in a complex way. Such causes vary from country to country, as well as within a country. Thus, strategies for anaemia control should be tailored to local conditions and take into account the specific etiology and prevalence of anaemia in a given setting and sub-population. In addition, policies and programmes for anaemia control that do not account for the spatial heterogeneity of anaemia in children may result in certain sub-populations being excluded, therefore restricting the effectiveness of the programmes. This study investigated the demographic and socioeconomic determinants as well as the spatial variation of anaemia in children aged 6 to 59 months in Kenya, Malawi, Tanzania and Uganda. The study made use of data collected from nationally represented Malaria Indicator Surveys (MIS) and Demographic and Health Surveys (DHS) conducted in all four countries between 2015 and 2017. A Bayesian geoaddivitive model, which included a structured and unstructured spatial effect, was used. The study revealed distinct spatial variation in childhood anaemia across the four countries. However, the spatial variation was predominantly due to district-specific factors that do not transcend boundaries. These factors may include a lack of access to good health care and poor nutrition, among other local factors. Therefore, efforts in assessing the local district-specific causes of childhood anaemia within each country should be focused on.

Keywords

Adjusted odds ratio; fully Bayesian approach; Hierarchical geoaddivitive model; HemoCue haemoglobin concentration analyser; spatial effect

1. Introduction

Anaemia is defined as a significant reduction in haemoglobin (Hb) concentration which decreases the amount of oxygen reaching the tissues and organs of the body. Anaemia contributes to adverse health problems in

children, and affects their cognitive and physical development. In severe cases, these effects are irreversible. According to the most recent estimates of the World Health Organization (WHO), the highest anaemia prevalence of 42.6% in 2011 occurred in children under the age of five years old, which translated to just over 273 million children suffering from anaemia globally. In Africa, the prevalence of anaemia in children was estimated at 62.3% in 2011 (WHO, 2011). Despite the decrease in the prevalence of anaemia in high-income countries, anaemia remains a significant public health problem in many low and middle income countries, particularly in sub-Saharan Africa where anaemia is a major contributor to childhood morbidity and mortality (Abdo et al., 2018).

The causes of anaemia are multifactorial and interrelate in a complex way. Such causes include iron deficiency, other micronutrient deficiencies such as folate, vitamin B12 and vitamin A; intestinal parasites such as soil-transmitted helminths (STH) and *Schistosoma*; malaria, HIV infection, and chronic diseases such as sickle cell disease. Many of these factors contribute to the etiology and as well as the severity of anaemia through several mechanisms, either through the direct destruction of infected red blood cells and/or through the lack of the ability of the red blood cells to absorb iron. While iron deficiency is the most common cause of anaemia in developed countries, there are many other contributing factors in less developed countries.

Since 2012, the WHO advocates for global nutrition targets by 2025 with a comprehensive implementation plan on maternal, infant and young child nutrition, where the WHO strives for goals of achieving a 50% reduction of anaemia in women of reproductive age by 2025 (WHO, 2014). However, childhood anaemia has no such direct goals in place and thus has not received adequate attention. Nevertheless, the WHO and UNICEF have recommended that strategies for anaemia control be built into a country's primary health care system and existing programmes such as maternal and child health, integrated management of childhood illness, roll-back malaria, deworming (including routine anthelmintic control measures) and stop-tuberculosis (WHO and UNICEF, 2004). These control strategies are expected to be tailored to local conditions by taking into account the specific etiology and prevalence of anaemia in a given setting and population group. Accordingly, studies on anaemia control should be cognisant and account for the spatial variation of anaemia in the population. Failure to account for the spatial heterogeneity of anaemia and the possible causes of the spatial heterogeneity can cause ecological confounding (see Mainardi, 2012 and references therein).

This study investigates the spatial variation of anaemia in children aged 6 to 59 months as well as determines the significant risk factors associated with anaemia in these children in 4 sub-Saharan African countries jointly, namely Kenya, Malawi, Tanzania and Uganda.

2. Methodology

- *Study Area and Data*

This study uses data collected in the Demographic and Health Surveys (DHS) and Malaria Indicator Surveys (MIS) carried out in Kenya, Malawi, Tanzania and Uganda between 2015 and 2017, namely the 2015 Kenya Malaria Indicator Survey (KMIS2015), the 2017 Malawi Malaria Indicator Survey (MMIS2017), the 2015-2016 Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS2015) and the 2016 Uganda Demographic and Health Survey (UDHS2016). These four countries are situated on the east of sub-Saharan Africa and together form one contiguous region. The surveys were nationally represented and utilised a stratified two-stage cluster design. Three questionnaires, the household, women and men questionnaires, were carried out in the selected households. These questionnaires were designed to collect information regarding the characteristics of the household and eligible women and men. All children under the age of five years old in the selected households were tested for malaria and anaemia, with the consent of a parent or guardian.

- *Outcome Variable*

In all the surveys, a child's haemoglobin concentration was measured by finger- or heel-prick blood specimens using a portable HemoCue analyser. For this study, a binary response variable was used, indicating whether the child was anaemic if their altitude adjusted Hb level was less than 11 g/dL, or not anaemic if their altitude adjusted Hb level was greater than or equal to 11 g/dL.

- *Explanatory Variables*

The explanatory variables considered in this study comprised of a number of demographic, socioeconomic and environmental factors. These potential risk factors are shown in Figure 1. Such factors included the gender and age of the child, number of members in the household (size of the household), mother's highest education level, the child's malaria Rapid Diagnostic Test (RDT) result, type of place of residence: rural or urban; cluster altitude, household wealth index, type of toilet facility, and the age and gender of the head of the household. In addition, certain geospatial covariates were also considered. As no information regarding intestinal parasites was collected in the surveys used in this study, certain geospatial covariates were used as a proxy. Specifically, the cluster level average day land surface temperature (LST) and the cluster level average Enhanced Vegetation Index (EVI) for 2015.

Furthermore, the spatial variation of childhood anaemia across the administrative levels of the countries was investigated. The administrative levels of each of the countries were chosen based on the levels for which public health decisions are made. Accordingly, administrative level 1 (called "counties" or "districts") for Kenya, which consisted of all 47 counties; administrative level 2

(called “districts”) for Malawi, which consisted of 26 out of 28 districts for which data was available; administrative level 2 (called “districts”) for mainland Tanzania, which consisted of 176 out of 184 districts for which data was available; and administrative level 1 (called “districts”) for Uganda, which consisted of 121 out of 122 districts for which data is available; a total of 370 districts were considered.

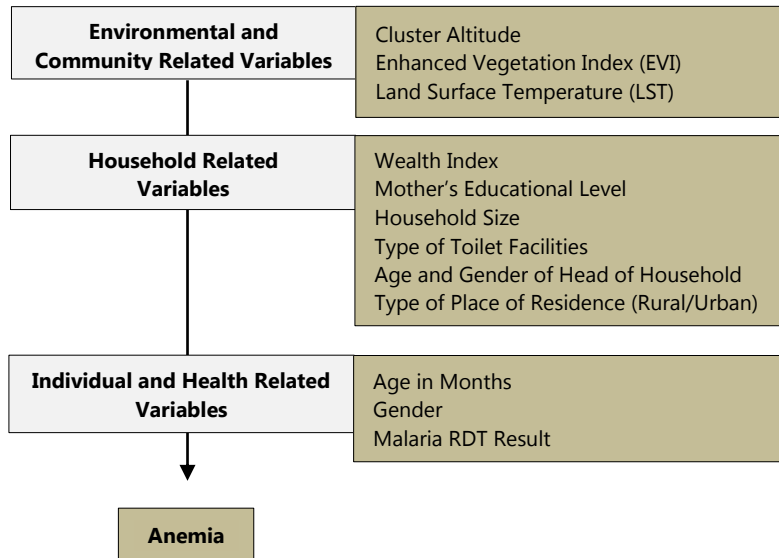


Figure 1: Conceptual framework for potential risk factors of anaemia among children (adapted from Ngnie-Teta et al., 2007)

• *Statistical Methods*

Univariate logistic regression was used to test for associations between each covariate and the child’s anaemia status. Covariates with associations that were significant at a 10% level were included in a hierarchical multivariable geoaddivitive model with a logit link function. A geoaddivitive model is a structured additive regression model that includes a spatial effect and is based on the generalised linear model (GLM) framework (Umlauf et al., 2015). For this study, Y_{hijk} follows a Bernoulli distribution where $P(Y_{hijk} = 1) = \pi_{hijk}$ is the probability that child k in household j within cluster i and district h is anaemic and $P(Y_{hijk} = 0) = 1 - \pi_{hijk}$ is the probability that the child is not anaemic. The hierarchical geoaddivitive model is given by

$$\text{logit} (\pi_{hijk}) = \mathbf{x}'_{hijk}\boldsymbol{\beta} + f_1(z_{hijk}) + \dots + f_p(z_{hijk}) + f_{spat}(z_{hijk}) \quad (1)$$

where the left side of Equation (1) is the logit link function and the right side is the geoaddivitive predictor. The parameter $\boldsymbol{\beta}$ is the vector of the linear fixed effects of the covariates that are modelled parametrically, and $f_r(\bullet)$, $r = 1, \dots, p$, are the unknown smooth functions that represent the non-linear effects of the continuous covariates which are modelled non-parametrically,

thus Equation (1) is a semi-parametric model. The spatial effect of district s_h in which the child resides, $s \in (1, \dots, 370)$, is given by $f_{spat}(s_h)$ which represents the effects of unobserved covariates that are not included in the model and also accounts for spatial autocorrelation (Kandala and Madise, 2004). This spatial effect may be partitioned into a spatially correlated (structured) and an uncorrelated (unstructured) effect as follows:

$$f_{spat}(s_h) = f_{str}(s_h) + f_{unstr}(s_h) \quad (2)$$

The structured spatial effect $f_{str}(s_h)$ accounts for the assumption that districts close in proximity would have similar observations. However, the unstructured spatial effect $f_{unstr}(s_h)$ accounts for the spatial variation due to effects of unmeasured local factors that are not spatially related.

In this study, inference was fully Bayesian, hence all parameters and functions were treated as random variables. The fixed effect parameters in θ were assigned vague Gaussian priors $(0, 1000)$, where the precision = $0.001 = 1/\text{variance}$. The Bayesian perspective of penalised splines (P-splines) was adopted for the unknown smooth functions f_r , where second-order random walk smoothness priors and third degrees splines were used (Lang and Brezger, 2004). For the structured spatial effect, $f_{str}(s_h)$, intrinsic Gaussian Markov random field (IGMRF) priors specified by Besag et al. (1991) were used. The unstructured spatial effect $f_{unstr}(s_h)$ was assigned i.i.d. Gaussian priors. The variance components of the random and spatial effects are unknown precision parameters that require estimation. Therefore, hyper-priors were assigned to them in a second stage of hierarchy. These hyper-priors are defined on a logarithmic scale and thus a log-gamma (1,0.001) distribution was used. A sum-to-zero constraint was imposed on the non-linear and spatial effects to ensure model identifiability between the intercept and these effects. Three types of models were fitted:

- Model 1: GLM model: Linear fixed effects of all variables, categorical and continuous.
- Model 2: GAM model: Linear fixed effects of categorical variables and some continuous variables, and non-linear effect of the child's age in months.
- Model 3: Geoadditive Model: Model 2 with the inclusion of the spatial effects.

The posterior distributions of the parameters in the models were estimated using Integrated Nested Laplace Approximation, and thus the INLA package in R was used (<http://www.r-inla.org/>) (Rue et al., 2009). The final geoadditive model was selected using the Deviance Information Criteria (DIC) and the effective number of parameters (p_D). QGIS 3.4 (<https://qgis.org/en/site/index.html>) was used to create maps displaying the

posterior mean estimates of the spatial effects for the different regions of the countries.

3. Results

Based on the univariate logistic regression with 10% level of significance for inclusion, the only independent variable not entered into the multivariable model was the age of the head of household. The variance inflation factor (VIF) was used to check for collinearity among the remaining continuous independent variables and all variables had a VIF < 4 and thus it was assumed that multicollinearity was not significantly present (Zuur et al., 2009). The non-linear effect of all continuous variables was investigated, however the only variable to display a significant non-linear effect on the log-odds of a child's anaemia status was their age in months. Thus, this was the only non-linear effect considered in the models fitted, while the remaining independent variables were included as linear fixed effects. Model 3 produced the lowest DIC, and thus the results of this study are based on this model, which includes both linear and non-linear effects as well as the spatial effects. In other words, the model given in Equation (1) is chosen and adopted.

Table 1 displays the adjusted posterior odds ratio estimates (AOR) with their 95% credible intervals for the linear fixed effects included in the final geosadditive model. Figure 2 displays the non-linear effect that a child's age in months has on the log-odds of being anaemic as well as the 95% credible band. There was an increase in the log-odds of anaemia from 6 to 10 months, after which the effect declined. Figure 3 displays the estimated means of the structured and unstructured spatial effects on the log-odds of anaemia, where the blue regions have a negative (or lower) spatial effect and the red regions have a positive (or higher) spatial effect, and thus are associated with an increased risk of anaemia. The structured spatial effect, which ranges from -0.0368 to 0.0316 , is weak in comparison to the unstructured spatial effect, which ranges from -1.3061 to 0.9780 . This suggests that the prominent driver of childhood anaemia in these countries consists of district-specific factors that are not spatially related and that do not transcend boundaries, such as a lack of access to good health care and poor nutrition.

Table 1: Adjusted posterior odds ratio estimates (AOR) and credible intervals

Variable	AOR	95% Cred. Interval
<i>Individual and Household Level</i>		
<i>Gender (ref = Male)</i>		
Female	0.873*	(0.818, 0.932)
<i>Malaria RDT Result (ref = Negative)</i>		
Positive	4.401*	(3.979, 4.871)
<i>Household Size</i>	1.019*	(1.008, 1.030)

Variable	AOR	95% Cred. Interval
<i>Type of Place of Residence (ref = Urban)</i>		
Rural	0.926	(0.835, 1.027)
<i>Mother's Education Level (ref = No Education)</i>		
Primary	0.857*	(0.773, 0.950)
Secondary and Higher	0.795*	(0.694, 0.911)
Unknown	0.845*	(0.742, 0.963)
<i>Gender of Household Head (ref = Male)</i>		
Female	1.003	(0.927, 1.086)
<i>Type of Toilet Facility (ref = No Facilities)</i>		
PIT Latrine	0.813*	(0.723, 0.914)
Flush Toilet	0.749*	(0.612, 0.916)
Other	0.711	(0.382, 1.325)
<i>Wealth Index</i>	0.858*	(0.807, 0.911)
Cluster Level		
<i>Cluster Altitude (in 100 metres)</i>		
	0.974*	(0.962, 0.987)
<i>EVI</i>	0.987	(0.927, 1.051)
<i>LST</i>	1.008	(0.994, 1.022)

*significant at 5% level of significance

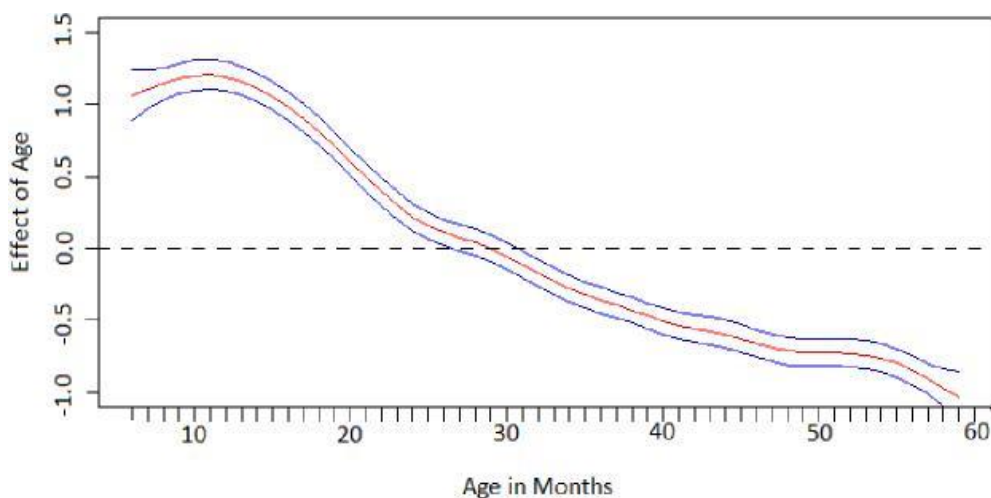


Figure 2: Estimated non-linear effect of child's age in months on the log- odds of anaemia. The posterior mean together with the 95% credible interval band are shown.

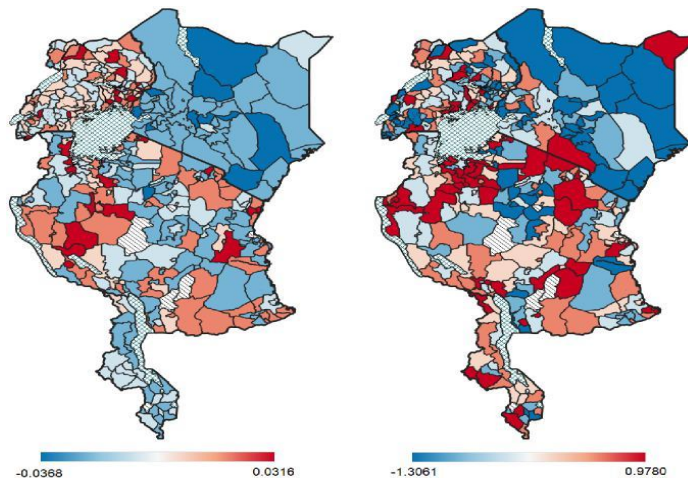


Figure 3: Estimated posterior means of the structured spatial effect (left) and the unstructured spatial effect (right) on the log-odds of anaemia (criss-cross pattern indicates water bodies; diagonal lines indicates districts with no data).

4. Discussion and Conclusion

Anaemia control measures need to account for the spatial heterogeneity that is evident in these countries, as well as take into consideration the potential factors and type of factors contributing to the spatial heterogeneity. Kenya and Malawi districts showed negative (low) structured spatial effects. On the hand, districts in Uganda and Tanzania displayed a mix of positive and negative structured and unstructured spatial effects, with the unstructured spatial effect being more prominent. Accordingly, efforts in assessing the local district-specific drivers of childhood anaemia within each country should be focused on.

References

1. Abdo, N., Douglas, S., Batieha, A., Khader, Y., Jaddou, H., et al. (2018). The prevalence and determinants of anaemia in Jordan. *East Mediterr Health J*, In press.
2. Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*, 343:1–20.
3. Kandala, N.-B. and Madise, N. (2004). The Spatial Epidemiology of Childhood Diseases in Malawi and Zambia. *African Population Studies*, Supplement B, 191–218.
4. Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *J Comput Graphical Statist*, 13:183–212. Mainardi, S. (2012) Modelling spatial heterogeneity and anisotropy: child anaemia, sanitation and basic infrastructure in sub-Saharan Africa, *Int J Geogr Inf Sci*, 26(3):387–411.

5. Ngnie-Teta, I., Receveur, O., and Kuate-Defo, B. (2007). Risk factors for moderate to severe anemia among children. *Food Nutr Bull*, 28(1):76–89.
6. Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J Royal Stat Soc*, 71(2):319– 392.
7. Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *J Stat Softw*, 63(21):1–46
8. WHO (2011). *The global prevalence of anaemia in 2011*. Geneva: World Health Organization. WHO (2014). Global nutrition targets 2025: Anaemia policy brief. https://apps.who.int/iris/bitstream/handle/10665/148556/WHO_NMH_NHD_14.4_eng.pdf;jsessionid=DC95FE26B76B825FA3407D855D3CBDC9?sequence=1 [Online; accessed November 2018].
9. WHO and UNICEF (2004). Focusing on anaemia. http://www.who.int/nutrition/publications/micronutrients/WHOandUNICEF_statement_anaemia/en [Online; accessed November 2018].
10. Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science, New York, USA.



Professional data scientists: Who are they and how do we train them?



Riaan de Jongh

North-West University, South Africa

Abstract

Much has been written on the divide between industry and academia, especially in the field of Statistics. This talk will propose some guidelines on how the gap between academia and industry may be bridged, both in teaching and research aspects. The guidelines will be illustrated by using a case study of a successful professional university programme.

Keywords

Training; Professional; Industry-University collaboration

1. What is Data Science?

Data science is a multi-disciplinary field consisting of a number of disciplines (e.g. applied mathematics, statistics, machine learning, operations research, artificial intelligence). It is used to solve problems in various application areas, for example health sciences, astrophysics, agriculture, telecommunications and finance. Its primary aim is to extract insight from data in various forms, both structured and unstructured. At the core is so-called “big data” that are stored in various ways and exhibit complex many-to-many relationships, which is made more challenging by the ever-increasing requirement to process these in real time to support “instantaneous” decision-making. The rise of data science can largely be attributed to advances in computer technology and processing speed, low cost storage of data, and the massive availability of data from the Internet and other sources. The access to big data and the advances in computer technology make possible the renewed application of machine learning and statistical techniques on problems, reporting huge successes in a wide range of applications. One of the subfields of data science is statistics, a branch of mathematics dealing with the collection, analysis and interpretation of data. Statistics have established itself firmly as an academic discipline and has been in existence since the eighteenth century. Because of the big data explosion, a number of the classical statistical approaches that perform reasonably well for small datasets fail when dealing with huge datasets. Despite this, many more recently developed statistical techniques are used successfully in a big data context. Examples are logistic regression, cluster analysis, and decision trees. Machine learning and artificial intelligence are relatively new subfields of data science

and concentrate on brute force computer power and complex optimisation algorithms to solve real-time prediction problems. Examples are the successful applications of neural networks and deep learning in the area of speech recognition and language processing (e.g. Siri and Google Assistant). It should be noted that machine learning is frequently concerned with prediction tasks and models in this context (e.g. recommender systems and factorisation machines). Unlike statistics, machine learning is not concerned with traditional aspects of statistical inference (e.g. about the significance of the estimates of model parameters). Although statistics and machine learning are different disciplines, there is some overlap, for example a technique like random forests are frequently quoted in both fields. A reader of the literature in both fields will quickly realise a difference in the terminology used for similar concepts. It is interesting to note that in a recent paper, a prominent researcher at Harvard University (Meng, 2018) warned about the big data paradox, i.e. he emphasised that data quality plays an enormous role and that having more data will fool us when making population inferences in a big data context.

Another subfield of data science is operations research (OR), which became popular in the early eighties. Spurned by the advent and wider availability of personal computers, OR, like data science now, was all about using the mathematical and computing sciences to solve real-world problems in a multi- and interdisciplinary way.

2. What is a data scientist?

Because data science is so wide in scope, many professionals may claim that they are data scientists, e.g. statisticians, operations researchers, engineers, computer scientists, actuaries, physicists and machine learners. From my own practical experience, it is clear that when solving data science problems, you need a range of people of which some can work in depth on theory and others can tend to application. It is a way to attempt to cover the whole spectrum. When solving complex problems in data science, one person cannot handle all aspects, but it could possibly be achieved with a group of people. Currently the main focus of data scientists is to use innovative techniques emanating from the subfields to solve problems in a particular application area of interest. It should be noted that the application areas and therefore the type of problems encountered are very different, frequently necessitating a deep knowledge of the particular subject matter. For example, consider astrophysics and the squared kilometre array. Apparently these telescopes will receive data at one terabyte per second and researchers are typically interested in detecting tiny signals engulfed in white noise. On the other hand, in finance, amongst others, researchers exploit large data bases to learn more about the credit behaviour of customers.

3. What can we learn from the established subfields of data science?

As stated before, statistics and operations research are two of the oldest subfields in data science and many practicing statisticians and operations researchers consider themselves data scientists. What can we learn from these fields that could help us in training the data scientists of the future? Universities all over the world have largely failed to deliver *professionally* trained graduates in the fields of OR and statistics. Although well trained academically, many newly appointed graduates find it difficult to immediately add value at their place of employment. Typically they lack subject matter knowledge of the application field (e.g. finance or physics) and struggle with real-world problem solving abilities, such as the formulation of messy problems, meaningful interaction with clients, interrelationships with team members and business communication. Some students also lack numerical and data handling programming skills that are not addressed adequately in many curricula.

Some of the lessons I learnt in the many industry projects I have been involved in, include:

- Always focus on the business value throughout the course of the project.
- Involve all role-players and instil trust and confidence about your ability as a consultant.
- Manage the client's expectations, communicate clearly and pay attention to fostering good interpersonal relationship skills.
- Test the client's understanding of his/her own problem and educate the client when necessary.
- Be sure that the problem to be solved is well formulated, because you do not want to solve the wrong problem.
- Always be cognisant of the importance of simplicity and when your solution is very complicated, seek a simpler solution, if possible.
- Do not be fixated on new untested technologies.
- Solve the critical aspects that will determine eventual success, first.
- Always revisit the scope and risks of the project and plan properly.

How do we, however, train students to ensure that they become professional data scientists? This is not easy and will be addressed in the last section.

4. What should we teach aspiring data scientists?

From the above it should be clear that a training programme should include training in the following:

- *The mathematical and computational sciences:* Topics could include courses in statistical and probability theory, artificial intelligence, machine learning, operations research, and computer science.

- *Programming skills:* Numerical programming skills in languages such as SAS, R and Python.
- *Data management skills:* Topics should include data bases and warehousing that concentrate data manipulation and merging skills in languages such as SQL, SAS, R and Python.
- *Subject matter knowledge* in selected fields of application.
- *Professional problem-solving skills.*

Assuming a sound knowledge of undergraduate training in the mathematical and computer sciences, one could include the following topics in a graduate programme: generalised additive models; regularisation (lasso and elastic nets); model selection; time series analysis; multi-variate statistics; cluster analysis; optimisation; neural networks and deep learning; support vector and factorisation machines; event stream processing; text analytics; database handling and extraction. All of these courses should have a practical element, where the techniques are programmed in one of the above-mentioned programming languages and applied to data and problems in a relevant application. Depending on the application areas, suitable courses on the important concepts in these fields should be included. For example, in astrophysics, it might be necessary to include courses such as signal processing and pattern recognition and basic concepts in astrophysics. Similarly, if the application area is finance, courses could include scorecard model building, risk management and other important financial concepts (e.g. value-at-risk). It is of course, not practical to cater for all the fields and possible topics, if not impossible. At my university we have spread the programme over two years, where all the technical courses are covered in an honours degree and half of the masters' degree. The remainder of the masters' programme addresses the professional training aspects.

5. Adding professionalism to the training programme

Teaching students the problem-solving skills necessary for the industry is a real challenge. The instructor should facilitate a mind set change among students to ensure they focus on the importance of solving the business problem and not a statistical or mathematical sub-problem. More importantly, these courses should be taught by people with the necessary experience in solving problems in the particular application area (see e.g. Coetzer & de Jongh, 2016). This suggests that data science programmes comprising only academics with no experience in solving industrial or business problems will make it extremely difficult to equip data scientists with the requisite skills to function effectively in industry.

In our Masters programme we follow an integrated hands-on training approach in solving problems in the area of application. This is done in the form of on-site (at the client company) internships where a student is assigned

to a specific problem posed by industry. The student has to complete the project over a six-month period with the assistance of an academic supervisor (responsible for academic quality) and a client project officer (responsible for business value add). Formal on-site project meetings are scheduled where all role players should be present to discuss progress. In this way the academic supervisors also gain industry experience and get a feel for the problems being faced by industry. This often leads to industry directed research projects by the supervisor for the company. The company has the benefit of screening the student for employment and potential problem-solving value add at relatively low cost. The demand for these students has increased dramatically, supported by the fact that project proposals outnumber available students 2:1. As a fringe benefit the programme has spurred a number of research imperatives between academia and industry and many papers have already been submitted, which have been co-authored by academics and practitioners. It should be noted that although student projects are classified confidential and although students have signed non-disclosure agreements with the assigned company, it is amazing how quickly client project officers can share sensitive information when they are offered co-authorship of a paper. Interestingly, but not surprisingly, alumni of this programme become future client project leaders and research collaborators.

Please see the references for more information about this programme.

References

1. Coetzer, R.J.L. & De Jongh, P.J. 2016. Discussion of 'industrial statistics: the challenges and the research'. *Quality engineering*, 18(1):63-68.
2. De Jongh P.J. & Erasmus, C.M. 2014. Industry-directed training and research programmes: the BMI experience. *S African journal of science*, 110(11/12), Art. #2013-0392, 8 pages. <http://dx.doi.org/10.1590/sajs.2014/20130392>
3. De Jongh, P.J. 2018. University-industry engagement in data science. <https://www.youtube.com/watch?v=QgsoRUuLhU> [Video].
4. Meng, 2018. Statistical paradises and paradoxes in big data. *The annals of applied statistics*, 12(2):685-726.



Advanced data collection – An outlook to the future



Irene Salemink

Statistics Netherlands, Heerlen, The Netherlands

Abstract

The demand from the society for data-driven fact-based information continues to rise. Technological possibilities and the increase of available and usable data offers possibilities to produce this fact-based information. Furthermore, conventional methods, using primary data collection, would increase the burden to the society and would be too time-consuming and costly to satisfy that demand. The use of administrative data and sensor data is a logical step towards the future. Statistical methods, legal frameworks and technology are being developed to maximize the added value of these data sources. Consequently, the nature of primary data collection for official statistics is bound to change. This paper gives an outlook to and a pathway towards this future of hybrid data collection.

Keywords

Administrative data; Hybrid data collection; Adaptive survey design; Integration by design; Metadata

1. Introduction

In general, the mission of National Statistical Offices (NSO's) is seen as provider of trusted official statistics often based on a mandatory program that consists of a set of consensus indicators describing economic, social and demographic phenomena in society. The awareness is growing that often these phenomena are so complex that using a single indicator or a limited set of indicators is not providing enough or accurate information anymore. Therefore, alternative ways are being developed to provide insight in these complex societal phenomena. Statistics Netherlands (CBS) has formulated this in its mission as; providing insight in complex societal phenomena by delivering "actionable intelligence" to enable evidence-based policy and decision-making. With the objective of continuous quantitative monitoring of developments and progress with the required aggregation level and timeliness.

The increasing demand from the society for data-driven fact-based information in combination with the fast evolving technological developments causes great challenges for statisticians. At the same time they have to deal with the reality of the ever-increasing difficulties to retrieve statistical data with the traditional approach based on surveys. As very well illustrated in the essay by

Jarmin (2019). The Netherlands show a similar trend compared to the international decrease in willingness to participate in social surveys, and the necessity of burden reduction for the business statistics. New methods and improvements of the data collection strategy are required, giving rise to a major change in one of the most important parts of the production process: data collection. As extensively deliberated by Groves (2017) and Bean (2016) there is broad agreement on the need to transform from a survey-centric model to a model that blends structured survey data with administrative and unstructured alternative data sources. The use of administrative data is characterized by its versatile applications. This paper deals with these various aspects of the use of administrative data in the production of official social and business statistics at CBS. In the following paragraphs, examples are given for its use in the replacement of survey data (2.1), facilitating the creation of new information products and more detailed statistics (2.2), better approach of target populations (2.3) and in establishing and improving population backbones (2.4). To make the most of it new methods on combining data need to be explored (2.5) as well as new technological developments (2.6). The increasing use of administrative data also creates new job types (2.7) and influences the organisation of NSI's itself in relation to data owners (2.8).

2. Aspects concerning the use of administrative data

2.1 Administrative data in official statistics

Since the mid-1990s, wherever possible CBS has switched from the system of conducting sample surveys to assembling data from the administrative sources of other public institutions and using the existing data in those registers rather than collecting the data it needed by distributing questionnaires. The Dutch census being a fine example of a population census for which no additional data collection is needed (0 enumerators). By bringing together all available data sources, combining registers and surveys a virtual census is conducted (<https://youtu.be/SLpDkcyenf0>).

The Statistics Netherlands Act required CBS to minimise the administrative burden it caused and at the same time provided for the use of government registers by CBS for statistical purposes. A second impulse on increased availability and use of government registers was in the year 2000 when the Dutch government launched the project "Streamlining Administrative Data" in which the infrastructure was developed to store administrative information in a network of so called basic registers. This administrative system of basic registers is part of the Generic Data Infrastructure (GDI) and is governed by the National Commissioner for Digital Government (<https://www.nldigitalgovernment.nl>). This network aims for better governmental service to the public and business sector. The basic idea is that citizens and businesses provide only once their personal information to a governmental body. Each basic register keeps its own

authentic data to fulfil its primary tasks and is supposed to share this information with the other registers in the network. To enable this sharing and re-use of information the network holds unique linking keys by which mutual relations between the datasets can be established. Those keys and relations provide interesting opportunities for CBS. Although CBS also has access to the single sources (by law), it is the mutual relations and for example the linking to the Statistical Business Register, that determines the real value of this system for statistical production.

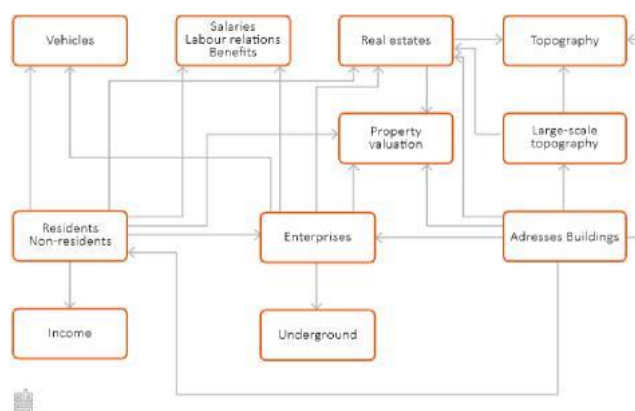


Figure 1: Basic registers and relations between them in the system of basic administrations

The Trade Register (TR, owned by the Dutch Chamber Of Commerce, CoC, part of the Department of Ministry of Economic Affairs, in Fig. 1 represented by the icon “enterprises) for example has relations with the municipal basic administration (inhabitants per municipality) via the social security number of natural persons. A link with the basic administration of “Addresses of Buildings” gives access to unique addresses and linking to “Salaries Labour relations Benefits” enriches data sets with information on Wages, Labour- and social security/welfare rates. Another unique identification number links the Trade Register data also to the registers of the tax authorities. In addition to government registers, in the last few years CBS has also started using business records, for example, scanner data from supermarket checkouts to calculate the Consumer Price Index and information concerning energy consumption from energy companies. Nowadays it is impossible to imagine the production of official statistics without the use of administrative data: for 17 statistical themes, we contact over 75 holders of registrations to obtain over 200 data sources to be used in official statistics production.

2.2 Product development

In addition to using administrative data in official statistics, these data also prove to be very valuable in combination with data from sample surveys and registrations. The administrative registers often contain far more detailed information than sample surveys can provide, generating new information. Combining data from a variety of registers also creates new possibilities. Hence new, more detailed and up to date statistical information is made available. For example, in recent years in the Netherlands, real estate drew a lot of attention, especially empty shops and offices. The National Monitor on Disused premises gives a complete view on all real estate (houses, offices and shops) at municipality level and for houses even on a district and local area level. Based on several registers like the Addresses of buildings, Valuation for tax purposes, Resident's registration and the Trade Register it was determined whether or not a property is disused. After a first national publication in 2017, the method was further customized by determining whether there was a matter of energy use in empty buildings as indicated by data from energy companies and reality checks were carried out to determine the relation between actual and administrative disuse. Our users consider the figures appropriate to monitor policy concerning property disuse. Another example is the combining of data on ownership of vehicles, characteristics of vehicles, drivers' licenses and travelled distances with characteristics of persons and households, to satisfy the needs on information on traffic and mobility related to trends in society, as well as to develop more regional data and information on specific population groups. On the international level, to get a better view on cross boundary payments CBS is investigating which data on payment transactions could be useful and is available at banks.

2.3 Adaptive survey design

The nature of primary data collection for official statistics is bound to change. In response to budget pressure due to gradual but persistent declines of response rates, designs like adaptive and responsive survey design have received a lot of interest over the last decade e.g. Chun (2018). During the past years, many surveys at CBS were redesigned to reduce cost and to increase or maintain response rates where also alternative approaches like adaptive survey design is investigated. Adaptive survey design assumes that differentiation of effort over relevant population subgroups is either effective in improving survey quality or efficient in reducing survey costs. Currently, adaptive survey design is a standard option in redesigns of persons and household surveys at CBS and was implemented in the Dutch Health Survey, e.g. Berkel et al. (2018). How does adaptive survey design relate to the use of administrative data? Adaptive survey designs have four main elements: quality and cost objectives and metrics, stratification of the target population, design features, and an

optimization and implementation strategy. See Schouten et al. (2017) and Tourangeau et al. (2017). It is within the domain of the stratification of the target population that administrative data prove their value added in improving survey designs. Determining target groups, also called segmentation or clustering of the target population, is done with a classification tree. People are divided into groups based on personal characteristics. Within the Dutch Health Survey, demographic and regional characteristics have been used that are known to have a different response distribution than the population. Examples are ethnicity, ethnicity of parents, age, income, urban character of the neighbourhood or municipality, education, household-type and size, marital status, wealth, gender, and home ownership. The strata were based on administrative variables that are used in post-survey adjustments.

2.4 Improving Statistical Business Register backbone function

Next to improving survey design for social statistics, administrative data also prove their value in the production of business demography and other statistics. By linking administrative data to administrative and statistical units stored in the Statistical Business Register (SBR) data sets can be enriched, enterprises can be characterised and sub populations can be determined. For example, Family Businesses (FB) are recognized to play an important role in economies of the member states of the European Union (EU). FB make up between 65 to 80% of all European companies, they make a significant contribution to Europe's GNP and employment (40 to 50% of all jobs), and tend to be great innovators, with a longer-term vision and specific commitment to local communities (<http://www.europeanfamilybusinesses.eu>). In order to measure the importance of FB's their performance and characteristics separating them from other kind of businesses they need to be characterized within a SBR. As described by Konen (2017) CBS identified FB's without sampling and surveying but by using information from the SBR and administrative registrations (Trade register, Payroll Tax register, management of relations of tax authorities, satellite of Self-employed Entrepreneurs, household register, alliance register and Child parent register). The research on detecting Family Businesses in the SBR fits into a broader field of research on 'profiling' enterprises thereby differentiating businesses based on certain characteristics. In addition, policy makers show interest in different typologies for Small and Medium Enterprises. Besides Family Businesses and the Self Employed, there is interest for Hidden Champions, Almost Failed Firms, Ambitious Entrepreneurs, (Un)-Consciously Constraint Entrepreneurs and Corporate Social Responsibility. These "sub-populations" can only be derived by combining the SBR with a multitude of various data sources (registers, administrative data, internet data etc.). At the Register Department of CBS

methodologies are developed to use several (new) sources to distinguish for example enterprises involved in internet economy, e.g. Ostrom et al. (2016). In order to be useful for statistical purposes, as for many other Big Data investigations, the data needs to be linked to statistical information. In this case, the characteristics of websites needed to be linked to the businesses behind the website. Therefore two key pieces of information were used; the websites as recorded in the SBR, and the businesses' CoC-registration number as published on the website. These identifiers provide the basis upon which websites can be linked to the respective businesses. Subsequently, when successfully linked, the SBR facilities further links to a variety of data sources available at CBS (see 2.2). Obviously, the future lies in extensively linking a multitude on data sources where the international dimension in characterizing enterprises (also small and medium) is unabated important. As Timothy Sturgeon (2013) stated: "Clearly, the assumptions behind current data regimes have changed and statistical systems are struggling to catch up. While it will be exceedingly difficult to fill data gaps without new data, and progress that relies only on existing data resources will always be limited, the most efficient approach will be to develop systematic links between key existing data, supplemented with a few additional variables, with data on enterprise characteristics drawn from administrative sources, all tied together by enterprise identifiers that make ownership clear, even when it extends across borders."

2.5 Methodology on combining data

In our ambition to increase our statistical output an important prerequisite for CBS is to make as much and diverse as possible data available. CBS does this by combining administrative data from registers, registrations, Big Data (i.e. sensor data), private data and survey data. The adequate combination of sources can be decisive regarding the outcome, implying that approach and way of working need to be adjusted. In fact, when combining multiple data sources from multiple modes the challenge is to develop methodology that helps to deal with issues concerning matching of data sources. Specific issues that can occur when matching various sources are; units to be matched do not equal source units (persons, businesses), sources do not contain overlapping units however one wants to estimate the correlation between variables occurring in both sources, matching errors resulting in bias of an estimator that one wants to correct and (assuring) the coherence between statistics. General techniques like probabilistic matching, matching with supervised machine learning and synthetic matching are extended and/or combined to solve these issues. In addition, combining registers and survey data comes with its specific challenges; variables can occur in multiple sources with different measurement errors for which methods are developed to come to consistent estimates. When

missing data needs to be imputed than preferably in such a way that the outcome is consistent with previous published outcomes and when similar aggregates are published in various tables one strives for micro-meso-macro consistency. The real challenge and paradigm shift will be the approach where in advance to making the observation and survey design, the information already available (registers, registrations, Big Data, etc.) is taken into account: integration by design. Instead of enriching surveys with administrative data the (combined), various administrative and Big Data sets are completed with survey data (only when necessary). In addition, the challenge can even be take one step further. What about the knowledge we have on events that have taken place and are known to us before we have collected the data? Event driven processing opens up a completely new area of possibilities and challenges.

2.6 Technology

For researchers there is the need to get easier and faster access to more data with better tools. The reality is that datasets become too big to copy, are not allowed to “leave the building”, need matching between multiple sources, require knowledge on the source and its metadata, etc. How to deal with these different challenges may be dependent on the type of “data sharing”. In its data architecture, CBS defined four patterns; 1) External data is coming to CBS to be matched with CBS data, 2) CBS data is brought to an allocated environment to be matched with other data, 3) Both external and CBS data are kept on premise, data is matched by virtual connecting and 4) The algorithm (and data) is send to the other data source to execute the matching. These patterns come with certain (distinct) capabilities that need further investigation. For example, Privacy Preserving Analytics enables analysis of privacy sensitive data of various sources without the risk to look into each other’s micro data but with the outcome of new statistics and insights. It is also possible to apply special encryption methods to enable various parties to execute computations with each other’s privacy sensitive data (secure multi party computing). The content of the records stays hidden and no third party is needed to keep the encryption keys. Often datasets consist of different formats and size, which makes it impracticable or undesirable to copy this data physically to one location. The technique of data virtualisation then offers the possibility to simplify access to data regardless where and how the data is stored. An important prerequisite to almost all techniques is presence of meta data (conceptual, technical, process, origin). In fact, meta data is the hub of the action and forms the basis for seeking and finding data. In order to help users find the right data, understand its semantic characteristics and use this information in data integration, data analyses and other statistical activities, CBS developed a meta data model. The model is based on a graph representation of characteristics that describe statistical datasets as well as relationships between datasets.

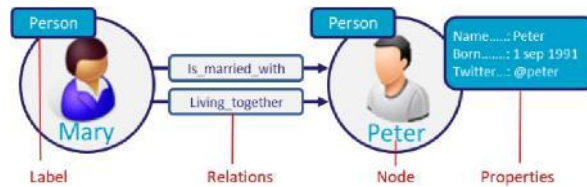


Figure 2: Example of the meta data model as a semantic representation of objects

Data lineage in turn makes it possible to trace which variables are used where and in what outputs. When quality issues occur, it is possible to identify the source of the error and which outputs are affected by it. A hot issue in a time where apparently everybody seems to have access to massive amounts of data are the fact vs fake discussions. Making it more and more necessary to file and save processing steps in permanent and invariable data sources. Block chain technology could be of great value for example in supporting third parties to verify whether the data really originates from an NSI and the quality of the processing steps. A process that starts at data collection and stops with the dissemination.

2.7 Data scouting

Although registers and registrations may be all around, it is not self-evident that CBS can have the data at its disposal. Firstly, it must be known that a data source is available and secondly it must be determined whether this data source is useful in making official statistics (whether or not combined with other data sources). It is at this point of interchange between hard skills on data handling and soft skills concerning relation management that CBS introduced the position of Data scout. The aim of our Data scouts is to organize and facilitate the acquisition and opening up of new (big) data sources in order to create new, or improve/enrich existing statistics. Thereby a Data scout is the linking pin between both internal and external stakeholders. Its portfolio comprises a wide range of tasks varying between identifying the CBS data requirements together with content experts, mapping possibilities of relevant new data sources, evaluating and testing possible new data sources on their usability for the intended purposes, working together with legal, technical and domain experts to discuss the (im) possibilities, building new relationships with relevant partner organizations, negotiating terms and conditions for data use, defining (joint) business models and making agreements with data owners. Challenges that our Data scouts encounter include the following areas: the need to anticipate continuously on future possibilities (knowing possible applications before knowing the data completely), long project lead-times, setting up joint business models with private corporations, dependence on external parties, issues

concerning data sharing (Safety, access, storage, etc.) and various legal issues (privacy, GDPR, collaboration with private partners etc.).

2.8 Data ecosystems

For a long time, CBS gathered the information required to produce statistics by distributing questionnaires to businesses, citizens and the government. Nowadays, most of the data come from registers and new data sources, which make it possible to produce far more intricate statistics that are also more identifiable. To meet the public demand for statistics, CBS is endeavouring to further expand its role as data hub and producer of statistics from and for the entire government. Therefore, CBS is working on so-called data ecosystems where CBS fulfils the role of a platform, enabling a broad cooperation between municipalities, scientific institutes, governmental bodies and businesses. The aim of a data ecosystem is to create an environment in which the cooperation between CBS and (decentral) governments results in innovative clusters of innovative enterprises and institutes making use of the rich data infrastructure that CBS has to offer. Ensuring essential guarantees like privacy, quality and consistency.

3. Discussion and Conclusion

Although surveying is still indispensable (e.g. for large and complex Enterprises due to globalisation, or measurement of specific behaviour and sentiments), CBS' influence on the content of the data that is captured is steadily diminishing. The technological possibilities increasingly determine what data are collected, and producing statistical information from these new sources is becoming more complex in conceptual and methodological terms. Validation of the results and measuring errors and biases are a big challenge for the future. Modern developments also create tremendous opportunities and by combining various sources, greater possibilities occur to produce new, more detailed statistics. The future of data collection will undoubtedly be "Advanced" and build upon these new developing fundamentals. Thereby having impact on multiple dimensions: Monetary (reduce costs and response burden, increase revenue and efficiency), Culture (change the way of thinking, operating and presenting outcome), Policymaking (rational, fact driven and evidence based, empowering local government) and Technical (data sharing across public and private organizations). Collecting through connecting.

References

1. Jarmin, R.S. (2019) Evolving measurement for an evolving economy: thoughts on 21st Century US economic Statistics. *Journal of Economic Perspectives*, 33 (1), 165 – 184.
2. Groves, R.M., Harris-Kojetin, B.A. (2017) National Academies of Sciences, Engineering, and Medicine. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/24652>.
3. Groves, R.M., Harris-Kojetin, B.A. (2017) National Academies of Sciences, Engineering, and Medicine. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/24893>.
4. Bean, C., (2016) Independent review of UK economic Statistics.
5. Chun, A.Y., Heeringa, S.G., Schouten, B. (2018) Responsive and adaptive design for survey optimization, *Journal of Official Statistics*, 34 (3), 581 – 597.
6. Berkel, van K., Doef, van der S., Schouten, B. (2018) Implementing adaptive survey design with an application to the Dutch Health Survey. *Journal of Official Statistics* (submitted).
7. Schouten, B., Peytchev, A., Wagner, J., (2017) *Adaptive Survey Design*, Series on Statistics Handbooks, Chapman and Hall/CRC.
8. Tourangeau, R., Brick, M., Lohr, S., Li, J. (2017) Adaptive and responsive survey designs: a review and assessment, *Journal of the Royal Statistical Society A*, 180 (1), 203 – 223.
9. Konen, R. (2017) Family businesses in the Netherlands, Meeting of the Group of Experts on Business Registers – jointly organised by UNECE, Eurostat and OECD, Paris.
10. Ostrom, L., Walker, A.N., Staats, B., Slootbeek-van Laar, M., Ortega Azurduy, S., Rooijackers, B. (2016) Measuring the internet economy in The Netherlands: a big data analysis, available at <https://www.cbs.nl/nl-nl/achtergrond/2016/41/measuring-the-internet-economy-in-the-netherlands>
11. Sturgeon, T.J. (2013) *Global Value Chains and Economic Globalization- Towards a new measurement framework*, Industrial Performance Center MIT, see also www.globalvaluechains.eu
12. Many thanks to: Barry Coenen, Paul Grooten, Florian Henning, Leanne Houben, Rico Konen, Johan Lammers, Annemieke Luiten, Marcel van der Steen and Jeroen van Velzen for their review, comments and suggestions.



Modernizing data collection in Canada

Stéphane Dufour, Geoff Bowlby, François Laflamme, Sylvie Bonhomme, Holly Mullin, Etienne Saint-Pierre, Fred Barzyk, Sevgui Erman
 Statistics Canada, Ottawa, Canada

Abstract

As traditional methods to collect data from households are becoming less effective and more costly, new innovative approaches are emerging and must be considered. Around the world, traditional primary data collection methods are becoming less effective and require more effort to achieve satisfactory results for household surveys. Technological and cultural changes have increased collection costs, as establishing contact with respondents and gaining their co-operation now require more effort. As a result, response rates for many surveys are trending downward. Finding new innovative ways of collecting the data necessary to create insights is very important for national statistical offices if they want to remain relevant. Statistics Canada has recently increased its emphasis on researching and introducing such innovative collection methods for household surveys. As a result, response rates have stabilized and costs have been managed effectively over the last few years. The first part of the paper will describe the initiatives that successfully contributed to alleviating the downward trend in response rates. However, continued research is required on new data collection methods and techniques, as the downward trend in response rates could return, along with resulting cost increases to limit it. As a result, Statistics Canada is researching more advanced approaches, which might change its primary data collection more dramatically by complementing or replacing traditional collection. The next steps are thought to lead towards completely new data collection techniques, such as sensor and scanner use, crowdsourcing, web scraping, automated voice interface use, and other innovative methods. The second part of the paper will describe some of the experiments, risks and opportunities that are being considered at Statistics Canada. It will also provide suggestions to identify and consider even more innovative and modern approaches.

Keywords

Costs; innovative; experiments; timeliness; relevance

1. Introduction

Statistics Canada, like many statistical organizations throughout the world, has observed a downward trend in household survey response rates. Changes

in the external environment (e.g., more cellphone-only households) and changes in respondent behaviour and their communication preferences have led to this steady decrease. Statistical organizations are asking: what types of initiatives can improve response rates? Statistics Canada has responded in two different phases.

The first phase, almost completed, focused on better managing current collection approaches. This phase saw most effort devoted to developing an electronic questionnaire platform that enables web-based and multi-mode data collection strategies, answering respondents' demand for more convenient electronic self-reporting modes. In addition, Statistics Canada has made important moves towards improving the management of cases in collection (such as case prioritization and implementation of responsive collection design (Laflamme et al. (2016))), improving the allocation of interviewers' workloads and managing survey operations more actively.

Statistics Canada recently began the second phase of its research, focusing on new data collection methods and techniques that might be more aligned with respondent preferences and reduce respondent burden. These new primary data collection modes aim to be easier to use, more efficient and less burdensome than the usual collection approaches, or even eliminate the need for surveying altogether. This paper also seeks to provide, in Section 2, an overview of Statistics Canada's recent successes in better managing its alternative data collection process and practices. It also briefly describes Statistics Canada's new data collection initiatives.

2. Phase 1: Better management of current collection approaches

This section presents an overview of initiatives that have been successfully implemented and that have contributed to alleviating the downward trend in response rates.

New e-questionnaire platform

Survey respondents in Canada increasingly expect an electronic self-response mode. Some years ago, Statistics Canada set out to build this option for its respondents, while at the same time replacing a myriad of data collection systems that were becoming increasingly difficult and costly to maintain. The resulting Integrated Collection and Operation System was first used for the 2016 Census of Population and later adapted for use by all business, household and agriculture surveys, as well as for Consumer Price Index data collection.

This new system has resulted in approximately 80% of Statistics Canada's surveys now offering an HTML-based, multi-mode-ready questionnaire, which can be delivered to a respondent's computer, laptop or other mobile device, and which can also be accessed by interviewers in homes or in a call centre.

The remaining 20% of surveys are planned for migration to the new system within the next 24 months.

The new e-questionnaire platform is achieving two goals. The first is to provide respondents with their preferred response mode. The second is cost savings, since the self-response mode is reducing the hours of interviewing required. The estimated annual savings from offering an e-questionnaire option (not including the census) are CAN 2.9 million so far.

Case prioritization and interviewer allocation

Case prioritization was developed to improve sample representativeness by targeting high-priority surveys or cases that belong to domains with lower response rates. In some circumstances, case prioritization might be used to target specific cases for various operational reasons. The objective is to monitor data collection while it is in progress to identify the cases to prioritize. It is one of two “adaptive” approaches (the other being responsive collection design), which use information available before and during collection to adjust the collection strategy for the remaining in-progress cases.

As part of the recently deployed collection platform, rules to deliver cases according to the highest priority have been introduced to govern work in Statistics Canada’s five call centres. These rules have various levels. For example, at one level, the rule assigns cases to call centre agents so that they work only on a given survey, or in proportions x , y and z on several given surveys. The call centre would pay attention to these particular surveys on that day. Next, the prioritization system targets specific operations, such as non-response follow-up or refusal conversion, within the priority surveys.

The allocation of interviewer efforts is related to case prioritization. Research at Statistics Canada had shown that staffing levels were not always well aligned with the workload sample and expected productivity (Laflamme (2008a); Laflamme (2008b)). In response, Statistics Canada has optimized interviewer efforts on cases where they will be more efficient.

Another initiative is to automate the delivery of specific cases, such as those eligible for responsive collection design.

Responsive collection design

Responsive collection design (RCD) is a technique Statistics Canada has used in production for all computer-assisted telephone surveys since January 2015, following a series of experiments in previous years.

Using RCD at Statistics Canada resulted in higher response rates and improved data quality, without increased costs or burden to Canadians. A typical RCD approach divides the collection operation into phases. The earliest phase begins the survey with a traditional, randomized allocation of questionnaires to interviewers. Next, the interviewers are asked to complete certain cases that

are more likely to bring about a successful response. The final phase emphasizes more difficult cases to reduce the differences in response rates between the domains of interest.

Active management

Traditionally, Statistics Canada surveys have been managed through the regional offices, where all survey taking takes place. Operations management was left entirely to those offices, with relatively little planning and support from the central office located in Ottawa. Recently, Statistics Canada changed this approach and introduced a new set of common plans and tools to centrally manage survey data collection in progress through an active management unit. The active management unit in Ottawa has three main objectives. The first is to determine data collection milestones where changes to the collection strategy are required. The second objective is to identify problems as early as possible and correct them (if required) before collection has finished. The third, which is a more global objective, is to use collection resources effectively to find the most appropriate balance between data quality, timeliness and survey costs. Active management is based on current, timely and empirical observations, and it is considered one of the main reasons Statistics Canada's response rates have stabilized.

Active management is best demonstrated by explaining the tools that managers now have at their disposal as a result of this program. For any given survey, all data collection managers have access to a national production plan before collection (which they have an opportunity to influence). They also have access to monitoring reports delivered centrally on a regular basis. The monitoring reports come with basic analytical information designed to identify collection issues and potential solutions to any issues that are noticed through monitoring.

The plan for active management is to refine available tools, including by implementing data visualization tools, and to continue to work towards establishing operational survey "command centres" in each region and at headquarters in Ottawa. The goal is to improve responsiveness and optimize data collection resources.

Expansion and improvement of respondent communication material

About five years ago, Statistics Canada focused on communication material to improve response rates. The idea was to "nudge" respondents using the latest research on behavioural economics and show, through various tools and integrated activities, that survey participation is useful. A framework was developed to prioritize needs for communication support based on the type of survey, the importance of the survey and the expected response rate.

The best example of an effective new communication strategy is the one used for the 2016 Census of Population. Based on results from the previous Census of Population, the population of Canada could be divided into five different groups, each with its own unique communication strategy. The groupings were based on the likelihood of a fast response to the census. One group, the easiest to reach, got relatively light communication. People who are more difficult to reach, on the other hand, received communications at various stages of collection. This segmentation strategy is an important reason why the 2016 Census of Population was considered the best ever in Canada, with the highest response rate on record, and with an impressive cost and quality performance.

3. Phase 2: Experimenting with new data collection methods

As mentioned earlier, focusing only on optimizing Statistics Canada's current collection operations would be insufficient. New ways of collecting data must be explored to reflect the new reality of a population less interested in completing surveys and to take advantage of technologies now available that could transform primary data collection operations. This section presents Statistics Canada's research focus areas, considering the anticipated operational implementation feasibility.

Developing a crowdsourcing service

Crowdsourcing has been an early success in the introduction of new primary data collection techniques. Crowdsourcing involves asking the population to proactively provide information rather than wait to be contacted when selected as a respondent. The risk of such an operation is obvious to the statistician—crowdsourcing data quality is difficult to assess, with metrics on quality near-impossible.

Nevertheless, Statistics Canada began to experiment with crowdsourcing in areas deemed relatively low-risk. The technique was first used for a project to improve the available information about dwellings in Canada. The "crowd" was asked to provide GPS locations for a set number of dwellings posted on the Statistics Canada website. That drew considerable interest from the population, who provided the requested information faster than expected. Secondly, Statistics Canada crowdsourced the price of cannabis through its StatsCannabis web application in the months before the legalization of cannabis in Canada in fall 2018, when cannabis consumption was still illegal (except for approved medical use). This resulted in over 20,000 entries to the questionnaire, and reasonable price estimates (i.e., within expectations, upon validation).

Subsequently, Statistics Canada implemented a crowdsourcing service within its survey operations branch. The service is relatively simple—a short e-questionnaire, accessible to all website visitors (i.e., there are no barriers such

as access codes to keep people from responding). For each crowdsourcing operation, there is a structured communication plan aimed at specific crowds. Active monitoring processes are also used to ensure projects are successful. In the last eight months, a number of new crowdsourcing operations have begun in Canada, mostly to collect qualitative information for Statistics Canada, as it designs new products and services. Public participation in Statistics Canada crowdsourcing has shown successful results.

Using SMS messages as a reminder to respondents

Statistics Canada is currently piloting the use of SMS (short message service) as a survey reminder strategy in an effort to encourage respondents to report their data after their initial invitation. This is part of a strategic plan to take a more user-centric approach to contacting respondents.

The first pilot survey saw 13,000 respondents receive text messages as their fifth (last) reminder, and it generated a response rate of 1.5%. This compares with the usual take-up rates of 1% for paper (mail) reminders and less than 1% for email follow-ups. More research is planned to assess the impact of using SMS for earlier reminders (first, second or third) to improve comparability with other modes and evaluate the cost effectiveness of this new way to contact Canadians.

The pilot survey was implemented in partnership with a major telecommunications company in Canada, using an SMS aggregation system. This tool enabled Statistics Canada to send automatic mass communications effectively at a reasonable cost of CAN 0.50 cent per SMS.

Before testing the use of SMS, Statistics Canada engaged in extensive public consultation. This consultation revealed that respondents are likely to view an unsolicited SMS from Statistics Canada (i.e., cold call via text) negatively, but would find it acceptable if they had already been in contact with Statistics Canada through another mode and were informed about the potential use of SMS. In addition, Statistics Canada consulted with the Office of the Privacy Commissioner, which reiterated the recommendation that Canadians be given advance notice that they might receive an SMS. This is why Statistics Canada has chosen to use SMS for reminder notices, rather than earlier contact with respondents.

Using a Statistics Canada data collection application on mobile devices

Statistics Canada is currently investigating the use of a mobile application to collect data for household surveys that require respondents to report information several times a day or on several days. This would give respondents a readily available, user-friendly collection tool for surveys that require repeated input, such as time use surveys or consumer expenditure reporting.

In addition to providing a convenient way for respondents to complete some of the most burdensome surveys, an application could take advantage of the option to ping respondents at strategic points in time to nudge them to respond. It could also benefit from the multiple sensors that smartphones currently use, including GPS and pedometers, as well as any connected devices with sensors, such as Fitbits. Data collected elsewhere on the device could be used by the application if the respondent permitted it and if the data suited the project using the app.

A first pilot project is planned for 2019/2020: a new survey measuring subjective well-being in Canada. Because of potential privacy issues, Statistics Canada is currently studying the legal and IT risks before completing specifications for the required solution and is working in collaboration with the Office of the Privacy Commissioner and IT security experts.

Testing cognitive interactive voice response technology

Later in the 2019/2020 fiscal year, Statistics Canada anticipates testing cognitive interactive voice response (IVR) technology as an interviewer or respondent monitoring tool. Cognitive IVR is a way for humans to interact with an artificial intelligence platform, such as IBM's Watson, Google Duplex and other similar products. This can be done with a number of methods, but as a first step, Statistics Canada is planning to explore ways to automate quality control and interviewer feedback through a cognitive IVR system. In addition, Statistics Canada hopes that the instant feedback of these platforms will enable it to successfully collect more data by being able to better tailor its tone and approach to each individual respondent.

This would be a first step towards using cognitive IVR in a way that would have bigger consequences for Statistics Canada's data collection operations. In addition to providing automated feedback to human phone operators, this technology could enable respondents to call a phone number and be interviewed by the cognitive IVR system, just like they would with an interviewer. This could be used for all or part of an interview process.

Using scanner data to collect information

Statistics Canada has completed the first year of a three-year plan to introduce scanner data into the production of the Consumer Price Index (CPI). The ultimate goal is to replace all traditional food price collection in the field (in-store collection). A simple implementation plan has been put in place whereby each field-collected quote is being replaced by an average price for the same or similar product using scanner sales data. Statistics Canada has successfully integrated one major retailer and is scheduled to introduce two more in the fall of 2019. Savings from the reduction of field collection costs are

being reinvested into the development of tools necessary to support collection from alternative data sources such as scanners.

This new source of data required new processing tools to be created outside the traditional system. The scanner data need to be pre-processed to link the products to the CPI classification—machine learning is now being used for this classification process.

Further developments are being considered, such as automated substitution of products and the use of multilateral methods of index calculation (i.e., the use of all data over time rather than a sample that mimics field collection). Both of these developments are beyond the scope of the three-year plan and would not be implemented before 2021.

Using sensors to collect information—satellite and telemetry

Satellite imagery is a key component of the Agriculture Statistics Program. Statistics Canada has collected vegetation index values from satellite imagery since the 1990s to support the Crop Condition Assessment Program, a web mapping application that depicts crop and pasture conditions across Canada in near real time. This data source, coupled with climatic data, is the foundation for the crop modelling project. The results of this project have been accurate enough to replace traditional collection methods for the September Field Crop Survey since 2016, eliminating more than 9,000 phone interviews. In 2019, Statistics Canada is looking to expand the modelling approach of the Field Crop Survey to further reduce response burden.

In addition, Statistics Canada successfully used a combination of crop insurance data and a crop classification map produced by Agriculture and Agri-Food Canada with medium-resolution satellite imagery to estimate crop area. Results were primarily used for validation at first, and this method is another potential way to reduce response burden for the Field Crop Survey and future censuses of agriculture. A 2019 pilot project is looking into updating crop area and yield on a weekly basis, at the parcel level, as the growth season progresses (in-season estimates) using near real-time satellite imagery, climatic data and crop insurance data.

Statistics Canada is also working on an innovative project with the Canadian Food Inspection Agency using the traceability data managed by the Canadian Pork Council, which collects group movement data. Statistics Canada is using data science and leading-edge methods to clean, process and use the data to create real-time modelled pig inventories by location, along with the probabilistic movements of each animal in the group. This partnership on traceability in the pork industry is an excellent opportunity to benefit each organization's goals. Statistics Canada will have information on pig movements and the inventory required for its statistical programs. The Canadian Food

Inspection Agency will benefit by having a framework on disease predictability to help monitor potential outbreaks of disease, such as African swine fever.

Exploring web scraping as a new mode of collection

Currently, Statistics Canada's Annual Survey of Manufacturing and Logging Industries uses available information to prefill components of the annual questionnaires to facilitate reporting of the commodities being produced, as well as the sales amount. The commodity data are prone to non-response and reporting errors but are crucial for measuring economic production in Canada. A pilot project will investigate the use of web scraping technology to collect website information to get a better picture of the types of commodities manufactured and to potentially improve sales estimates. A generic web scraper will be used to collect text data, which will be transformed into a list of commodities from each company's website. This information will then be used to prefill the questionnaires described above and improve the high non-response. In addition, the data will be used to improve auto-coding rates and quality.

Automated web scraping involves important ethical and privacy issues for national statistical offices. In response, Statistics Canada is developing a directive on web scraping to establish recommendations about using API when available, consulting the robots.txt of websites, respecting website controls and protocols, not capturing personal information, being transparent, and addressing other issues. All of this work will be done in close collaboration with the Office of the Privacy Commissioner.

Statistics Canada is also exploring web scraping from news platforms to collect information that can support enterprise profiling, financial variable coherence analysis, merger and acquisition event detection, and sentiment indicators for measuring business tendencies based on news analytics.

4. Conclusion

While it may be a challenge for national statistical offices to collect data with response rates declining around the world, Statistics Canada has implemented changes to stabilize response rates and adapt to respondents' preferences in terms of collection or contact mode. In addition, Statistics Canada feels that the new tools and techniques that are now available show great promise in supporting its work. These changes and techniques were summarized in this paper. Statistics Canada is happy to partner with other national statistical offices interested in experimenting with and developing such solutions, to improve the future of data acquisition.



Sensor data at the heart of innovation in official statistics



Sofie de Broe¹, Ger Snijkers¹, Barry Schouten¹²

¹Statistics Netherlands, ²Statistics Netherlands and Utrecht University

Abstract

This paper deals with two approaches of using sensor data in official statistics. Both approaches lead to hybrid data collection in which sensor data is combined with additional survey data. The first approach invites sample units to collect sensor data through mobile devices, wearables and/or IoT type sensors. The second approach assumes that sensor data have already been collected by sampling units and these units are asked for consent to use these data for a specified time period. The two approaches can be applied to both business statistics and social statistics. However, the type of sensors that are employed and the resulting sensor data are typically different for the two target populations. In the paper, we introduce criteria to determine the utility of sensor data for official statistics. We do so from the perspective of output, from the perspective of the sensors and from the perspective of the target populations. We illustrate the criteria for two realistic case studies; one for social statistics and one for business statistics.

Keywords

Sensor data; Mobile device; Wearable; Survey; Measurement equivalence

1. Introduction

Sensor data have become omnipresent in business processes and in daily life. Mobile devices have become standard tools for communication, daily archiving and personal administration within the time span of just a decade and have a high population coverage worldwide. Simultaneously, authorities and businesses have implemented all kinds of sensors for monitoring and improving of events and activities. Because of their high population coverage and daily life use, mobile devices, wearables and IoT sensors have become tools that may supplement surveys with automated data from sensors. Some of these sensor data may also replace survey data or may even have the potential to introduce revised concepts and views on statistics.

Naturally, the potential to record or link sensor data is not sufficient reason to also do so. Criteria are needed that identify combinations of survey topics

¹ The authors like to thank Ole Mussmann for his input on mobile device and wearable sensors and Tim Punt for his help in applying the various criteria to the agriculture case study.

and sensor data that are promising. In this paper, we will identify such combinations in an official statistics context.

We see two approaches to employing sensor data: self-initiated and existing sensor data. Under the first approach, sensor data do not yet exist but are collected by responding businesses, persons or households. Under the second approach, consent is asked to link existing sensor data that businesses, persons or households have collected or are collecting. The two approaches are not disjoint; in fact, they can be combined in a single data collection. They can be further enriched using survey data. As a result, hybrid data collections arise.

In the subsequent sections, we introduce criteria to evaluate the utility of sensor data, briefly describe various types of sensors, illustrate ideas through two examples and end with a discussion on future research.

2. Criteria for sensor data

Obviously, the mere possibility to collect sensor data is not enough reason to also do so, as it requires a separate and new architecture and infrastructure and respondents may not be willing to share the data. On the data collection side, sensor data, especially when they are collected invitation by respondents (i.e. primary data collection), demand for new data collection channels. These channels demand for new and/or additional processing tools and skills, for expansion of existing monitoring and analysis tools and for a redesign of survey estimation methodology. Such changes are costly and time consuming. On the respondent side, sensor data may still be burdensome and/or may be privacy intrusive. Hence, a strong business case for mobile device sensor data is needed and respondents need to benefit as well.

Sensor data are candidates to enrich or replace survey data when a survey is relatively costly and/or inaccurate, when sensors are relatively cheap and/or accurate, and when respondents are willing to share sensor data. We, therefore, consider criteria from three perspectives: the survey quality-costs, the sensor quality-costs and the respondent.

From the survey point of view, existing survey topics may be candidates for enrichment or replacement with sensor data, when they satisfy at least one of the following criteria:

- Burden: The survey topic(s) are burdensome for a respondent, in terms of time, cognitive effort, or data retrieval;
- Centrality: The survey topic(s) are non-central to respondents, i.e. the average respondent does not understand the question or does not know the answer;
- Concepts: The survey topic(s) do not lend themselves to a conceptualisation through a survey question-answer approach to begin with;

The three criteria refer to the survey topic properties that are prone to measurement error and/or nonresponse. To provide examples: Examples that satisfy the first criterion are topics that require keeping a diary for a specified time period, say a week or a month, and provide details about all time periods. Other examples are surveys that require consultation of administration and archives, for example about assets, investments and finances. The second criterion is satisfied, for instance, for travel surveys where respondents need to provide exact coordinates of locations they have visited, for health surveys where respondents need to describe sleeping patterns and STS and SPS surveys where businesses need to provide detailed information about their activities and output. The third criterion applies to complex socio-economic or psychological concepts such as happiness, health or wealth, or research and development where many questions are needed to derive latent constructs.

From the sensor point of view, the main criteria are:

- Omnipresence: The sensor(s) are available to most population units. i.e. as sensors in contemporary devices and wearables or as sensors in IoT systems;
- Data access: Data generated by the sensor(s), as well as metadata about the properties and accuracy of the sensor data, can be accessed and processed;
- Quality: The sensor data is comparable, reproducible and accurate;
- Costs: Any costs associated with the sensor(s) or implementation of the sensor measurement process are affordable for an NSI;

The four criteria all link to the utility of the resulting sensor data. The omnipresence criterion refers to the coverage. In theory, tailored instruments can be developed that record complex phenomena and behaviours, but these are until now used only in lab settings. Smartphone sensors are examples of omnipresent sensors, whereas sensors in wearables have a much lower population coverage and pose challenges with regard to data access. The data access criterion means that sensor data can be stored, manipulated, processed, evaluated and interpreted. For instance, location data can be stored and processed, but it is not always clear what sensor, GSM, Wi-Fi or GPS, produced the data and how accurate the data are. For businesses there is the issues that so far, only few innovative companies have installed sensors because of costs issues. In order to produce business statistics, a representative number of companies should have sensors installed. The quality criterion originates from the statistical objective to derive accuracy of statistics and to be able to compare statistics between persons or companies and in time. Sensor accuracy is the equivalent of survey accuracy and should be evaluated on measurement and missing data properties. For instance, location data can be used to estimate travel distances but are subject to missing data, measurement errors, sensor data drift and potentially also device

effects. As such, sensors are just like other data collection instruments. The final criterion applies when sensors need to be provided to respondents and in companies and refer to costs associated with their installation and (operational) use.

From the respondent point of view, sensors may vary in their intrusiveness. Four criteria follow:

- Willingness: Persons/businesses are willing to consent to provide the sensor data;
- Data handling: Persons/businesses can retrieve, revise and delete sensor data on demand, i.e. they have ownership over the data;
- Burden: Persons/businesses are willing to devote the effort needed to collect and handle the sensor data;
- Feedback: Respondents may retrieve useful knowledge about themselves;

In order to employ sensors, respondents/companies need to be asked for consent to activate sensors and to store and send data. Most mobile device sensors require consent by default. Exceptions are the various motion sensors that can be activated in Android without consent. However, even without the technical necessity to ask for consent, there are legal and ethical reasons why consent is imperative. Willingness to consent varies per type of sensor and depends on the context and purpose of the measurements. Recent literature, see Struminskaya et al (2018) for an overview of studies, has investigated willingness and confirmed differences between sensors and settings. Obviously, the more intrusive a sensor measurement is, the more respondents will refuse and the larger the potential damage of missing sensor data. Recent European legislation require that respondents can get copies of their data and can request deletion of their data at any time². This requirement puts constraints on the storage of and access to sensor data. Next, sensor measurement themselves, such as photos or sound recordings, require some respondent effort. This effort may be too great for respondents so that missing data and/or lower data quality result. Finally, the sensor data may be fed back to respondents in an aggregated form and may provide valuable information to respondents and companies. In the last case, this offers great potential for a data circle between the NSI and companies.

Sensor data can be collected passively or actively. Passive sensor data is collected without respondent intervention or feedback, apart from consent. In active sensor data, respondents are asked to check, revise, accept and/or supplement sensor data, i.e. the respondent is involved in data collection. Motives for active sensor data collection are increased response rates,

² See <https://gdpr-info.eu/issues/right-of-access/> and <https://gdpr-info.eu/issues/right-to-be-forgotten/>

increased data quality and hybrid forms of survey and sensor data. An example is a travel survey in which respondents' locations are stored whenever a device is in motion. These data can be collected passively. The data may also be shown to respondents for quality checks and for enrichment of sensor data with stop motives and other context information. Active data collection is much more demanding as it requires real-time data handling and a careful design of a user interface.

Summarizing, there are various criteria from the person/company, sensor and surveyor points of view that need to be confronted with costs and logistics of sensor data collection and processing.

3. Mobile device sensors

We discuss the two forms of sensor data: self-initiated sensor data and existing sensor data.

3.1 Mobile device sensors and wearable sensors

The following elements and sensors are supported by many contemporary smartphones and tablets (see Mussmann and Schouten (2018) for details about the sensors): 3D touch, motion sensors (accelerometer and gyroscope), ambient light, Bluetooth, camera, camera flash, cellular (or GSM), fingerprint, GPS, heart rate, humidity, magnetic field, microphone, NFC (near field communication), pressure, proximity, screen, speaker, thermometer, vibration, Wi-Fi and wireless charging.

Wearable devices are mobile devices that one can wear or attach to clothes. Their functionality is more specific and targeted than smartphones and tablets in order to reduce size and weight. The motivation to wear a device is often health-related, but wearables do support other functions. Most common wearables are activity trackers, fitness bands and smart watches, which are intended for continuous use. Less common wearables are shoe clips, smart glasses, clothes with sensors and jewellery with sensors, which are intended for temporary use. Due to their more specific functionality, wearables often require other mobile devices to communicate with users. A dedicated app needs to be installed on a smartphone or tablet that provides a user interface to set or alter wearable settings and to read summary data and statistics. Wearable sensor data can be sent to the user directly or indirectly through the producer of the wearable. In the latter case, often only edited and aggregated data are available and the raw sensor data remain at the side of the producer. Consequently, wearable sensor data lie somewhere in between primary and secondary data collection. They are mostly self-initiated but often are maintained and owned by another party. Due to the closer proximity to a respondent's body and the fact that they can be worn 24 hours per day and 7 days per week, wearables can measure data that smartphones and tablets

cannot. Examples include detailed sleeping patterns, calorie usage, more detailed activity tracking or electrocardiography. Typical sensors that are included in wearables are: motion sensors, Bluetooth, GPS, heart rate, screen and vibration. Less common sensors are: Blood oxygen levels, camera, light sensor, magnetic field, NFC (near field communication), speaker, thermometer and Wi-Fi.

3.2 Existing sensor data

Sensor data may already be collected and owned by various parties and not necessarily for statistical purposes. We term these data secondary sensor data. We distinguish data based on user-owned sensors and data based on the public Internet-of-Things (IoT) sensors (i.e. not owned by private users). Although a clear distinction is hard, we view privately owned IoT type sensors, such as weather stations or burglar protection systems, as user-owned sensors. So the main distinction is between individual and public use. In this project, we consider linkage of secondary sensor data to samples of a target population after respondent consent.

Within user-owned sensors, a distinction can be made between commercial and non-commercial data collectors. A range of companies produces apps for mobile devices, in particular wearable devices, to provide paid services to individual customers. Other companies, such as Google, also provide unpaid services in exchange for the right to use the resulting data for commercial purposes. These services are usually continuous and, consequently, create a dynamic, vast stream of sensor data. The resulting data are mostly based on self-selection, i.e. the initiative is with the user, and not based on invitation. Non-commercial parties may collect mobile device sensor data for research or policy making motives. Such data collection often has a finite time horizon and is invitation-only. In all these cases, however, the sensor data are self-initiated by users, but the data is stored, handled and owned by others. Apart from privacy and legal constraints, such user-owned sensor data can potentially be linked to individual respondents in surveys. After linkage, a hybrid data collection follows where part of the data may be asked through questions, part of the data may be measured on respondent mobile devices and part of the data may be linked. An example is activity tracker data owned by a third party linked to persons in the sample supplemented by survey data on health perceptions and health determinants. Another example is budget expenditure diary data linked to supermarket scanner data and supplemented by scanned receipts for other types of stores. IoT sensors are usually owned by government authorities and implemented for very specific purposes, such as monitoring of weather, pollution or traffic. The sensors are attached to objects, often have a fixed location, operate continuously and measure activities of multiple persons. Data is stored, handled and owned by the same

authorities independently of persons' consent, but subject to privacy constraints and legislation. IoT sensor data do not necessarily include identifying information about the persons to which data correspond. As a consequence, linkage of IoT sensor data to individual respondents may be hard or even infeasible without additional information such as time stamps and/or location coordinates. Nonetheless, hybrid forms of data collection may arise where respondents are asked for additional information that enables linkage.

Nowadays, a lot of data within businesses are already available in electronic format. A first example regarding System-to-System (S2S) data collection for statistical purposes involves sensor data. Increasingly, electronic sensors are used to run a business, e.g. by agricultural businesses (like dairy farms using milking robots). A second example is financial data in a fully integrated and digitalised business information chain which makes S2S data communication for financial, tax and statistical reports possible. Drivers for switching from questionnaires to S2S are: working towards smart business statistics, (i.e. timely and new statistical output integrated in business processes), the reduction of response burden, and monitoring and benchmarking businesses to their counterparts.

4. Examples of both types of sensor data

We consider two examples: two surveys mandatory in the European Statistical System (ESS): The Household Budget Survey (HBS) and the ICT survey, and sensor data on potatoes crops as a potential replacement for mandatory ESS survey on crop yields at an innovative farm in the Netherlands.

4.1 Household Budget Survey and ICT survey

The HBS records all household expenditures and purchases during a week and large household expenditures and purchases during a longer period up to a month. The HBS is very burdensome due to the duration of the diary keeping. Detailed expenditures and purchases are non-central to respondents. The HBS does not involve complex latent concepts that relate to many survey questions. The HBS deals with all kinds of purchases and expenditures, both small and large, and both frequent and infrequent. Some of these purchases are done on site, such as shops, restaurants and cinemas. Time-location sensor data may be employed to assist respondents in memorizing or recalling locations where products or services have been purchased. Some purchases are done online and part of those may be done through a mobile device. The use of certain online shopping apps may be tracked to again assist the respondent. In all of these cases, direct access to the type, amount and cost of products and services will, generally, not be possible to privacy restrictions on the apps. Another option is to use the camera to scan shopping receipts. This

is mostly useful for purchases that involve many products/services simultaneously and that are burdensome to insert into a diary/questionnaire. The HBS might make use of scanner data from shops, as well as bank transaction data and loyalty card data. Scanner data would be limited to large chains and can be matched to a respondent via time-stamp and purchase amount. Bank transaction data can cover purchases independent of shops but have varying coverage depending on the penetration of electronic payments in the target country. Requesting bank transaction data is of course more intrusive to the respondent's privacy. In project @HBS, financed by Eurostat, the utility of these data sources is explored.

The ICT Survey collects survey data on the frequencies, durations and purposes of use of all forms of contemporary ICT (TV, internet, social media, mobile devices). The survey is not burdensome and a survey type instrument is well-suited due to the lack of complex latent constructs. However, ICT use, especially frequency and duration, and details about the respondent ICT facilities are non-central. As mobile devices are ICT by themselves and connect to other ICT, they can provide sensor data about the use of the mobile devices and connected devices. Obviously, the type of device and operating system (OS) may be derived. Furthermore, the measurements may consist of the frequency and type of use of the mobile device, such as social media, online browsing, SMS texting, gaming, video streaming, taking pictures and phone calls. Apart from the presence of apps on the device that are used for these purposes, one may also consider the frequency and amount of use. However, access to app (meta)data is, of course, limited for privacy reasons and the potential amount and detail varies between apps, even when respondents would consent. The mobile device may also provide insight into other ICT through its Wi-Fi connection(s) and through Bluetooth connections. The mobile device can, for example, measure the type, speed and strength of the Wi-Fi at home or provide a list of Bluetooth connected devices. Although it is not the purpose of the ICT, the mobile use data also present a view on the general ICT profile of a person. The ICT surveys could benefit from social media data or mobile phone provider data or internet provider data. Social media data would be provided by global companies like Facebook, Twitter or Instagram. Data access could be requested from the respondent for public data by providing user names or handles, or from the companies for private data. Mobile phone provider data would be contributed by a mobile phone provider local to the country of the respondents. Possible data include phone call, SMS and internet usage as well as time-location data with a coarse resolution in space and time. The respondent's internet provider could supply data covering the internet use.

Tables 1 to 4 show our scores on the sensor and respondent criteria for, respectively the self-initiated and the existing sensor data. In some cases, question marks are given indicating that any score would be speculative.

Table 1: Assessment of the sensor criteria. A flag means that the pair scores positively on the criterion.

<i>Survey</i>	<i>Sensor</i>	<i>Omnipresence</i>	<i>Data access</i>	<i>Quality</i>	<i>Costs</i>
ICT	Device use	✓		✓	✓
	Device properties	✓		✓	✓
HBS	Location	✓	✓	✓	✓
	Camera	✓	✓		✓
	Device use	✓		✓	✓

Table 2: Assessment of the respondent criteria. A flag means that the pair scores positively on the criterion.

<i>Survey</i>	<i>Sensor</i>	<i>Willingness</i>	<i>Data handling</i>	<i>Burden</i>	<i>Feedback</i>
ICT	Device use	?		✓	✓
	Device properties	?		✓	
HBS	Location	✓		✓	
	Camera	?	✓		✓
	Device use			✓	✓

Table 3: Assessment of the sensor criteria. A flag means that the pair scores positively on the criterion.

<i>Survey</i>	<i>Sensor data</i>	<i>Omnipresence</i>	<i>Data access</i>	<i>Quality</i>	<i>Costs</i>
ICT	Social media	✓		?	✓
	Mobile provider	✓		?	✓
	Internet provider	✓		?	✓
HBS	Scanner data	✓		✓	✓
	Bank transactions	✓		✓	✓
	Loyalty card			✓	✓

Table 4: Assessment of the respondent criteria. A flag means that the pair scores positively on the criterion.

<i>Survey</i>	<i>Sensor data</i>	<i>Willingness</i>	<i>Data handling</i>	<i>Burden</i>	<i>Feedback</i>
ICT	Social media	?		✓	
	Mobile provider	?		✓	✓
	Internet provider	?		✓	✓
HBS	Scanner data	?	✓	✓	✓
	Bank transactions	?	✓	✓	✓
	Loyalty card	?	✓	✓	✓

4.2 Sensor data at an innovative farm in the Netherlands

As a starting point for investigating the use of business sensor data we opted for the agricultural sector as a first candidate for two reasons. First of all

getting high enough response rates for agricultural surveys has always been cumbersome. Secondly the agricultural sector is innovative and developments like smart and precision farming seem promising within the context of official statistics (De Vlieg, 2018; Vonder, 2017; Thomas & McSharry, 2015).

In order to run his farm, in the context of precision farming the innovative farmer uses many sensors throughout the year. During winter he scans his fields for soil characteristics like humidity, sun coverage and nutrients. During spring, new crops are planted in such a way that it matches the yield potentials of field sections. In the summer, crop growth is carefully monitored, by measuring soil humidity, nutrients in the plants, and the need for pesticides. When the crops are harvested in autumn, the crops are weighted on the spot, providing detailed insights in field sections with high and low yield.

The farmer's data are stored in two different data platforms: a platform for precision farming data, and the so called 'harvest registration system'. The precision farming platform is specifically tailored to the farmer's needs and houses the sensor-generated data. A first exploratory analysis of this database shows that parts of questionnaires can be imputed using these data, while other parts still have to be completed manually: data on fields and harvest cover the requested information, while for the planting, plant treatment, and number of employees parts of the requested data are missing in this database; financial information requires other data sources altogether. As a result, we concluded that parts of these sensor data are associated with some of the concepts that are operationalised in questionnaires. However, before the data could be analysed two steps had to be taken. First of all, a complete and accurate description of the data was missing. The next step was data cleaning: the data suffered from e.g. measurement errors (incorrect values), and duplicate records. Some errors could be corrected by checking the data, but in order to get a complete understanding of the content of the data the farmer's expertise about the metadata and the data generation process was indispensable. While precision farming platforms such as this are promising, it currently lacks the required data management and omnipresence necessary to produce high-quality statistics. This of course is subject to change based on developments in the industry. In addition, other challenges need to be addressed when using these kinds of decentralized platforms: localizing them in the first place can prove to be challenging, as well as getting access to the data; this relates to issues like data ownership, data sharing, and trust, among others. When getting data from various providers data harmonisation may also be an issue: sensors produced by various manufacturers and used by various farmers may be generating different kinds of data. These issues could be overcome by the establishment of data service enters (DSCs). The second data platform is a so called 'harvest registration system', which is used by many farms in the Netherlands to import, store, interpret and share relevant data

concerning their fields. The number of generally used systems is limited to 3 or 4. The input for these systems can be both manual (H2S) and digitally (S2S) covering many subjects NSIs are interested in. These include: parcel and crop registration, pesticide usage, fertilization and harvest yields.

Even though this platform still has to be explored, harvest registration systems show great potential to be used as an alternative to surveys, as they are a rich source of information. Furthermore, the systems already have a build-in export function. A function that is being used to pass on information to stakeholders and a number of regulatory agencies alike. Here the farmers remains in full control which data are shared and which are not, acting as a gatekeeper. As for NSIs, these systems offer great opportunities to close the data circle by S2S gathering of data and returning valuable statistical input back into the systems.

5. Discussion

We contribute to existing literature in two ways: We propose a set of criteria to support cost-benefit assessments of sensor measurements and sensor data and we illustrate the criteria for two real-life cases. The criteria are constructed from three viewpoints. The first viewpoint is from the perspective of the survey itself; does the survey contain topics or questions that may benefit from automated measurements. The second viewpoint is that of the sensor: What are accuracy and costs of the sensor options. The final viewpoint is the respondent: How does the respondent react to a request for sensor data. One side remark is in place. Our assessments of the survey, sensor and respondent criteria are subjective. This is partly for the very reason that they are new and have not been tried in practice. It is very hard, for example, to predict willingness to provide sensor data (or to consent to sensor data linkage) independently of the context. Another reason is that sensors and wearables by themselves show variety in accuracy and costs, even within mobile devices, so that it is hard to judge about quality. It is imperative that the assessments are made more rigorously by consulting multiple experts. Also this exercise will be part of an in-depth follow-up paper.

In our inventory, we distinguish sensor measurements initiated by the survey institute and existing sensor data. Although the data may originate from the same type of sensors, the context of the two is very different. Self-initiated sensor data require a data collection infrastructure and lead to direct data collection costs. Obviously, other aspects such as data processing, data storage, privacy and legislation are very different as well. In this paper, we consider secondary data as complementary to survey data, i.e. we ask respondents to consent to linkage. Consequently, a hybrid form of data collection arises. However, such a combination of data sources is no goal by itself; secondary data may provide the sole source of data for specific topics.

It is likely that from the big data perspective, survey data may be used as complementary data in some settings. In the near future, the two starting points may, in fact, lead to similar hybrid forms of data collection.

The most promising combinations of sensor data and survey data are those that score well on all criteria. For the two examples, we found pairs of survey topics and sensor measurements that potentially have a positive business case. We make two recommendations: First, we advise to replicate our assessments of the various criteria with experts in mobile device and wearable sensors, especially for quality. Second, we propose to empirically test respondent willingness to provide sensor data and to consent to linkage to existing secondary data. Such experiments have emerged, but are yet at early stages.

Statistics Netherlands has defined an innovation strategy with a clear focus on making policy relevant statistics based on new methods and new secondary data sources. Sensor data offer great potential, but for an NSI the challenge lies in data access (as data are often in the hands of private partners), concept validation (do the sensor data measure the same as the survey data) and data quality. In order for sensor data to be of any use for end users (and to become part of an official statistical process), these issues still need to be addressed in order to use sensor data as a secondary data for official business statistics. To some extent NSIs lose control over the data collection. In order to gain some control and address the challenges, we foresee another business model for NSIs like Statistics Netherlands: Be a stakeholder and partner in emerging data hubs from the very start, and be in close communication with the end-user. The role of an NSI in these data hubs should be to enrich these data with relevant statistical information, and assist in the use of statistical information.

References

1. De Vlieg, J. (2018), 'A huge push in technology is coming' (in Dutch: 'Er is een enorme push in technologie in aantocht'). *Boerderij*, 03: 19 (februari 2018).
2. Thomas, R., and P. McSharry (2015), *Big Data Revolution: What Farmers, Doctors, and Insurance Agents teach us about discovering Big Data Patterns*. Wiley, Chichester, West Sussex, UK.
3. Van Dijk, C., and C. Kempenaar (2016), Open data for precision farming in the Netherlands (in Dutch: Open data voor precisielandbouw in Nederland). Wageningen University & Research, report 662. Wageningen, Netherlands.
4. Vonder, M. (2017), Sensors going smart. Presentation at 'Big Data Matters' Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (TNO Netherlands)



INEGI's statistical autonomy: Institutional governance and some ever-present risks



Mario Palma Rojo¹

Mexican National Institute of Statistics and Geography, Mexico City, Mexico

Abstract

Mexico's National Institute of Statistics and Geography (INEGI) is one of the few National Statistical Offices (NSOs) in the world that have been granted legal autonomy, and probably the only one that has it written in the country's constitution. In 2008 a legal institutional framework was set up to provide INEGI independence from all branches of government, as well as a coordinating role in the overall production of official statistics in the country. This existing regulation is characterized by establishing various legal and administrative provisions that help to safeguard the professional independence of the Institute -e.g. technical and managerial autonomy, own patrimony (resources), appointment of high officials only after agreement of both the legislative and executive powers, etc. The guaranteed publication of data on sensitive subjects that may in some instances contradict government claims, such as figures on national accounts and on victimization rates, are examples of the soundness of the institutional framework that regulates and protects INEGI's work. The case of Mexico allows to reflect on the benefits of granting autonomy to NSOs, but also on the ever-present risks of political interference that may threaten their professional independence and integrity.

Keywords

Professional Independence; Statistical Autonomy; Mexico; INEGI.

1. Introduction

All over the world, NSOs have made important advancements in the provision of official statistics to society. Just to cite an example, via the survey sampling foundation NSOs have been able to produce with confidence statistical data in a varied array of subjects -many of them sensitive and complex to measure-, to compare them with other sources of information -such as administrative records- and to integrate time series -which has allowed for the monitoring and evaluation of policies and development. Today most societies depend on economic, social, demographic and environmental data that usually comes from NSOs.

¹ This paper was prepared in collaboration with Victoria Bonilla Veliz (Advisers Office, INEGI). The opinions expressed and arguments employed herein are solely those of the authors and do not necessarily reflect the official views of INEGI.

At the same time NSOs are under tremendous pressure to improve and to protect their work. They face accelerating demands for faster and more disaggregated information, meanwhile globalization requires them to respond to emerging needs of internationally comparable statistics (e.g. monitoring of the Sustainable Development Goals -SDGs-), and the unparalleled changes in technology, challenges them to integrate external sources of information (such as Big Data) into their production of information. Additionally, they face increasing competition from private data providers, as well as the extension of fake news and misinformation campaigns which requires them to publicly defend the use of accurate official statistics.

This paper, however, focuses in one additional and more political challenge that NSOs face and that sometimes we may risk not noticing: the threat of political interference from their own national governments. Despite the international endorsement of the Fundamental Principles of Official Statistics (FPOS)² and the advancement in the institutional governance of NSOs, there are worrying examples from around the world of governments undermining the professional independence of NSOs and threatening the integrity of official statisticians.

The cases of Andreas Georgiou – former President of ELSTAT (Greece’s NSO), prosecuted by the Greek government for the alleged crime of inflating the 2009 Greek fiscal debt and deficit figures (IAOS, 2018)-³ and Anar Meshimbayeva -former Chairperson of the Kazakh Agency of Statistics, accused of artificially inflating the costs of the 2009 population census’ materials, sentenced to 7 years of imprisonment and condemned to pay the state over 1.5 million USD (Baer, 2018)- are just two examples that add to other instances in which there have been attempts to politically interfere in the work of NSOs and undermine their independence.

Some of these cases are Argentina -where political intervention in the processes and results of inflation data produced by INDEC (Argentina’s NSO) led the IMF to issue a declaration of censure on its Consumer Price Index and Gross Domestic Product official data in 2012, which was removed until 2016 once important remedial measures to improve the quality of the data were implemented-;⁴ Canada -where the government eliminated the mandatory long-form census in 2010 and replaced it with a voluntary household survey despite the warnings of its Chief Statistician (who resigned over the issue) of the consequences in the quality of the data (The Globe and Mail, 2014)-; and

² The FPOS were approved unanimously by the member States of the United Nations General Assembly on 29 January 2014.

³ More information on Andreas Georgiou’s case can be found in: (Aizenman, et al., 2017); (Walker, 2017); (Bloomberg, 2017); and (The Economist, 2016).

⁴ (The Economist, 2014) and (The Economist, 2017). For the IMF statements on Argentina, see (IMF, 2012) and (IMF, 2016).

more recently Puerto Rico -where the country's institute of statistics (PRIS) has been fighting multiple dismantling attempts by the State government, which replaced four of the seven PRIS Executive Board members in 2017 and presented a plan (approved by Puerto Rico's legislature) to incorporate PRIS under the Department of Economic Development and Commerce, where data collection would be consolidated and outsourced (Acevedo, 2018).

How to protect statistical independence? Undoubtedly the most important mechanism is through governance arrangements that legally set the autonomy of NSOs and shield them as much as possible from potential intervention by governments. One governance scheme is provided by the United Kingdom which established in 2008 the UK Statistics Authority as an independent regulator of the UK's official statistics system, operating at arm's length from government and reporting (accountable) to the UK Parliament rather than to the government through a ministry (Shah, 2018).⁵ And another one is provided by Mexico whose NSO -INEGI- has since 2008 been granted legal autonomy, becoming one of the few NSOs having this status.

This paper reviews the experience of INEGI and analyses the institutional framework set up to provide it with legal autonomy from all branches of government, as well as the legal and administrative provisions that help to safeguard its work. The case of Mexico and INEGI allows to reflect on the benefits of granting autonomy to NSOs and -even when institutional governance mechanisms are established- on the ever-present risks of political pressure and interference that may threaten the production of official statistics.

2. INEGI's autonomy: What does it entail?

Article 26 of Mexico's Constitution grants INEGI an autonomous status and charges it with the responsibility of both producing official statistics and coordinating the activities of the National Statistical and Geographical Information System (SNIEG) of the country. The autonomy established by law encompasses four elements: the technical independence of the Institute, its managerial autonomy, the recognition of legal personality and the guarantee of an own patrimony (resources). To operate these elements and safeguard in practice the autonomy of the Institute, the law establishes various legal and administrative provisions. Without the aim of being exhaustive, it is worth to highlight the following.

First, the operation of the SNIEG is governed by the National Statistical and Geographical Information System Act 2008 (SNIEG Act). This Act regulates the work of INEGI -as producer of information and coordinator of the SNIEG- and establishes the information responsibilities for all government units

⁵ For more information on the UK Statistics Authority and legislation that governs the UK Statistical System see the Statistics and Registration Service Act 2017.

integrating the national statistical and geographic system. It is based on the principles of access to information, transparency, impartiality and independence, which are set as the foundation of the whole operation of the SNIEG and as fundamental attributes of INEGI's work. By recognising these values in the law and regulating according to them, Mexico honours the implementation of the FPOS, and in particular of the principles of professionalism and independence.

Second, the law determines that a Board of Governors -composed of five members: one president and four vice-presidents- will direct the work of INEGI. The appointment of these high officials is only after agreement of both the legislative and executive powers. They are proposed by the President of the country and ratified by the Senate; and must fulfil a series of academic and professional experience requirements. The tenure of the Board of Governors' President is 6 years, while the one of the vice-presidents is 8 years. All of them can be renewed in their job for a maximum of two periods, and their renewal is set to be staggered to guarantee continuity in the working of the Institute and to avoid all (or most) members to change during one presidential administration. The Board of Governors decides by majority the causes for dismissal of any of its members. With these provisions the law restricts the possibility of the government unilaterally attempting to appoint or replace Board members.

Third, it is instituted in the law that all data produced by the SNIEG is considered as official information. Moreover, INEGI's Board of Governors has the faculty of determining what information should be regarded as information of national interest, based on a series of criteria covering their relevance for the design and evaluation of public policies, methodological soundness and quality, timeliness and periodicity, etc. The attribute of national interest implies the obligatory publication of this information and its use is also mandatory for all (federal, state and municipal) authorities. For this reason, each year INEGI must make public in advance its Publications Calendar, which sets the publishing dates of the information of national interest. These dates are not subject to change, hence giving certainty to users and the public that the information will be provided and shielding INEGI of potential political pressure not to publish it or delay their publication due to political concerns.

Finally, the law also establishes that the Federal Government Budget should include enough funds (resources) for INEGI to conduct National Censuses, integrate the System of National Accounts and produce National Price Indexes, as these are considered by law information of national interest. This arguably applies to all information judged of national interest to produce by the Board of Governors. Although it does not shield the Institute of

potential budget cuts,⁶ it does ensure that the provision of information by INEGI on subjects deemed of national interest cannot be affected by an unilateral decision of the government to substantially cut its funds, as the provision of this information is protected by law.

In sum, both its legal mandate and autonomous status have facilitated the Institute to operate as a neutral -external to all governmental activities and state units- producer of official statistics on relevant economic and social phenomena, as well as guardian of the quality of the information made available by the whole national statistical and geographic system. The legal institutional framework supporting that regulates and protects INEGI's work has helped to insulate the Institute from political interference. The guaranteed publication of data on sensitive subjects that may in some instances contradict government claims, such as figures on national accounts, inflation, unemployment, homicides and victimization rates, are examples of the soundness of this institutional framework. Notwithstanding, there are some ever-present risks.

3. Ever-present risks

As it has been mentioned above, NSOs face today increasing competition from private data providers. Many of them look to sell their work to the government and it may be in the interest of governments to hire them, especially if the statistical information provided by NSOs contradict their claims. The risk of outsourcing the production of statistics is that there is no guarantee the information produced by private data providers will meet the impartiality, quality, transparency, confidentiality and independence standards that NSOs follow in their production of official statistics.

Also, there is always the risk of budget reductions. NSOs already attempt to respond to huge demands to produce more and better statistical information, while working in an environment of constrained resources. Budget cuts exert even more pressure to their work, and if they are substantial they may seriously undermine the capacity of NSOs to comply with their remit, reducing consequently their production of relevant and quality official statistics.

Finally, despite a legal framework recognising the professional independence of NSOs, there might always be incentives/attempts by the governments to make political nominations for the high official positions at the NSOs, or even modify the law to influence the statistical offices' work. In this kind of cases, the defence of the NSOs independence must be in the

⁶ As it has recently happened at the beginning of 2019, situation which led the Institute to make personnel cuts, as well as administrative and programmatic changes (such as the cancellation of some surveys and other programmes). (Animal Politico, 2019)

interest of all. The international statistical community has played an active role in the condemnation of cases in which the professional autonomy of NSO's and the integrity of official statistics have been under threat.⁷ Although this has been effective to raise awareness of the situation, it is however not enough. To the international leverage in favour of NSOs' autonomy, necessarily should be added national commitment to legally institutionalize it. 7 Two of the most recent examples are the international statistical community condemnation of Andreas Georgiou prosecution (IAOS, 2018) and the dismantling of Puerto Rico's Institute of Statistics (IAOS, 2018a).

4. Conclusion: How to defend statistical independence?

From a structural point of view, a crucial quality for the work of NSOs is to have a legal administrative framework that establishes its autonomy and support their activities. Autonomy implies the possibility for NSOs of pursuing technical work without interference of any kind, but it also means for the governments that an agency -external to the government function- is measuring the social and economic phenomena in which the former is involved. While this restrains the evident conflict of interest, it is no surprise that governments may be reluctant to grant autonomy to NSOs, and, even if this one has already been granted to them by law, to politically influence their work.

INEGI's autonomy status insulates it from interference by the executive branch. International experience shows us that without a clear legal mandate and institutional framework supporting autonomy, it is at the very least complicated for a technical government agency to resist political pressures. Administrative provisions -such as the appointment of high officials only after agreement of both the legislative and executive powers, the obligatory publication of information of national interest, and the budgetary allocation of resources mandated by law for the information bestowed with this attribute, among others – have undoubtedly been crucial for INEGI to independently provide official statistics on relevant and sensitive economic and social phenomena, as well as to coordinate the work of other producers of information in the national statistical and geographic system of the country.

Therefore, if there is an institutional and legal framework explicitly supporting the autonomy of NSOs, as is the case in Mexico, it will be all for the best. This is not always possible as it requires a lengthy political and legal process, but in any case, it is indispensable that professional independence is guaranteed in some legal form.

⁷ Two of the most recent examples are the international statistical community condemnation of Andreas Georgiou prosecution (IAOS, 2018) and the dismantling of Puerto Rico's Institute of Statistics (IAOS, 2018a).

Still laws have to be implemented in real life and for that to happen NSOs have to become skilled at convincing of the implicit and technical merits of good statistics and at working together with institutions with quite different objectives and outlooks towards a common objective: the quality production of official statistics.

For this reason, it is safe to say that the permanent defense of statistical independence should be made also through the provision of quality information useful for the development of public policies; communication and coordination with government agencies; awareness of risks; and international communication.

References

1. Animal Político. Redacción, "El INEGI cancela o suspende 14 encuestas al no obtener los recursos que pidió para 2019", *Animal Político*, 24th January 2019. (<https://www.animalpolitico.com/2019/01/recorte-presupuesto-inegi-cancela-proyectos/>). [Reviewed on 29th April 2019].
2. Acevedo, Nicole, "Truth in numbers: Groups, government clash over accurate stats in Puerto Rico", *NBC News*, 23rd April 2018. (<https://www.nbcnews.com/storyline/puerto-rico-crisis/truth-numbers-groups-government-clash-over-accurate-stats-puerto-rico-n864416>). [Reviewed on 24th April 2019].
3. Aizenman, Anbar, Anisha Chinwalla and Benjamin A.T. Graham, "How does jailing the statisticians fix Greece's financial crisis? It doesn't.", Analysis, *The Washington Post*, 13 March 2017. (https://www.washingtonpost.com/news/monkey-cage/wp/2017/03/13/how-does-jailing-the-statisticians-fix-greeces-financial-crisis-it-doesnt/?noredirect=on&utm_term=.370d4ad832b7) [Reviewed on 24th April 2019].
4. Baer, Petteri, "The Odd Case Against Statisticians in Kazakhstan", Paper prepared for the Special Meeting on "NSOs Professional Independence: Threats and Responses", 18 September 2018. (<http://iaosi.org/index.php/latestnews/233-special-meeting-on-nsos>) [Reviewed on 24th April 2019].
5. Bloomberg's Editorial Board, "The Scandalous Persecution of a Greek Whistle-Blower", Opinion, Bloomberg, 4th August 2017. (<https://www.bloomberg.com/opinion/articles/2017-08-04/the-scandalous-persecution-of-greece-s-budget-whistle-blower>). [Reviewed on 24th April 2019].
6. Constitución Política de los Estados Unidos Mexicanos. Last amendment on 27-01-2016. (<http://www.diputados.gob.mx/LeyesBiblio/htm/1.htm>). [Reviewed on 26th April 2019].
7. International Association for Official Statistics (IAOS), "80 Former Chief Statisticians Condemn Prosecution of Andreas Georgiou", 18 June 2018.

- (<http://iaos-isi.org/index.php/latestnews/221-80-former-chief-statisticians-condemn-prosecution-of-andreas-georgiou>) [Reviewed on 24th April 2019].
8. ---, "IAOS one of the 47 Organisations Urging Puerto Rico's Governor not to Dismantle Statistical Agency", 16 May 2018a. (<http://iaos-isi.org/index.php/latestnews/220-iaos-one-of-47-organisations-urging-puerto-rico-s-governor-not-to-dismantle-statistical-agency>). [Reviewed on 26th April 2019].
 9. International Monetary Fund (IMF), "Press Release: Statement by the IMF Executive Board on Argentina", Press Release No. 12/319, 18th September 2012. (<https://www.imf.org/en/News/Articles/2015/09/14/01/49/pr12319>). [Reviewed on 24th April 2019].
 10. ---, "IMF Executive Board Removes Declaration of Censure on Argentina", Press Release No. 16/497, 9th November 2016. (<https://www.imf.org/en/News/Articles/2016/11/09/PR16497-Arentina-IMF-Executive-Board-Removes-Declaration-of-Censure>). [Reviewed on 24th April 2019].
 11. Ley del Sistema Nacional de Información Estadística y Geográfica. (<https://www.snieg.mx/contenidos/espanol/normatividad/marcojuridico/LSNI EG.pdf>). [Reviewed on 26th April 2019].
 12. Shah, Hetan, "How to save statistics from the menace of populism", *Financial Times*, 21 October 2018. (<https://www.ft.com/content/ca491f18-d383-11e8-9a3c-5d5eac8f1ab4>). [Reviewed on 24th April 2019].
 13. Statistics and Registration Service Act 2007, (<https://www.legislation.gov.uk/ukpga/2007/18/part/1>). [Reviewed on 24th April 2019].
 14. The Economist, "Called to account. The disturbing prosecution of Greece's chief statistician", Finance and Economics, *The Economist*, 3rd September 2016. (<https://www.economist.com/finance-and-economics/2016/09/03/called-to-account>). [Reviewed on 24th April 2019].
 15. ---, "Don't lie to me, Argentina", Leaders, *The Economist*, 20th June 2014, (<https://www.economist.com/leaders/2014/06/20/dont-lie-to-me-argentina>). [Reviewed on 24th April 2019].
 16. ---, "Welcome back. Argentina's new, honest inflation statistics", The Americas, *The Economist*, 25th May 2017. (<https://www.economist.com/the-americas/2017/05/25/argentinas-new-honest-inflation-statistics>). [Reviewed on 24th April 2019].
 17. The Globe and Mail, "Ending mandatory long-form census has hurt Canada", Editorial, *The Globe and Mail*, 6th November 2014. (<https://www.theglobeandmail.com/opinion/editorials/ending-mandatory-long-form-census-has-hurt-canada/article21486149/>). [Reviewed on 24th April 2019].
 18. Walker, Marcus, "Greece's Response to Its Resurgent Debt Crisis: Prosecute the Statistician", World-Europe, *The Wall Street Journal*, 6th February 2017.

(<https://www.wsj.com/articles/greeces-response-to-its-resurgent-debt-crisis-prosecute-the-statistician-1486396434>) [Reviewed on 24th April 2019].



The requirements for a well-functioning statistical system in a modern democratic society



Roeland Beerten

Flemish Statistics Authority, Statistics Flanders, Brussels, Belgium

Abstract

This paper provides a discussion of the characteristics of a modern democratic society, and which elements in a democratic system are necessary for official statistics to properly function in order to contribute to the quality of democratic decision-making. It will do this by first discussing which institutional conditions are necessary for an official statistical system to contribute to strengthening democratic decision-making, accountability and transparency. Secondly, it will look at how official statistics systems relate to the branches of democratic government and how these branches can shape the conditions for official statistics to function effectively and efficiently. It will analyse the conditions and requirements to achieve organisational and technical independence and the professional autonomy of statisticians, by giving some examples of how different countries organise their official statistics systems within government.

Keywords

Official statistics; government; democracy; institutional framework; professional independence

1. Statistics and democratic society

Etymologically, the word 'statistics' refers to the state. In the past, statistics were used by autocrats as 'the state's science' in order to exert, consolidate and expand their political and military power. The concept of a population count (census) has existed for more than two thousand years: it was the only way to find out how much manpower was available for an army, or how much tax could be collected (Pullinger, 2013).

With the arrival of the Enlightenment, a new interpretation of power took hold, which crushed the autocratic view of the state. Along came a shift in ideas about the goal of 'the state's science'. Statistics were no longer regarded as a measure of a state's strength and power, but rather as a measure of 'society's happiness'. For example, an early data collection at the end of the 18th century in Scotland was described as (Sinclair, J. (1798), as quoted in Pullinger (2013)): *"an inquiry into the state of a country for the purposes of ascertaining the quantum of happiness enjoyed by its inhabitants and the means of its future improvement"*.

With the rise of democratic state systems and, more recently, a stronger consumer perspective on the relationship between state and citizen, statistics themselves have also become more democratic. Figures are not only important for those holding a position of power, but for every element of a democratic system. Throughout time, statistics have evolved from a tool of government to a public good for society as a whole.

Today, the principle of official statistics being a public good is strongly endorsed by the international community at large. The United Nations, in their Fundamental Principles of Official statistics, state (United Nations, 2014): *“Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens’ entitlement to public information.”*

For Statistics Flanders the principle of official statistics being a public good is the first of the four strategic principles underpinning its work (Beerten, 2017): *“Official statistics serve society as a whole. That is why we not only take into account the government’s needs for statistical information, but also the needs of Flemish society at large. Although we are part of the Flemish Government, official statistics should be relevant and accessible to the broadest possible range of users: official statistics are a public good. Government policy should certainly indicate priorities, but it is not exclusively up to the government to determine which statistics should be developed, produced and published.”*

2. Conditions for an official statistics system to contribute to strengthening democracy

Any official statistics system is part of the machinery of government, and thus interacts with it, and depends on it. So which are the conditions to ensure that the official statistics system is able to produce numbers which meet the public good? The first chapter of the European Code of Practice gives a description of the institutional context and its conditions in which official statistics systems should operate (Eurostat 2017):

- Professional independence
- Coordination and cooperation
- Mandate for data collection and access to data
- Adequate resources
- Commitment to Quality
- Confidentiality
- Impartiality and objectivity

From this list it is clear some of the conditions are controlled by – and are thus the responsibility of – the institutional framework of government, but other conditions are rather characteristics of the official statistics system itself, and so are the responsibility of the official statistics system. The table below splits out the conditions between the two types of actors¹:

Table 1 – Institutional context factors by type of actor

Institutional framework	Official statistics system
Professional independence	Coordination and cooperation
Adequate resources	Quality
Mandate for data collection	Confidentiality
	Impartiality and objectivity

In what follows we will discuss the conditions of the institutional framework in some more detail; for the purpose of this paper we will not cover the dimensions which are largely the responsibility of the official statistics system. Of course this is not to say they are not important in supporting and shaping the democratic debate (see Bumpstead and Alldritt (2011) for a good discussion on relevant conditions which are largely within control of the statistics system itself, such as utility, accessibility and relevance).

2.1 Professional independence

A key condition for earning credible official statistics which meet the criteria for being a public good is the fact that they are produced in a professionally independent manner. Citizens should be able to have confidence in impartial figures that do not result from a political ideology, or from too close a connection with the policy that needs to be justified to the public. Moreover, even if there is no actual political interference with the production or dissemination of statistics, the perception of independence of the official statistics system should be strongly guaranteed. The trustworthiness of our official statistics is of a fragile nature and it can be easily jeopardised as a result of negative perceptions.

¹ The proposed split of responsibilities between the two types of actors is, of course, not entirely clear-cut, and it should be recognised some of the responsibilities will to some extent be shared between the two.

The principle of professional independence of official statistics is explicitly recognised by the international political community, although the principle is implemented in different ways (see section 3 below). Within the context of official statistics production by government, professional independence implies political independence. The mere suspicion of any political influence on the development, production or dissemination of official statistics is detrimental to the trust in and trustworthiness of the figures.

In practice, this means that decisions on the development, production and dissemination of official statistics should be taken by the official statistics producers and not by political or other administrative bodies. This professional independence reveals itself in purely professional choices regarding statistical methods, standards and procedures that are applied, and regarding the content and the moment of publication of statistical information.

For the content of statistical information, this means that the figures themselves should be produced by statistical experts in an independent manner. The task of statistical experts consists in describing the figures in an impartial way. Providing statistical commentary is an integral part of the job, as not everyone has sufficient knowledge of statistics to be able to read and interpret the figures. Moreover, the commentary should be descriptive, neutral and impartial (although it should be recognised this is an ideal to strive for rather than a gold standard which will always be achieved – see Spiegelhalter, 2019, p. 68). A clear distinction should be made between statistical comments formulated by statistical experts and political comments on the figures, made by politicians or policy experts. Statistical comments are produced independently of political interpretations and policy comments, and are clearly separated from them at the time of publication.

Decisions about when and how the figures and their statistical commentary are published also belong to the exclusive competence of the official statistics producers: political motives or newsworthiness cannot play any role in determining the moment of publication. The only determining principle can be that statistics have to be published as soon as possible after their collection, allowing for the necessary quality control.

Guaranteeing the professional independence of official statistics is therefore not just the responsibility of statistical experts or analysts. The importance of this independence should be recognised and pursued by anyone working within the government's political and administrative system. Only by building and maintaining a wide support base for professional independence can the trustworthiness of official statistics be realised and maintained.

2.2 Adequate resources

A second condition for an official statistic system to operate efficiently is the recognition of the need for adequate resources. This means there should be sufficient personnel and working budgets for the statistical service. Of course, in a context of competing demands within government official statistics needs to get their voice heard. It is therefore important that we - statistics producers - should not refrain from pointing out the value of our work. This means that we have to be able to prove the relevance of our statistics to policymakers within government who ultimately decide on the resources we work with. We have to be clear about the value and usefulness of official statistics for government and for society at large. We should present positive arguments for receiving sufficient resources in order to be able to develop and maintain a strong official statistics infrastructure.

However, the often hard-won resources available within an official statistics system should also be deployed in the most efficient way possible. In programming the work of an official statistics system, we should set clear priorities for what is needed, but we should also stop producing statistics that are not or hardly needed, in order to release funds for the production of new statistics with higher priority.

2.3 Mandate for data collection

The data sources that can be used for the production of official statistics have been changing in the past decades. Traditionally, official statistics have always strongly relied on censuses or surveys among persons, companies or institutions to generate data. This classic model of censuses and sample surveys is increasingly replaced by strategies in which two other data types play an increasingly important role in addition to censuses and surveys: administrative data held within government and - more recently - big data. This evolution is sometimes described as an evolution from 'designed data', in which the statistical expert can autonomously determine the data collection, towards 'organic data', in which statistical expertise contributes very little, and already existing data are relied upon (Groves, 2011).

This evolution has had significant consequences for the data availability for official statistics. With this shift the locus of data ownership has moved away from statistical offices to a range of diverse players outside the statistical office: other government departments or agencies, and increasingly private sector actors. In some countries recent legislation recognises this shift, and provides for compulsory access to these data by the statistical office (e.g. UK Statistics Authority, 2017).

3. The organisation of statistical systems in democratic societies

As explained before, official statistics systems are by default part of government. However, there is some debate – and difference in practice – as to which branch of government an official statistics system should belong to.

An in-depth review of governance models for official statistics is not within the scope of this paper, but it appears that in the majority of countries the official statistics system is part of the executive branch of government, with a reporting line to a minister. In some countries however, the statistical system reports to parliament; for example the UK Statistics Authority reports to a Statistics Board which is independent from the executive branch, and which reports to a parliamentary committee (see Laux, Alldritt and Young, 2008).

There is some considerable debate within the statistics community as to what the ideal model is in order to meet the principle of official statistics being a public good. Mostly these debates focus of the core issue of professional independence, and how it can be guaranteed. Georgiou (2018) argues that "in order to fully and sustainably meet the standards of professional independence, impartiality and objectivity of producers of official statistics, the production of such statistics should not be part of the executive branch of government – or any other branch of government – but a separate branch."

However, Tavernier (2018) argues that the presence of a statistical system within the executive branch of government does not inhibit an independent functioning or the production of statistics which serve the public good. He lists a number of advantages of the presence of the INSEE (and others statistics producers) within the executive branch in France, for example easier access to ministers and decision-makers who decide on funding for statistics, and the ease of making sure the statistics are relevant to and sensitive to policy needs.

It is clear there is no single right answer to these debates, and other issues will be of importance in determining the ideal setup, for example:

- the political culture
- the maturity of a particular democratic system
- the strength of the official statistics system itself
- the presence of a strong wider ecosystem around statistics

These factors will to some extent help determine the ideal model for the setup of the official statistics system within a specific country setting. However, it is clear the three conditions discussed in the previous section (professional independence, adequate funding and a mandate for data access) are key in the considerations for establishing an efficient statistical governance model.

References

1. Beerten, R. (2017) *Trustworthy numbers for a strong democracy. A strategy for Flemish official statistics*. Statistics Flanders.
2. Bumpstead, R, and Alldritt, R. (2011) *Statistics for the People? The role of official statistics in the democratic debate*. Paper for the 58th World Congress of the International Statistical Institute, Dublin, 21-26th August 2011.
3. Eurostat (2017) *European Statistics Code of Practice for the national and Community statistical authorities*
<https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142>
4. Georgiou, A. (2018) *The production of official statistics needs to be a separate branch of government*. Statistical Journal of the IAOS 34, 49–160, IOS Press.
5. Groves, R. (2011) *Three Eras Of Survey Research*, Public Opinion Quarterly, Vol. 75, No. 5, 2011, pp. 861–871.
6. Laux, R., Alldritt, R. and Young, R. (2008) *Independence for UK official statistics: the new UK Statistics Authority*. Paper for the Seminar of the Société Française de Statistique.
7. Pullinger, J. (2013) *Statistics making an impact. Presidential address to the Royal Statistical Society*, J.R. Statist. Soc. A, part 4, pp. 819-839.
8. Sinclair, J. (1798) *Statistical Account of Scotland*, vol. XX. Edinburgh: Creech.
https://archive.org/stream/statisticalacco01sincgoog/statisticalacco01sincgoog_djvu.txt
9. Spiegelhalter, D. (2019), *The Art of Statistics. Learning from Data*. London, Pelican Books.
10. Tavernier, J.-L. (2018) *A statistical system integrated in the national central administration: The French experience. A commentary paper to Andreas Georgiou's paper "The production of official statistics needs to be a separate branch of government"*. Statistical Journal of the IAOS 34, 161–163. IOS Press.
11. United Nations (2014) *United Nations Fundamental Principles of Official Statistics* <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>
12. UK Statistics Authority (2017) *Better Use of Data: Statistics and Research* <https://www.statisticsauthority.gov.uk/about-the-authority/better-useofdata-statistics-and-research/>



Statisticians misbehaving: The ethical dimensions of an essential profession



Alphonse L. MacDonald

General Bureau of Statistics (ABS), Paramaribo, Suriname

Abstract

In pre-modern, pre-democratic times the administration of societies was ruled by the will and whims of the "Sovereign". In the modern democratic society civil servants or public officials carry out their functions on the basis of explicit rules and regulations for the benefit of society as a whole. These rules and regulations are based on the principles of democracy, transparency and accountability. Statistics is a key element of the national information system and society expects statistics to be scientifically correct, of high quality and timely. To produce this requires the collaboration of the government, the statisticians and the citizens. Technical and operational independence of national statistical organisations are key requirements for a well-functioning statistical system. Legal provisions should exist to ensure that the national statistical organisations have the required degree of independence and the required, human and financial, resources. Similarly, the scientific basis of the profession requires that the statisticians adhere to the principles of science and apply established methodologies. Citizens have to provide the required information truthfully and as complete as possible. Occasionally, the Government or the citizens are not in agreement with the outcome and the results of the statistical operations and blame statisticians of not carrying out their duty properly. In the last decades there have been cases where statisticians were accused of un-professional and even criminal behaviour and have been disciplined, dismissed or taken to court. These cases have been widely reported in the media. The international statistical community has come to the defence of these colleagues with variable success. There are also instances in which statisticians commit errors, or deviate from established procedures producing substandard or even useless statistics. This undermines the confidence of the citizens in the statistical system and could have negative consequences for the standing of the country in the international community. To assist statisticians to correctly carry out their duties several professional and scientific statistical organisations have issued codes of conducts which focus on both technical requirements and ethical behaviour. However, these codes of conduct have a limited diffusion among statisticians. It is suggested to include exposure to the principals of ethical behaviour in the formal training of statisticians of all levels.

Keywords

Official statistics; National statistical organisation; Technical and operational independence; Codes of conduct; Ethical behaviour

1. Introduction

In pre-modern, pre-democratic times the administration of societies was ruled by the will and whims of the "Sovereign". Modern societies require valid and reliable statistical information based on the principles of democracy, transparency and accountability to function. Its production requires a national statistical organisation with the necessary legal provisions to allow it to function with technical and operational independence, free from governmental and societal interference. The information required can, in many cases, only be provided by the citizens, which they should do truthfully and as complete as possible. Statistic laws should guarantee that such individual information will be used for statistical purposes only. The statistics are common goods and should be made available to all users without distinction. The scientific basis of the profession requires that the statisticians are well trained and adhere to the principles of science and apply established methodologies. To assist them to correctly carry out their duties professional and statistical organisations have issued codes of conducts which focus on technical requirements and ethical behaviour. However, these codes have a limited diffusion among statisticians.

Not all governments enacted legislation to establish the technical and operational independence of the national statistical offices. Even in cases where legislation exists, occasionally government directly or indirectly, interfere with the technical and operational independence of the statistical office. In the last decades there have been cases where statisticians were accused of un-professional and even criminal behaviour (Chile and Greece) and have been disciplined, dismissed or taken to court. The international statistical community has come to the defence of these colleagues with variable success.

2. Development of modern statistics

In Europe, since the Enlightenment individuals with an interest in the advancement of knowledge of the society and science established "learned societies" in which topics of scientific and societal interest were discussed and numerical studies on population, social, economic and health phenomena carried out. Parallel to the compilation of numerical information, in mathematics new theories (probability, measurement and errors), methods and techniques, were devised which provided the basis for the emerging science of statistics. Given the political fragmentation of Europe and the great variety in measures and weights early statistical compilation used a wide

variety of methods, and the results of studies on the same subject were very often not comparable.

The 19th century was a period of rapid political, economic and social change. The transformation of authoritarian regimes into liberal democracies changed the nature of official statistics. They were no longer instruments for the government for policy formulation, but became means of verification of governmental compliance with policies by parliament, the electorate and ultimately by the population at large. In the first half of the 19th national statistical commissions, bringing together independent scholars and senior civil servants, were established, which led to the creation of statistical units in the ministries and eventually national statistical offices. The creation of statistical societies (the Royal Statistical Society of London, 1834 and the American Statistical Association, 1839) promoted the standardisation of methods and procedures. These efforts culminated in a series of International Statistical Conferences between 1853 and 1874, organised by Adolph Quetelet. In these conferences international standards for a wide range of statistics (census and civil registration methodology, classification of causes of death, transport statistics, etc.) were established. The International Statistical Institute (ISI) was created in 1885 in London. In USA early attempts were made to measure "public opinion".

The 20th century saw important developments in statistical theory and techniques, and the establishment of the principles of official statistics. In 1991 the Fundamental Principles of Official Statistics were approved by the Conference of European Statisticians. They established the principles regulating the relations between Governments and official statisticians, to ensure that the citizens are provided with valid, reliable and timely information to enable the functioning of a democratic society for common wellbeing and prosperity. See <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>. In 1992 they were adopted by the United Nations Economic Commission for Europe (UNECE) and in 1994 they were endorsed by the Statistical Commission of the United Nations and on 29 January 2014 by the General Assembly of the United Nations.

3. The Codes of Conduct

3.1 Introduction

National, international, official and voluntary statistical organisations have issued codes of conduct. In this paper the codes of selected voluntary national and international professional statistical organisations will be considered. The codes of conduct of two public opinion and market research organisations the American Association for Public Opinion Research (AAPOR) (<https://www.aapor.org/>) and the European Society for Opinion and Market Research (ESOMAR) (<https://www.esomar.org/>) will be reviewed. They are

chosen, because their activities are the most scrutinised statistical operations, as they are involved in election research. In addition the codes of conduct of three long established national and international statistical organisations the Royal Statistical Society (<https://www.rss.org.uk/>), the American Statistical Association (ASA) (<https://www.amstat.org/>) and the International Statistical Institute (ISI) (<https://www.isi-web.org/>) will be reviewed. They have well established records promoting the development of statistics and are at the forefront in the promotion and defence of the independence of statistics and statisticians.

3.2 The codes of AAPOR and ESOMAR

AAPOR was created in 1947, and currently is said to have in excess of 3,000 members based in the USA and worldwide. Since its inception AAPOR has been concerned with standards in public opinion research and the ethical behaviour of its members. AAPOR's first code was adopted in 1970 and subsequently revised and updated. The AAPOR membership was in general against the imposition of detailed technical standards and instead preferred to promote ethical behaviour and adherence to "the tradition of all sciences, [that] survey researchers should be required to describe adequately just what they did so that their findings could be objectively evaluated." Hollander (1992) These concerns and the minimum disclosure requirements are still reflected in the current (2015) AAPOR Code of professional ethics and practices.

(<https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx>)

ESOMAR started as a European organisation in 1948; at present it is a worldwide organisation with a membership of in excess of 5,000 professionals and 500 companies in 130+ countries. It concentrates on market research, but uses a broad definition of market research, which includes social and opinion research, and recently included "data analytics" as well. Since its establishment it has taken the public's confidence in survey activities as a key factor for the success of market research. In 1948 it published its first version of the Code of marketing and social research practice. Since 1976 it works closely with the International Chamber of Commerce (ICC) (<https://iccwbo.org/?s=esomar>) and they jointly establish and public their guidelines and codes. Their most recent code dates of 2016.

(<https://iccwbo.org/publication/icesomar-international-code-market-opinion-social-research-data-analytics/>)

3.3 The codes of RSS, ASA and the ISI

The RSS was established in 1834 in London. At present the RSS has a worldwide membership in excess of 9,000 individual members of different categories, from interested individuals without formal statistical training to

chartered and registered statisticians and scientists and graduated statisticians. RSS also has a number of corporate partners, which are government agencies. RSS has voluntary accreditation systems for statisticians. It promotes the use of statistics in policy formulation and decision-making, statistical literacy, and the development of statistics as a science and profession. It has developed a large number of guides including a series to assist judges, lawyers, forensic scientists and other expert witnesses in dealing with statistical evidence in the administration of criminal justice. RSS issued a code of conduct in 1993 which was revised in 2014. The code is mandatory for all professionally qualified members but is recommended to all the members.

(<http://www.rss.org.uk/Images/PDF/join-us/RSS-Code-of-Conduct-2014.pdf>)

ASA was established in 1839 in Boston. At present ASA has a global membership in excess of 18,000 persons with an interest in statistics and has a large network of US based chapters and sections. ASA also includes organisations as members. Since its inception ASA has promoted excellence in the development and application of statistics as a science, and the dissemination of statistical information to the benefit of the society. Recently ASA started a voluntary accreditation programme along the lines of the chartered statistician system of the RSS. ASA started discussions on ethical guidelines in 1949 and after a number of trials, the first version of the Ethical guidelines for statistical practice was formalised and published in 1989. It was revised in 1999 and the current version was approved in April 2018. ASA introduced the concept of the "ethical statistician" and postulates that "all practitioners of statistics, regardless of training and occupation or job title, have an obligation to work in a professional, competent, and ethical manner and to discourage any type of professional and scientific misconduct". Furthermore ASA states that "Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations." (<https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx?hkey=85085cd1-5dfc-4fb9-b526-e3c6d45abc0d>)

The ISI was established at the Jubilee session of the Royal Statistical Society of London in 1885, bringing together governmental and academic statisticians with the objective of promoting the development of administrative (official) and scientific statistics. It continued the organisation of international statistical congresses, through it continued the development of international standards and new methods and techniques. Members of the ISI first expressed their desire for the promulgation of an ISI declaration of professional ethics in 1979. The first ISI Declaration of professional ethics was issued in 1985 and the current revised and updated version was issued in 2010. ISI considered the Declaration to be applicable to all persons who are involved in or use statistics and statistical information. The Declaration is not a set of mandatory rules, but

rather a consolidation of shared values that present options for the obligations and responsibilities of statisticians towards the society, employers, clients, funders, colleagues and research subjects.

(<https://www.isi-web.org/index.php/activities/professional-ethics/isi-declaration>)

3.4 Review of findings

The five codes although different in style and organisation they cover the same principles and identify the same responsibilities researchers and statistician should abide by. AAPOR and ESOMAR use the term “researcher” as the subject of their codes, while RSS, ASA and the ISI explicitly refer to “statisticians”. The codes specify the obligations and responsibility towards the science, the profession, the society, the public, employers, funders, clients and research subjects, and maintenance and promotion of professional competence and integrity. All have disciplinary procedures to deal with professional misconduct and violations of the codes, including for non-members. All stress the adherence of the professional principles and recommend that researchers and statisticians do not exaggerate or exceed their technical capabilities, and only undertake activities for which they are qualified. All, except the code of the RSS, are self-regulatory, but claim that they should be valid for all social science surveys practitioners, in the case of AAPOR and ESOMAR, and applicable to all statisticians and statistical practitioners, in the case of ASA and ISI. The codes do not contain specific technical standards, but AAPOR places great emphasis on the standards of disclosures and gives detailed indications for the information that should be provided for the different types of studies and the ASA code contains a detailed discussion of the type of information the statistician should provide to safeguard the integrity of the procedures and the results. ASA includes specific sections on responsibilities towards colleagues, other statisticians and statistics practitioners, as well as a section on the responsibilities of employers, including organizations, individuals, attorneys, or other clients employing statistical practitioners. ESOMAR recommends that researchers “conform to the general accepted principle of fair competition”.

4. Statisticians misbehaving

All human activities are subject to mistakes and applying statistical methods and techniques are not exempted. Comprehensive information on violations of the codes is lacking. There is evidence of violations of methodological standards in research design, execution analysis and interpretation. Moreover, statistical procedures are incorrectly used by non-statisticians. There also cases of blatant misuse of statistical information for commercial or political reasons. Although statisticians are aware of

misapplication of established methodology, the public at large normally learns of the failure of statistics from media. Two examples will be presented to illustrate the issues involved in cases of alleged misconduct of statisticians.

Miscarriage of justice: Netherlands. On 24 March 2003 Lucia de Berk, a paediatric nurse, was convicted to life imprisonment accused of the murder of four patients and attempted murder of three others. The case was brought to the attention of the police by hospital authorities where she worked after it was alleged that a large number of incidents (deaths and resuscitations) took place during her shifts at the hospital. The probability of such occurrences by chance was considered very low (1 in 342 million). Her conviction was officially based on two non-natural causes certified by a medical expert, but in the media and the proceedings repeated reference was made to the low probability that the number of incidents could occur by chance. The calculations had been made by a court witness/specialist, a professor of psychology of law, who some 30 years before had obtained a Master's degree in mathematical statistics, but who had abandoned statistics and continued advanced studies in geography, psychology, economics and law. On 18 June 2004 she was convicted by a Court of Appeal to life imprisonment for the murder of 7 patients and the attempted murder of three others. The conviction was based on two cases for which the prosecution claimed to have sufficient proof that she had poisoned them. For the other cases no proof of her guilt was apparent, but by implication (chain-link reasoning) she was also held responsible because they also had occurred on her shifts. The case had considerable media exposure nationally and internationally, but there were considerable doubts about the soundness of the procedures and her conviction. A Committee of Support was created led by a Dutch philosopher of science and a medical practitioner who doubted the correctness of the medical and statistical evidence and the way these were used. Prominent Dutch and foreign statisticians also expressed serious reservations about the statistical calculations presented by the prosecution's statistical expert witness. Following a widely supported public petition, at a retrial in 2009/2010 Lucia de Berk was acquitted and in 2010 received an undisclosed amount of compensation from the State.

Although the retrial allegedly was decided purely on legal arguments observers had throughout stated that the process was incorrect because the data were flawed, they had been collected post-hoc; the calculations of the statistical prosecution witness/expert were incorrect; the Court committed a prosecutor's fallacy, and the same team of experts was used to establish suspicion, measure the magnitude of the suspicion and acted as expert witnesses in the court. Gill (2006) Derksen (2007), Meester (2007)

Incorrect census methodology, 2012 population census. Chile is one of the Latin American countries with an excellent statistical organisation and

tradition. Since its independence in the early 19th century Chile systematically carried out national population censuses, from 1835 and since then roughly every ten years. The rate of under-enumeration of the Chilean census during the period 1950 – 1990 in general was below the regional average. But in the 2000 census round the under-enumeration in Chile was 3.8 % while the regional average 3.3 %. Up to the 2000 census round the census in Chile was carried out in one day by a large group of volunteers on the basis of the de facto concept, using a basic questionnaire. For the 2010 census round this was changed to a census carried out by paid enumerators, using the de jure concept, with a complex questionnaire and a multi-week field enumeration period.

Fieldwork for the 2012 census was carried out from 9 April to the end July 2012; the results of the census were released on 2 April 2013; on 26 April 2013 the Director of the National Statistical Office (INE) resigned, to be replaced on 29 April 2013, and on 2 May 2013 the results of the 2012 were withdrawn. The resignation of the Director and the withdrawal of the published census results were caused by the allegations by senior INE staff members that the data had been manipulated on the instructions of the Director. Bezama y González, M. (2013), *Economist* (2013)

Two review commissions, one national and the other international, were appointed to investigate the allegations and assess the quality of the census. Their findings were as follows: The census methodology was not correctly applied due to a dysfunctional atmosphere in INE, causing the marginalisation of knowledgeable technical staff; lack of understanding of the de jure concept, insufficient time for the preparation and training of staff, and lack of adequate funding. The census did not cover about 9.3 % of the population, either by not contacting the households, or because dwellings were wrongly classified as unoccupied, or because of refusals. The missing data were imputed. The non-coverage could be higher in particular geographical areas, administrative regions, communities, or in specific social categories. Coverage of the census population could not be properly assessed because the Post Enumeration Survey (PES) was not properly designed and executed and was only attempted in urban areas. Training of the field staff was uneven, and inadequate. The census reference date was not clearly identified or used. The documentation on the census procedures was defective; there were no methodological or administrative reports. (INE (2014))

5. Conclusions and way forward

“Statisticians misbehaving” appear to be rare events. Since the second half of the 20th century the scope of statistics has greatly expanded and currently it is fashionable to refer to “data and statistics”. The term “statistician” not only refers to persons with formal degrees in statistics, but also to persons

who apply statistical methods and technique in a wide variety of disciplines. Training of "statisticians" has been adjusted to accommodate the increasing number of students, due to the demand for specialists to handle emerging new sources of data. The value systems of societies have been modified with the increased competitiveness in the labour market, emphasis on production and productivity and the surge and spread of utilitarian principals. Also, in spite of the generalised use of technology in society, there is an increasing anti-science attitude among populations and political leaders of developed nations.

The professional organisations have programmes of divulging their codes and the benefit of their application, reaching existing statistical practitioners. To ensure that all future statisticians and statistical practitioners behave professionally and ethically, it is necessary that they are properly trained in basic scientific methodology and the principles of ethical behaviour of the profession.

References

1. Bezama, B. y González, M. (2013) El director del INE habría manipulado las cifras del Censo 2012. (<https://ciperchile.cl/2013/04/25/el-director-del-ine-habria-manipulado-las-cifras-del-censo-2012/>)
2. Derksen, T. (2007). The Fabrication of Facts: the Lure of the Incredible Coincidence. (Available at <http://luciadeb.nl/english/archives/fabricationoffacts.html>)
3. Economist, The (2013). Statistics in Chile. How many Chileans? <http://www.economist.com/blogs/americasview/2013/04/statistics-chile>
4. Gill, R. (2006) Lying Statistics Damn Nurse Lucia de B. (Available at: <https://www.math.leidenuniv.nl/~gill/lucia.html>)
5. Hollander, S. (1992). Survey standards, in: Sheatsley, P.B, and Mitofsky, W.J. (eds). A meeting place: The history of the American Association for Public Opinion Research. American Association of Public Opinion Research: 65 – 103.
6. Instituto Nacional de Estadística (INE). The 2012 population and housing census of Chile. United Nations, New York. (https://unstats.un.org/unsd/statcom/statcom_2014/seminars/2010_Census_Decade/Presentacion%20-%20Chile.pdf)
7. MeesterR, R. et.al. On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Law, Probability and Risk*. (Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.926.3867&rep=rep1&type=pdf>)



Safeguards for the professional independence of Statisticians in Europe



Pilar Martin-Guzman

Universidad Autónoma de Madrid, Madrid (Spain)

Abstract

The professional independence of statisticians is an essential component for the healthy functioning of a democratic system. Political interferences in the production of official statistics will undermine its credibility, so making them useless for the main purpose for which they are produced: providing impartial information for policy decision making. The European Statistical System, ESS, (consisting on EUROSTAT and the National Statistical Offices of the Member States) is aware of the risks inherent to political interferences. For that reason, a number of instruments have been devised for the preservation of independence in the statistical production. Legislation is a fundamental instrument. The ESS presents to the European Parliament and the Council regulations that, when approved, are of compulsory application in all Member States. The Amendment of Regulation 223 dealing, among other issues, with the appointment and removal of the Chief Statistician, is an example to be mentioned. Another useful instrument is the European Statistics Code of Practice, based on the UN Fundamental Principles of Official Statistics, and in which the professional independence plays a prominent role. This Code has been the precedent and inspiration for several Statistics Codes of Practice adopted in other regions, such as ECLAC and the South Mediterranean countries. In order to check the correct implementation of this Code a system of peer reviews, compulsory for all Member States, and also adopted by EFTA countries, is being organized roughly every five years. Then, there is a European Statistical Governance Advisory Board, ESGAB, in charge of providing an independent overview of the ESS as regards the implementation of the Code of Practice, and enhancing the professional independence, integrity and accountability of the ESS. The effectiveness of these instruments will be discussed. Still, the EU is a very complex supranational organization, embodying countries with very different traditions and administration systems, and working on the principle of subsidiarity. That explains that, in spite of all these safeguards, problems eventually arise. Finally, there are a number of countries geographically located in Europe that are neither members of the E.U, nor of EFTA, For several of these countries, Georgia, Armenia, Moldova, Azerbaijan, Belarus and Ukraine, Global Assessments are conducted with the financial and technical support of EUROSTAT.

Keywords

E.U legislation; code of practice; ESGAB; peer reviews

1. Introduction

The professional independence of statisticians is an essential component for the healthy functioning of a democratic system. Political interferences in the production of official statistics will undermine its credibility, so making them useless for the main purpose for which they are produced: providing impartial information for policy decision making.

Official statistics are a public good, to be used not only by governments, but also by social forces, financial market participants, companies, citizens and many other stakeholders. Accountability and transparency, as well as the professional integrity and the compliance with ethical principles and good practices have to be ensured.

The European Union has been aware, for a long time now, of the risks inherent to political interferences, even in democratic regimes. Already the introductory background of the Fundamental Principles of Official Statistics adopted by the Conference of European Statisticians, UN, in 1991 specifies that “the need for a set of principles governing official statistics became apparent at the end of the 1980s when countries in Central Europe began to change from centrally planned economies to market-oriented democracies. It was essential to ensure that national statistical systems in such countries would be able to produce appropriate and reliable data that adhered to certain professional and scientific standards”.

For that reason, a number of instruments have been devised for the preservation of independence in the statistical production. Although they have been produced by the European Statistical System (ESS) to be used within the realm of the European Economic Area (EEA), some of them are being employed in other European countries, and also in other Regions, mainly through the support of cooperation programs. Therefore, they have now a worldwide use and recognition. Some of them refer to legislation, or to compilation of good practices, while others consist on monitoring activities or on the performance of specific advisory bodies.

This paper intends to review these instruments and their effective implementation

2. The European Statistical System

The ESS is the body in charge of the design and application of these instruments. It is defined as a partnership between EUROSTAT and the national statistical institutes or other national authorities in each EU Member State responsible for developing, producing and disseminating European Statistics.

It functions as a network, with EUROSTAT playing a leading role in harmonizing statistics in cooperation with national statistical authorities.

The roots of the ESS can be found in the Statistical Service of the Coal and Steel Community, created in 1952 with the task to harmonize statistics that were already available at national level and had been collected for national purposes. A long journey has been travelled since. In 1992 the European Economic Area was created with the aim of extending the EU single market to non EU countries, in particular, to the European Free Trade Association (EFTA) countries. The foundational Agreement of EEA establishes the need to produce and disseminate comparable statistics for describing and monitoring all relevant economic, social and environmental aspects of the EEA and to enact statistical legislation similar to that passed in the EU.

Three out of the four current EFTA countries (Iceland, Lichtenstein and Norway) have joined the EEA. The fourth, Switzerland, although not being part of the EEA, has signed in 2007 a number of bilateral agreements with the EU, concerning access to and harmonization of statistics. As a result, representatives of the four EFTA countries attend the meetings of the EES with participation as active as the representatives of any of the 28 UE Member States.

Moreover, Albania, the Republic of North Macedonia, Montenegro, Serbia and Turkey are candidate countries, and are actively involved in adjusting their national statistical systems to the rules and practices of the ESS, so that they benefit from the application of some of the safeguards that will be described.

3. EU legislation as a safeguard

Legislation is a fundamental instrument for preserving independence in the statistical production. The EU is a very complex supranational organization, embodying countries of very different sizes and with very different traditions, practices and administration systems. It works on the principle of subsidiarity, which rules out EU intervention when an issue can be dealt with effectively by Member States, but specifies that the EU is justified in exercising its powers when Member States are unable to achieve the objectives of a proposed action satisfactorily and added value can be provided if the action is carried out at EU level.

Although Member States have national Statistical Acts guaranteeing functional independence, it sometimes happens that national legislation is not sufficiently specific in describing the conditions under which this independence has to be effectively implemented, and eventually some problems have arisen. Therefore the ESS comes to the rescue by providing more effective legislation at EU level. When deemed necessary, the ESS presents to the European Parliament and the Council regulations that, once approved, are of compulsory application in all Member States.

A good example of this safeguarding exercise has been the Amendment of Regulation 223, published in April 2015. It deals, among other issues, with two crucial points for the safeguard of the independence of statisticians: the appointment and removal of the Chief Statistician and the coordination of national statistical systems as a protection against political interference.

As it is now, practically all Member States specify in their legislation that the appointment of the Chief Statistician should be made on the basis of professional criteria, academic degree, career record or management expertise. But these requirements are in some cases rather vague. A fixed term of office for the Head of NSIs, generally with a limited renewability, is established in most Member States, but not in all of them. Moreover, not all countries enumerate in their regulations possible reasons for his/her dismissal, which obviously leaves a door open for political interference. And direct appointment by the government is still the norm in some of them, thus setting some restrictions to the application of the principle of equal opportunities.

Article 5a (4) of the Amendment to Regulation 223 specifies that "Member States shall ensure that the procedures for the recruitment and appointment of Heads of NSI's and, where appropriate, statistical Heads of ONAs producing European statistics are transparent and based only on professional criteria. These procedures shall ensure that the principle of equal opportunities is respected, in particular with regard to gender. The reasons for dismissal of heads of NSIs or their transfer to another position shall not compromise statistical independence". Now all Member States will have to adapt their current legislation in order to comply with these requirements.

The risks of political interference are higher in Other National Authorities (ONA's) producing official statistics. In many cases, the responsible for the production of European statistics in these ONA's are heads of units in ministries, and they often assume some other responsibilities too. In these cases the producers of statistics can be induced to follow the rules and practices of the corresponding ministry, which do not necessarily comply with the norms established by the statistical legislation.

The ESS has been fully conscious of this problem. The Amendment to Regulation 223, in article 5 (1), establishes that the National Statistical Authority in each Member State shall be the sole contact point for the Commission (EUROSTAT) on statistical matters and will assume the responsibility for coordinating all activities at national level for the development, production and dissemination of European statistics, determined in the statistical program -with inclusion of the statistical activity carried out by ONA's. This is meant to extend the usual good practices of NSI's to ONA's and provide a better safeguard for the independence of all producers of official statistics.

4. The European Statistics Code of Practice

A good legislation frame is a necessary, but not sufficient, condition for the preservation of independence. Standards for developing, producing and disseminating statistics have to be set for the appropriate application of the law. To this end, the Statistical Programme Committee of the ESS produced in 2005 a Code of Practice, based in the Fundamental Principles of Official Statistics, and in which the Principle of Independence plays a relevant role. This Code has been updated in 2011 and in 2017.

In its current version the Code of Practice includes 16 key principles (one of which, referred to Coordination, was included in the last update) for the production and dissemination of European official statistics and the institutional environment under which national and Community statistical authorities operate. A very positive feature of it is that each principle is accompanied by a set of indicators, aimed as a reference for monitoring compliance with the principle.

The Code of Practice has proved to be a most useful benchmarking instrument, and has, therefore, received worldwide attention. It has been the starting point and the model for several similar Codes of Practice adopted in other regions of the world, such as ECLAC, South Mediterranean countries and Afristat.

5. Advisory bodies: ESGAB

Once the Code of Practice had been adopted, it was deemed advisable to implement ways to overview and promote compliance with it by Member States. To this end, an advisory body and some monitoring techniques have been designed.

The European Statistical Governance Advisory Board was established in 2008 with the aim to boost the professional independence, integrity and accountability of the ESS. The members of the Board are seven independent high-qualified statisticians, appointed by the European Parliament and the Council for a fixed period of time, with the assignment to provide an independent overview of ESS as regards the implementation of the Code of Practice.

The main tasks of ESGAB are: a) to prepare an annual report, to be presented to the European Parliament and the Council, on the effective implementation of the Code of Practice in the EU, as well as the foreseeable risks that could arise from new technical or social developments, and b) to advise EUROSTAT on appropriate measures for the implementation of the Code of Practice and on its updating. ESGAB acts as a warning watchdog on matters concerning the compliance with the Code of Practice in the ESS.

6. Monitoring activities: The peer reviews

In order to check, and to improve, the correct implementation of this Code a system of peer reviews, compulsory for all Member States, and also adopted by EFTA countries, is being organized roughly every five years. Two rounds have been implemented up to now, in 2006-2008 and in 2014-2015, and a third one is under preparation.

The peer reviews are used as an instrument of benchmarking and monitoring on the basis of the explanatory indicators added to each principle of the Code of Practice. They are carried out by teams of three persons, comprising two high-level experts from NSI's and one from EUROSTAT. The process starts with the countries filling a self-assessment questionnaire, which when completed is sent, together with some complementary information, to the team for preparation of the interviews. Then a series of interviews, lasting from three to five days, take place in the country, including the top-management of the NSI, heads of main units, and providers of information from administrative sources, as well as a selection of users, such as government, the media, enterprises, social forces, academics and researchers. The second wave of peer reviews covered not only the NSI but also the ONA's in the Member State.

As a result, the peer review team writes a report that is made public in the EUROSTAT website. Also, where full compliance with the Code has not yet been achieved, NSI's, in agreement with the reviewers, identify improvement actions and indicate a timetable. The list of improvement actions and deadlines and its fulfilment are monitored by EUROSTAT.

The peer reviews introduce an external element in the implementation of the Code of Practice that contributes to the transparency of the process. The diversity of cultures and administrative organizations in Europe contributes to the cross-fertilization process that derives from being monitored by external eyes but from a peer's perspective, and this is a source of added value to the exercise

This practice has been extended to candidate countries, and also to other European countries, such as Armenia (2014 and 2019), Azerbaijan (2010 and 2017), Belarus (2013), Georgia (2013), Moldova (2013) and Ukraine (2012), through cooperation programs with ENP countries financed by the EU. Peer reviews following the ESS pattern were also conducted in eight countries of the ECLAC region (Colombia, Ecuador, Dominican Republic, Jamaica, Panama, Paraguay, Peru and Uruguay) in 2014-15, as part of a project financed by the Inter-American Development Bank, but in this case the reports have not been made public.

7. Conclusions

Since the adoption of the Fundamental Principles of Official Statistics the ESS has been designing several instruments for the safeguard of the independence of statisticians that have proved to be effective, some of which have been extensively used in other regions of the world. But technological development and social changes are a permanent source of new challenges that require its frequent updating.

References

1. Fundamental Principles of Official Statistics, United Nations Statistical Commission (1991)
2. Regulation (EU) 2015/759 of the European Parliament and the Council amending Regulation (EC) 223/2009 on European Statistics (Text with relevance to the EEA and Switzerland)
3. Agreement on the European Economic Area, part V, Horizontal Provisions, Art 76 (1992)



Measuring the digital economy: Malaysia's experience



Hasnah Mat, Norul Anisa Abu Safran, Mazreha Ya'akub
Department of Statistics Malaysia

Abstract

In 2019, Ministry of Communications and Multimedia Malaysia (MCMC) has stated that the policy to empower digital economy is being formulated and is expected to be announced in the third quarter of this year. A program based on strategic trusts to advance Malaysia towards a developed digital economy by 2020 was established by Malaysia Digital Economy Corporation (MDEC) and other relevance agencies since year 2012. The aim is to build an ecosystem that promotes the extensive use of Information Communication Technology (ICT) in all aspects of the economy to create communities connected globally and interacting in real-time. In an effort to build a vibrant digital economy for Malaysia, reliable and comprehensive statistics that are internationally comparable is essential to monitor the progress towards achieving digital economy and formation of better policies. Thus, Department of Statistics, Malaysia (DOSM) has took the initiative to produce the significance statistics through the surveys on Usage of ICT and e-commerce by establishment (ICTEC), ICT Use and Access by Individuals and Households survey (ICTHS) and Information & Communication survey (AES-ICT). Therefore, this paper aims to share the compilation of digital economy statistics and its contribution to Malaysia. Generally, in 2017 contribution of ICT to the national economy steadily grew by 10.3 per cent from 2016 which registered RM247.1 billion. ICT industry also recorded RM68.2 billion for compensation of employees with a share of 38.2 per cent to GDP. While the value added of e-commerce increased to RM85.8 billion as compared to RM75.0 billion in 2016.

Keywords

Digital economy, ICT, e-commerce

1. Introduction

Digital economy and the broad application of ICT nowadays have driven the borderless world to spark rapid competition and dissemination of information through the virtual world. This transformation has brought many changes in the society and has affected the pattern of human life. The spreading and sharing of information around the world become easier with ICT technology. The importance of the digital economy has attracted

enthusiasm of economist and policy makers to quantify the enormous consequence to the economy.

The ICT revolution will have a great impact on all aspects of growth, equity and governance for countries at all levels of development. Technical advances in many ICT areas continue apace and could level or change the playing field for developing countries, provided policy and institutional changes are made to capitalize on these advances (Hanna, 2011).

The government of Malaysia is committed in governing the development of Malaysia's ICT sector. In 1996, Malaysia launched the Multimedia Super Corridor (MSC) initiative in which it was envisioned that the ICT sector would play an essential role in transforming this country into a developed economy by 2020. As a result, Malaysia has enacted the outline of Legal Framework and policies that lead to the increase of advanced telecommunications infrastructure and higher broadband penetration. Malaysia has a comprehensive legal framework relating to the use of the internet and digital technology in which to ensure the advantages of digital economy could benefit the users including the industry players. Moreover, e-commerce has been growing in Malaysia due to the supportive policy environment and the advanced infrastructure.

Technology has changed the interaction and function of a society or community. The current advances technology has indirectly caused wide and great change in every forms of human activities in the world (Carayannis, Campbell, and Rehman; 2015).

The World Bank Group (2018) reported that from 2010-2016, the digital economy of Malaysia grew by 9 percent per year in value-added terms, faster than overall GDP. It is expected that the growth will reach 20 percent by 2020. E-commerce is expanding very fast, and expected to surpass RM110 billion, where it will reach 40 percent of the digital economy by 2020.

The compilation of digital economy statistics in Malaysia covers industry players and household. To gather this information, Malaysia conducts a few surveys namely ICTEC, ICTHS and AES-ICT. The main output of these surveys is for the compilation of statistics on ICT Satellite Account (ICTSA) for Malaysia. Eventually, digital economy statistics is produced.

2. Methodology

Generally digital economy could be referred as a wide range of economic activities that use digitised information and knowledge to deliver the production. The internet, big data, cloud computing, cryptocurrency and other new digital technologies are used to collect, store, analyze, and share information digitally and transform social interactions.

In Malaysia, a command understanding among agencies on e-commerce is a form of transaction either the sale or purchase of goods or services, through

a network of computers that have been designed for this purpose. E-commerce transactions can occur between enterprises, households, individuals, governments and public or private organization to another. It is also includes orders placed on websites, extranet or Electronic Data Interchange (EDI).

The classification of ICT industry is based on Malaysia Standard Industrial Classification (MSIC) 2008 Ver.1.0. which is in concordance with International Standard Industrial Classification of All Economic Activities (ISIC) Rev. 4. ICT consists of industries such as manufacturing, trade, services and content & media industries. The classification of ICT products is based on Malaysian Classification of Products by Activity (MCPA) 2009 which conforms with Central Products Classification (CPC) Ver. 2. ICT industry includes ICT services and Non ICT services as well.

Data on Usage of ICT and e-commerce by establishment are obtained from the Survey of ICTEC 2018 for reference year 2017. It covered establishments identified in all economic sectors classified under the MSIC 2008 Ver. 1.0. The concepts and guidelines are obtained from the Organisation for Economic Co-operation and Development (OECD) Model Survey on ICT Usage by Businesses (2nd Revision), 2015 published by OECD. While the sampling design of the survey was a one-stage stratified random sampling with a total of 57,194 establishments.

Information on household perspective was obtained from ICTHS 2017. Among the objectives of this survey is to calculate ICT indicators to measure development of national ICT and digital economy. The concept and definition of ICTHS 2017 is based on the Manual for Measuring ICT Access and Use by Households and Individuals 2014 Edition published by the International Telecommunication Union (ITU). Two-stage stratified sampling design was adopted with sample size of 25,909 living quarters. Living quarters are defined as independent and separate structures, are usually used as place of abode.

In Annual Economic Survey 2018, data on information and communication services were collected for reference year 2017. It covered 35 industries at 5-digit level under Section J¹ and the definition adopted is based on the recommendations of the MSIC 2008 Ver. 1.0. One-stage stratified random sampling is applied in this survey with a total of 8,699 establishments.

¹ Section J is referring to Information and Communication which comprises of activities for Publishing; Motion picture, video & television programme production, sound recording & music publishing; Programming & broadcasting; Telecommunication services; Computer programming, consultancy & related activities; and Information services.

3. Results

The main findings from those surveys are as follows:-

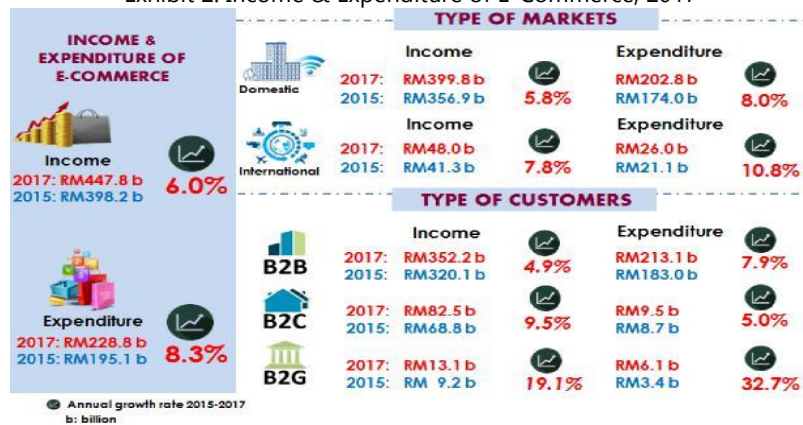
a. CTEC 2017

Based on survey of ICTEC year 2017, 78.9 per cent of establishments were used computers. Meanwhile, 73.3 per cent establishments were used internet and 37.8 per cent had web presence in their business (Exhibit 1).

Exhibit 1: Usage of Computer, Internet and Web Presence, 2017



Exhibit 2: Income & Expenditure of E-Commerce, 2017



Usage of intranet, extranet and local area network (LAN) was dominated by manufacturing sector at 47.1 per cent, 37.6 per cent and 64.7 per cent respectively. For wireless local area network (WLAN), services sector recorded the highest usage at 38.0 per cent, while for wide area network (WAN), Mining & quarrying sector recorded the highest usage of 35.1 per cent. In overall, the highest computer network infrastructure used by businesses was LAN with 55.1 per cent, followed by WLAN (35.7%) and WAN (30.0%). Meanwhile internet access via fixed broadband recorded the highest with 80.5 per cent, followed by mobile broadband (37.0%) and narrowband (2.4%).

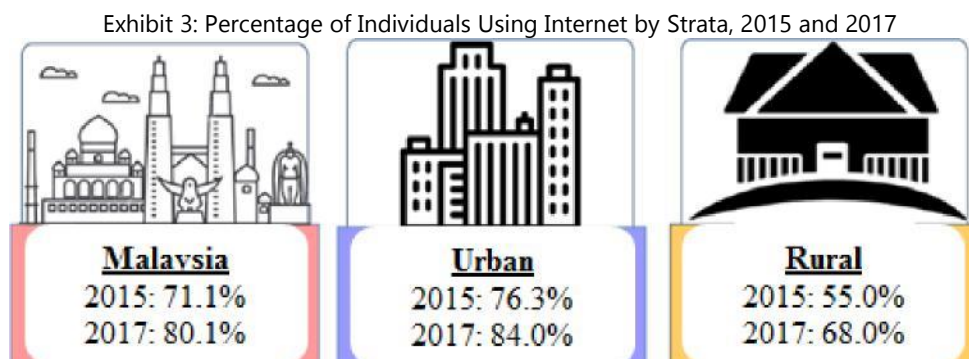
Sending or receiving email recorded the highest percentage for the purpose of internet usage of 92.1 per cent. This was followed by internet banking, 70.9 per cent and for obtaining the information of goods and services, 67.3 per cent.

The e-commerce income amounted to RM447.8 billion accounted for 15.9 per cent from the total gross output of establishments (Exhibit 2). Manufacturing sector was the main contributor with RM287.5 billion with a

share of 64.2 per cent. The income of e-commerce in Malaysia was led by domestic market with RM399.8 billion (89.3%) as compared to international market of RM48.0 billion (10.7%). The e-commerce expenditure recorded RM228.8 billion accounted for 13.5 per cent from the value of intermediate input of establishment in 2017. It was recorded RM202.8 billion in domestic market and RM26.0 billion in international market. In 2017, business to business (B2B) recorded the highest e-commerce expenditure of RM213.1 billion or 93.2 per cent. This was followed by business to consumer (B2C) (RM9.5 billion; 4.2%) and business to government (B2G) (RM6.0 billion; 2.6%).

b. ICTHS 2017

The report showed that percentage of individuals using internet increased by 9.0 percentage points from 71.1 per cent in 2015 to 80.1 per cent in 2017. Internet usage in urban area showed an increase of 7.7 percentage points to 84.0 per cent in 2017 from 76.3 per cent in 2015. The internet usage in rural area also rose to 68.0 per cent in 2017 from 55.0 per cent in 2015 as shown in Exhibit 3.



In 2017, among the internet users, three main locations of internet usage were at home (94.7%), various places via mobile devices (90.6%) and work place (50.0%). Among the popular internet activities carried out by the internet user were participating in social networks (86.3%), followed by downloading images, films, video or music; playing or downloading games (81.2%), getting information about goods and services (80.4%), downloading software or applications (74.5%) and sending or receiving e-mails (70.4%). Other activities carried out by internet user were internet banking (37.6%) and purchasing or ordering goods or services (e-commerce) (23.2%).

Percentage of individuals using computer increased 1.1 percentage points from 68.7 per cent in 2015 to 69.8 per cent in 2017. The percentage of individuals using computer in urban area increased by 1.1 percentage points to 75.0 per cent in 2017 from 73.9 per cent in 2015. Similarly, the percentage of individuals using computer in rural area increased by 0.6 percentage points

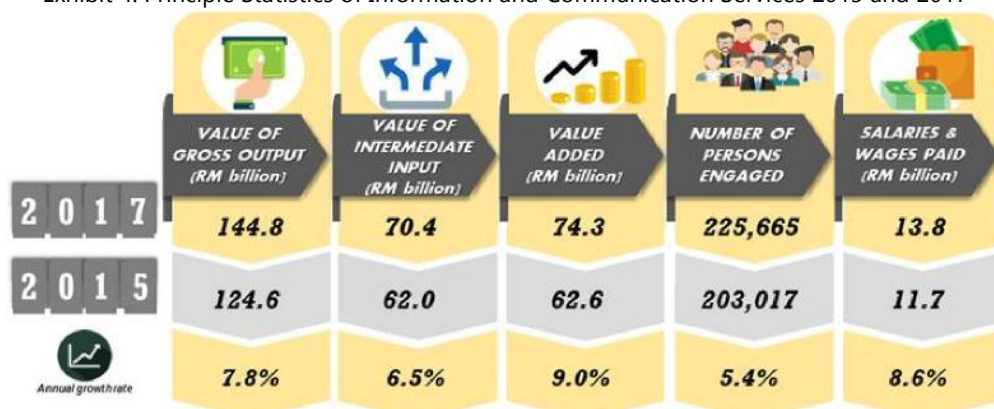
from 52.6 per cent in 2015 to 53.2 per cent in 2017. The main computer activities carried out by the computer users were copying or moving a file or folder (79.5%), using copy and paste tools to duplicate or move information within a document (79.1%), connecting and installing new devices (69.3%) and sending e-mails with attached files (61.7%).

Percentage of individuals using mobile phone in 2017 increased by 0.2 percentage points to 97.7 per cent in 2017 as compared to 97.5 per cent in 2015. Percentage of household access to internet increased 15.6 percentage points to 85.7 per cent in 2017 while household access to computer also increased by 6.5 percentage points from 67.6 per cent in 2015. Percentage of households using mobile broadband in 2017 increased by 22.3 percentage points to 84.2 per cent. Similarly, percentage of households using fixed broadband in 2017 was 29.1 per cent increased by 4.4 percentage points as compared to 24.7 per cent in 2015.

c. AES-ICT and ICTSA 2017

The information and communication services recorded gross output value of RM144.8 billion in 2017 as compared to RM124.6 billion in 2015 with the annual growth rate of 7.8 per cent. The value of intermediate input also increased by RM8.4 billion to record RM70.4 billion with the average annual growth rate of 6.5 per cent, thus resulting a value added of RM74.3 billion for the year 2017. The number of persons engaged in this sector also reported an increase of 5.4 per cent to 225,665 persons as compared to 203,017 persons in 2015. Meanwhile, the salaries & wages paid in 2017 amounted to RM13.8 billion compared to RM11.7 billion in 2015 (Exhibit 4).

Exhibit 4: Principle Statistics of Information and Communication Services 2015 and 2017

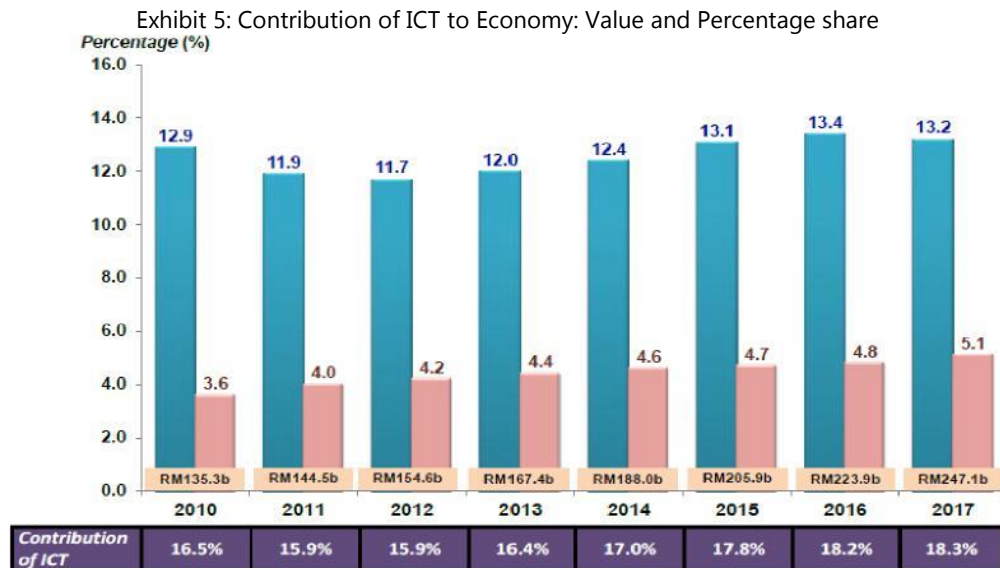


Telecommunications services was the largest contributor of gross output value with RM87.4 billion (60.4%) and the second largest contributor was computer programming, consultancy and related activities with RM35.1 billion (24.2%). In term of value added, telecommunications services also recorded the highest value added in 2017 which amounted to RM49.5 billion (2015:

RM38.3 billion). This was followed by the computer programming, consultancy and related activities of RM14.4 billion (2015: RM14.3 billion).

Computer programming, consultancy and related activities registered the highest number of persons engaged of 111,896 persons or 49.6 per cent (2015: 48.4%). The second highest contributor was telecommunications services with 58,163 persons or 25.8 per cent (2015: 25.4%) followed by publishing activity with 17,293 persons or 7.7 per cent (2015: 8.1%).

The total salaries & wages paid in information and communication services was top in computer programming, consultancy and related of RM6.1 billion and followed by telecommunications services with RM4.7 billion.



On overall, in 2017 contribution of ICT to the national economy continued to expand at RM247.1 billion registering a growth of 10.3 per cent (2016: 8.7%). ICT contributed 18.3 per cent to GDP comprising of ICTGDP (13.2%) and e-commerce for non ICT industries (5.1%). The total value of ICTGDP increased by RM13.9 billion amounting to RM178.2 billion and grew by 8.4 per cent (2016: 8.2%). The growth was impelled by ICT services industry with a share of 40.5 per cent followed by ICT manufacturing industry 36.1 per cent (Exhibit 5).

In 2017, compensation of employees for the ICT industry registered RM68.2 billion with a share of 38.2 per cent to GDP. The gross operating surplus recorded a share of 57.2 per cent followed by taxes less subsidies on production and imports, 4.6 per cent. ICT industry employed 1.09 million persons and contributed 7.6 per cent to the total employment.

The value added of e-commerce registered an increase to RM85.8 billion as compared to RM75.0 billion in 2016. E-commerce recorded a growth of 14.3 per cent led by non ICT industry with a share of 80.3 per cent. The contribution

of e-commerce to GDP recorded 6.3 per cent attributed by e-commerce for non ICT industry 5.1 per cent and e-commerce for ICT industry 1.2 per cent.

4. Conclusion

Basically, the ICT contributions can be obtained or derived from the statistics/breakdown of Gross Domestic Products. However, the comprehensive information and overall performance of digital economy can be gained through extraction of the three surveys compilation in perspective of industry player and household contribution. Besides the ICT contribution, value of e-commerce, information regarding usage of ICT among businesses and household can be obtained from this paper.

The contribution of digital economy in Malaysia provides details and comprehensive information on the specific economic activity and benefit to the people in which facilitates the policy makers, researchers, industry players and citizens to monitor, measure and formulating sound policies for the transformation digital programmes. These data are significant to measure the performance of overall digital economy in Malaysia.

References

1. Carayannis, E. G., Campbell, D. F., & Rehman, S. S. (2015). "Happy accidents": Innovation-driven opportunities and perspectives for development in the knowledge economy. *Journal of Innovation and Entrepreneurship*, 4(1). doi:10.1186/s13731-015-0021-9
2. Department of Statistics Malaysia (2017). Information and Communication Technology Satellite Account 2017
3. Department of Statistics Malaysia (2018). Annual Economic Statistics Information and Communication Services 2018
4. Department of Statistics Malaysia (2017). ICT Use and Access by Individuals and Households Survey Report 2017
5. Department of Statistics Malaysia (2018). Usage of ICT by Businesses and e-Commerce 2018
6. Hanna, N. K. (2011). Implications of the ICT Revolution. *Transforming Government and Building the Information Society: Challenges and Opportunities for the Developing World* (pp. 27-65). Retrieved from http://www.springer.com/cda/content/document/cda_downloaddocument/9781441915054-c1.pdf?SGWID=0-0-45-879448-p173923431.
7. OECD (2011). *OECD Guide to Measuring the Information Society 2011*, OECD Publishing. <http://dx.doi.org/10.1787/10.1787/9789264113541-en>
8. OECD (2012). *OECD Internet Economy Outlook 2012*, OECD Publishing. <http://dx.doi.org/10.1787/9789264086463-en>
9. World Bank Group (2018). *Malaysia's Digital Economy: A New Driver of Development*



Record linkage for statistical business register data



Maria Denise M. Peña
Asian Development Bank

Abstract

Data sources for Statistical Business Registers typically have different structures and several typographical errors - risking the data integrity of the database. Organizations can address this challenge by implementing record linkage techniques. These techniques intend to minimize duplicate records and to identify similar entities between different datasets, enabling smoother data integration. This study will explore record linkage methods and preferred specifications on data cleaning, deduplication, data matching, and validation of record pairs of Statistical Business Register data using R or RStudio.

Keywords

Fuzzy match; Deduplication; Entity resolution; Data matching; Data deduplication

1. Introduction

The ADB Statistical Business Register (SBR) serves as a central database for national statistics offices to store and retrieve historical and current information on businesses. This information contributes to the evidence-based decision- and policy-making of a particular territory, which entails the importance of the comprehensiveness and accuracy of the stored data. Since the information will come from various sources, a crucial challenge to optimize data quality would be the varying data collection formats, varying naming conventions, and data entry errors.

Government agencies may allocate resources to clean the data manually but this method may be unnecessarily time-consuming and susceptible to human error. This study will utilize recent technological advances in software and programming techniques to automate, or at least expedite the process of addressing data quality issues, with relatively accurate outcomes. The primary objective is to determine an extensive framework for data cleaning and identifying similar records between different datasets, specifically for the ADB SBR system.

2. Methodology

Scope and Data. The chosen software would be R or RStudio, a programming language and free software environment for statistical

computing and graphics. Apart from its customizability and transparency, R's Record Linkage package includes commonly used string comparators and employs probabilistic matching algorithms. The string comparator algorithms that will be tested are limited to those available in the PHP programming language, mainly because the existing ADB SBR system is based on this. Actual SBR data sources were used for testing. Due to confidentiality purposes, this paper can only reveal the overarching trends as a result of testing the record linkage framework using these datasets.

Framework. The outcome of data matching implementation depends on the source of the data and the year of documentation. There are four possible scenarios for Statistical Business Register data – deduplication, time-series, compilation, and consolidation. Datasets coming from the same data source and documented on the same year only requires identifying and deleting duplicate records. This outcome is commonly termed as deduplication. Since most datasets have numerous duplicate records, deduplication is a prerequisite step before the implementation of records linkage methods with other sources. After deduplication, the record linkage results between each dataset can produce a time-series that shows how businesses changed through time, a compilation of initially localized information about each business, or a consolidation of business information coming from various resources across multiple years. All datasets will undergo the proposed data matching framework detailed below to produce a cleaner and more accurate version of the latter three outcomes.

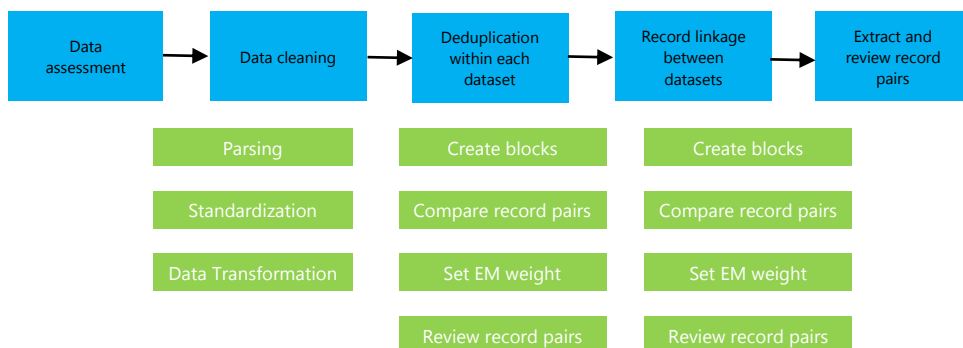


Figure 2.1: Proposed record linkage or data matching framework for ADB SBR data

Methodology. Given that the 2016 Tax data and 2016 Company Registry data are the only datasets available, this paper will only cover deduplication and compilation. The resulting trends will be validated by randomly generated datasets. Deduplication will be a pertinent step for all datasets before proceeding with record linkage. This is to ensure that there are no duplicate records that may cause multiple and equally-matched record pairs between datasets.

Data Assessment. Before any data cleaning can be done, it is imperative to assess if the language can be understood for data quality evaluation and if the dataset format is compatible with the software. Otherwise, the dataset will have to be translated and/or converted into a compatible format. Once the dataset is understandable, the next step is to identify the variables or fields for blocking and record comparison within and between the datasets. The dataset must contain pertinent variables for data matching and must be compatible with the chosen software to implement the succeeding steps of the Record Linkage framework.

Data Cleaning. Raw data would normally have a lot of typographical errors and variations, missing or out-of-date values, and different coding schemes. String comparison algorithms are highly sensitive to these, even to slight variations like capitalization and one-character differences. While phonetic comparison algorithms aren't as sensitive, these will completely omit the likelihood of similarity between two records if one is slightly misspelled. These are addressed by applying parsing, standardization, and data transformation techniques on the dataset.

Deduplication within each Dataset and Record linkage between Datasets. An additional step to data cleaning is to remove redundant records or duplicates within each dataset. This will mitigate the chances of having duplicated record pairs upon record linkage and inevitably enable smoother data integration to a central database. Clerical review and classic data cleaning techniques, such as smart functions, filtering, and conditional formatting, can be used to perform deduplication. Another option is to use the record linkage package in R since the process aims to identify similar strings.

The first step is to identify a categorical variable that can group the records into broad subsets. The similarity of records within the same subset will be quantified using a string comparator. By setting an EM weight, the software will produce a ranking of record pairs based on their likelihood of being matched pairs. Record linkage between datasets follows the same steps as deduplication but will be implemented between two different data sources.

Extract and Review Record Pairs. The results of deduplication and record linkage can be exported from R and reviewed by the researcher. While the algorithm can shortlist potential record matches and eliminate record-by-record clerical review, accurate validation of record matches will require a subject-matter expert of the datasets.

3. Results

Data Assessment. The framework was tested on Tax data and Company Registry data. Both are in English and in a format that is importable to RStudio - indicating that there is no need for a translation or a file conversion before data preprocessing.

Data Cleaning. The Tax data originally had a total of 40,038 records and 20 fields. Fields that are only 0%-50% populated were deleted; decreasing the number of fields to 17. The Company Registry data originally had a total of 219 records and 143 fields. Fields that are only 0%-50% populated were deleted and information in date format (DAY/MONTH/YEAR) were separated into date day, date month and date year; decreasing the number of fields to 29. For both datasets, all strings were capitalized, abbreviations were elongated, and unnecessary punctuations and extra spaces were removed.

Deduplication with a Dataset. Finding potential duplicates within datasets can only be done in RStudio. The tax data was split into eight sections (5,000 records per file) due to the processing limit of the hardware. There will be two rounds of deduplication. The first one will be within each dataset and the second will be between each dataset. The blocking variables used for the company registry data would be District and City. Business TIN, Business Name, and Business Unit Name were used for comparison. 77 exact duplicates were detected through deduplication, 72 within datasets and 5 between the datasets. The new record count is now 39,961. The maximum match rate (measure of similarity) is 1 for both the Levenshtein and Jarowinkler algorithms. The minimum match rate for Jarowinkler vary between 0.111 to 0.194. For Levenshtein, the minimum match rate does not deviate from 0.833. The frequency distribution of the match rates follow a normal distribution. The Levenshtein match rates are slightly more skewed to the right, indicating the increased likelihood of having lower match rates compared to Jarowinkler.

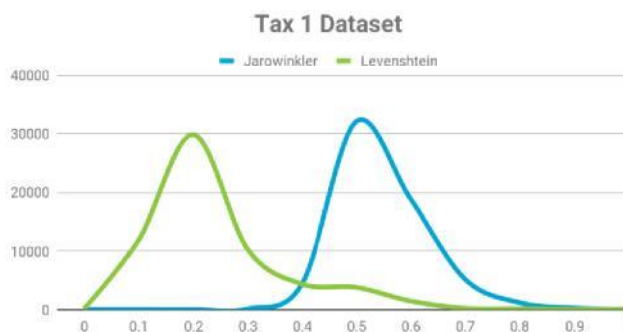


Figure 3.1: Sample frequency distribution of string comparator results of Tax Data

The blocking variables used for the company registry data were the Location Code, the Establishment Role, and the Reporting Unit. Business Name, Registration Number, Registered Name, and Tax Identification Number were used for comparison. Both string comparator algorithms (Levenshtein and Jarowinkler) were used for comparison purposes. 1115 out of 3234 record pairs were 100% matched; however, only 56 were validated to be true duplicates. The frequency distribution follows the same trend as the tax data frequency

distribution, with the Jarowinkler string comparator showing more optimistic match rates compared to the Levenshtein string comparator.



Figure 3.2: Sample frequency distribution of string comparator results of Company Registry Data

Record Linkage between Datasets. Now that both datasets have gone through preprocessing and deduplication, there will be less chances of records having multiple pairs. Tax data initially had 40,038 records. With 77 confirmed duplicates, only 39,961 records will be used to compare with the company registry dataset. The company registry’s record count dropped from 219 records to 163 after deduplication and confirming 56 exact duplicates within the same dataset through manual checking. The identified blocking variables would be District and City and the identified comparison variables would be Business Name, Tax Identification Number, and License Number since these variables exist and follow similar formats in both datasets.

Extract and Review Record Pairs. No exact duplicates were found between the tax data and the company registry data. This is largely due to the small sample size of the company registry data. The maximum match rates between the two datasets were 0.763 and 0.499 for Jarowinkler and Levenshtein, respectively. The minimum match rates were 0.113 and 0 for Jarowinkler and Levenshtein, respectively.

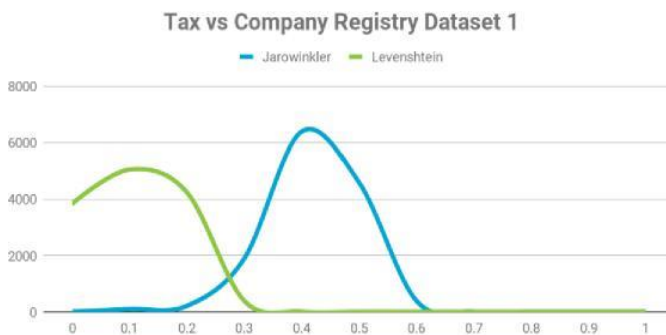


Figure 3.3: Sample frequency distribution of string comparator results between Tax Data and Company Registry Data

4. Discussion and Conclusion

Data Cleaning. This study covers the comparison of data matching results between datasets that have gone through data cleaning or preprocessing and datasets that did not. Results show that missing characters can decrease the match rate from 1 to as low as .875, missing words from 1 to 0.625, wrong capitalization from 1 to 0.625, and minor misspellings from 1 to 0.875. Records that would have easily been detected as similar entities through manual observation would be demoted to a lower match rate simply due to minor typographical errors. This further supports that it is pertinent to clean and standardize the datasets through removing punctuations and special characters, removing extra spaces, capitalization of all characters, and spelling out common abbreviations.

Deduplication. The company registry data, through deduplication, has shown to have 56 duplicates out of 216 records. The tax data has shown to have 77 out of 40,038. This means that the software will have 56 less records to compare with the rest of the records in the company registry data and 77 less records for the tax data, significantly decreasing the number of multiple duplicate record pairs and eventually the amount of time required to allocate for manual validation.

Blocking Variable. A dataset with more than 1,000 records is likely to have more than 1,000,000 records pairs since all records will be automatically compared. Given a large dataset and for practical purposes, it is preferable to assign a one to three blocking variables. Using categorical variables for blocking would be ideal, in the case of the available datasets. Variables that refer to unique characteristics may be too narrow and numerical variables may be too wide of a subset. In this study, the best options were establishment role, reporting unit, and location data (entity district and city).

Comparison Variable. Comparison variables are ideally unique characteristics of each record, such as business names and tax identification numbers. Based on manual observation of the record pairs between the tax data and the company registry data, it would be recommended to assign weights on each comparison variable based on the likelihood of uniqueness. Business names, for example, would have a larger weight compared to license number or tax ID since the latter two may have the same first characters by default due to a standard naming convention. This results in a high cumulative match rate between record pairs even if the business names are starkly different.

Dataset	BusinessName	TIN	License Number	Match Rate
Company registry	Army Welfare Project Limited	AAC00013	W2002686	
Tax	A B ENTERPRISE	AAP00434	W2002313	
	0.53	0.80	0.85	0.73

Table 4.1: Example of match rate result

String Comparison Algorithms. Given a large enough dataset, both algorithms produce match rates with a normal distribution, with the Levenshtein distributions relatively more skewed to the right. As observed with the minimum and maximum match rates and the match rate frequency distribution, the Levenshtein algorithm is more sensitive to differences between strings. The usage of either one will depend on the objective of the researcher. To obtain more potential matches and if there is ample time and bandwidth for a comprehensive clerical review, it would be preferable to use the Jarowinkler algorithm. If the main objective was to mitigate the risk of having false-positive matches and deleting potentially unique records, it would be preferable to use the Levenshtein algorithm. Another alternative would be to get the average of the results of both algorithms but this would still require manual assessment to ensure that the system captures the correct record pairs.

Match Rate Threshold. The defined match rate threshold for this study is 100% and will eventually be further refined if the results show that lesser match rates can pertain to real similar records. This study shows, however, that there is still a likelihood that records that have been detected as 100% similar based on the comparison variables are actually different records. Out of record pairs that have a 100% match rate, the probability of these being actual pairs is 0% to 51%, at least based on the results of this study. This suggests that manual validation is still a crucial step to ensure that the detected record pairs with relatively high match rates are indeed similar. Furthermore, most of the duplicates from deduplication are records with consecutive record numbers. In company registry dataset, for example, the 56 duplicates were all consecutive records. This implies that these records may pertain to different entities but the differentiating characteristic or variable between these records are simply not available with the given information. This further asserts the need for manual validation or assessment of the record pairs, at least during the initial stages of record linkage implementation for SBR datasets.

References

1. Murciano-Goroff, R.R.: Probabilistic Record Matching, <http://cs229.stanford.edu/proj2013/Murciano-Goroff-ProbabilisticRecordMatching.pdf>.
2. Borg, A., Sariyar, M.: Package 'RecordLinkage', <https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf>.
3. Harron, K.: Introduction to Data Linkage. Administrative Data Research Network (2016).
4. Herzog, T.N., Scheuren, F.J., Winkler, W.E.: Estimating the Parameters of the Fellegi–Sunter Record Linkage Model. *Data Quality and Record Linkage Techniques*. 93–106 (2007).
5. Köpcke, H., Rahm, E.: Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*. 69, 197–210 (2010).
6. Murray, J.S.: Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. *Journal of Privacy and Confidentiality*. 7, (2015).
7. Scholtus, S.: Probabilistic Record Linkage (Theme), https://ec.europa.eu/eurostat/cros/content/probabilistic-record-linkage-theme_en.
8. Winkler, W.E.: Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Statistical Research Report Series*. RR2000, (2000).



The insights of e-commerce in Malaysia

Hasnah Mat, Norul Anisa Abu Safran, Mazreha Ya'akub
Department of Statistics Malaysia

Abstract

E-commerce is a medium or platform of buying and selling goods and services over the Internet. This mechanism is widely used by all economic activities in Malaysia where it is transforming the traditional way of doing business. One of the components in boosting of e-commerce growth is improvement of ICT infrastructure. Anticipating the importance of ICT used and e-commerce growth and its contributions to the economy, the Department of Statistics, Malaysia (DOSM) took the initiative to embark on the surveys regarding the Usage of ICT and E-Commerce by Establishment (ICTEC) and ICT Use and Access by Individuals and Household survey (ICTHS) to obtain the required statistics. Therefore, this paper aims to share the findings of the surveys and the compilation of e-commerce in Malaysia. Income on e-commerce transactions recorded RM447.8 billion with an annual growth rate of 6.0 per cent while expenditure on e-commerce was RM228.8 billion with an annual growth rate of 8.3 per cent. The percentage of individuals using the Internet increased by 9.0 percentage points from 71.1 per cent in 2015 to a record 80.1 per cent in 2017, while for establishments it increased 11.8 percentage points from 61.5 per cent in 2015 to 73.3 per cent in 2017. Statistics pertaining to ICT used and e-commerce have been included in this paper. This result was in tandem with the target of Sustainable Development Goals (SDGs): Goal 9 to build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation. Apparently, Malaysia has improved on the ICT infrastructure, especially on the access and development of information and communications technology and strives to provide universal and affordable access to the Internet.

Keywords

ICT, e-commerce, SDGs

1. Introduction

Technological infrastructure is the key strength of the e-commerce industry. Information and communication technologies are playing a vital role in improving e-commerce practices in Malaysia (Muhammad Jehangir et al, 2011). The authors pointed out that strong ICTs and IT infrastructure are enhancing e-commerce capabilities in Malaysia. The increasing rate of internet

users, online spending and adopting of new technologies are the key drivers for the development of e-commerce in Malaysia.

E-commerce developing rapidly worldwide and it offers lucrative and vast opportunities. In general e-commerce transaction is the activity of buying or selling of goods or services through the Internet which bringing significant benefits to economies and societies. The convergence of the informational economy through ICT, Internet and electronic commerce has become more important transformation towards economic growth. It's not a surprise that selling and purchasing also have taken to the Internet and provide to the opportunities that are abound to make a significant presence in the global market. E-commerce become a trend, and has brought many changes in the society which has affected the daily lives of consumers and transformed the standard operating procedures of many businesses. In Malaysia, these significant transformations of e-commerce have influenced the specific economy such as manufacturing and services sectors.

Internationally, Organisation for Economic Co-operation and Development (OECD) plays a vital role in producing the recommendation for ICT, Internet and e-commerce. This recommendation is used to measure the development of e-commerce statistics where it is part of reliable source of comparable economic statistic and social data amongst the National Statistical Office (NSO). The OECD also monitors trends, analyses and forecasts economic developments, and researches social changes and evolving patterns in trade, environment, agriculture, technology, taxation, and other areas. Moreover, OECD is to promote policies that will improve the economic and social well-being of people around the world by providing a forum in which governments can work together to share experiences and seek solutions to common problems and identify good practice and coordinate domestic and international policies, (OECD, 2011). The Malaysia's NSO namely Department of Statistics Malaysia (DOSM) is a premier government agency under the Ministry of Economic Affairs entrusted with the responsibility to collect and interpret the latest statistics in monitoring of national economic performance and social development.

Based on mid-term review of the Eleventh Malaysia Plan 2016-2020, the share of e-commerce to Gross Domestic Product (GDP) was targeted at 20.8 per cent by 2020. The National E-Commerce Strategic Roadmap will assist all e-commerce traders to increase their contribution to the overall GDP to reach RM211 billion by 2020. Realising the importance of e-commerce activities to the Malaysia's economy, DOSM has took the initiative to embark on the surveys regarding the Usage of ICT and E-Commerce by Establishment (ICTEC) and ICT Use and Access by individuals and Household survey (ICTHS) to obtain the e-commerce statistics. Therefore, this paper aims to share the findings of the surveys and the compilation of e-commerce in Malaysia.

The statistics obtained from the surveys were in line with the target of Sustainable Development Goals (SDGs): Goal 9 to build the resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation. As a result, Malaysia has improved on the ICT infrastructure especially on the access and development of information and communications technology and strive to provide universal and affordable access to the Internet. The findings of the surveys reflects Malaysia's efforts to achieve the target of goal 9 especially on the inclusiveness and competitive economic forces in order to generate employment, income, facilitate international trade and enable the efficient use of resources. The fast growth of e-commerce business and ICT's technology will accelerate economic growth, connectivity, mobility and well-being of the people.

2. Methodology

ICT refers to the technologies and services that enable information to be accessed, stored, processed, transformed, manipulated and disseminated, including the transmission or communication of voice, image and/or data over a variety of transmission media. E-commerce transaction is the sale or purchase of goods or services, conducted over computer networks by methods specifically designed for the purpose of receiving or placing of orders.

The definitions and classifications are adopted from the guidelines as stipulated in the OECD Model Survey on ICT Usage by Businesses (2nd Revision), 2015 and Internet Economy Outlook, 2012 published by OECD. Thus, a standard concept and definition has been used to measure the impact especially on the e-commerce growth and its contributions to the economy. The classification of industry is based on Malaysia Standard Industrial Classification (MSIC) 2008 Ver.1.0. which is in concordance with International Standard Industrial Classification of All Economic Activities (ISIC) Rev. 4.

ICTEC is a structured survey to collect the necessary information on ICT used jointly with income and expenditure of e-commerce transactions. It serves as essential information to answer the research questions and objectives pertaining to ICT used and e-commerce in Malaysia. In general, this survey used a sampling design approach which covers all economy sectors which are Agriculture, Mining & Quarrying, Manufacturing, Construction and Services.

Sampling design of the survey is a one-stage stratified random sampling. Categories of industries and state level have been classified as stratum, and the establishment as the sampling unit. Each stratum (industry) has been set up to four substrata to ensure the distributed sample takes into account the economic characteristics of the industry. The main substratum is heterogeneous, was fully covered. Whereas, other substratum that are

homogeneous were sampled. Main substratum include large establishments that have a significant total revenue in the industry while for the second to fourth substratum are based on small and medium enterprise (SME) categories.

Meanwhile, ICTHS survey to collect the comprehensive information on ICT used and access by individuals and households in order to provide ICT indicators in development of national ICT. The concept and definition of this survey based on the Manual for Measuring ICT Access and Use by Households and Individuals 2014 Edition published by the International Telecommunication Union (ITU).

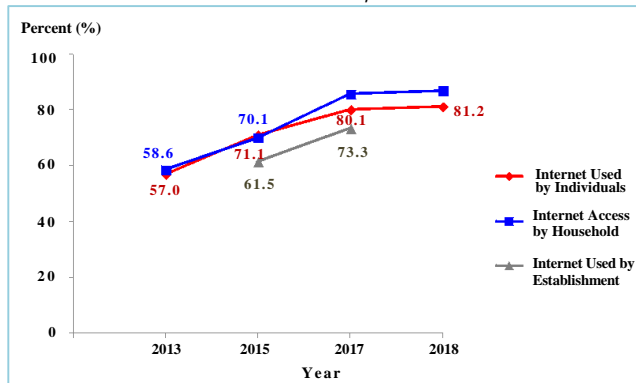
The sampling frame used for the selection of sample ICTHS was based on the Household Sampling Frame which made up of Enumeration Blocks (EBs) created for the 2010 Population and Housing Census which was updated from time to time. EBs are geographical contiguous areas of land which identifiable boundaries created for survey operation purposes, which is on average, contains about 80 to 120 living quarters (LQs). All EBs are formed within gazette boundaries i.e. within administrative, districts or local authority areas.

Two-stage stratified sampling design was adopted in ICTHS survey which primary strata covered all states in Malaysia, while for secondary strata covered urban and rural area which classification of areas according to population of gazetted, build-up areas and special development area.

3. Result

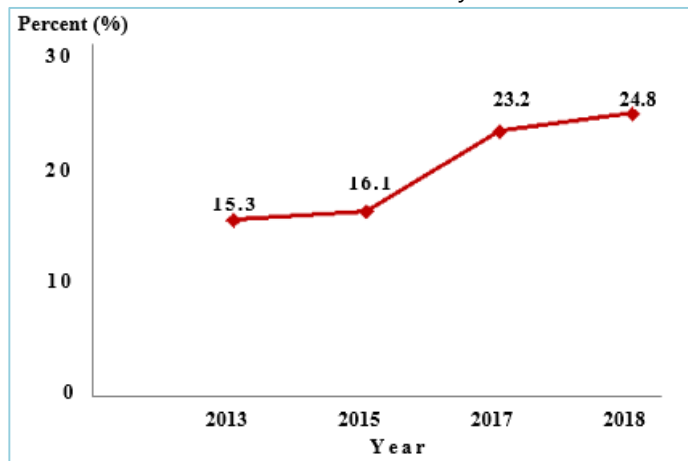
The boom of e-commerce highly related with access to the Internet, which encompasses internet accessible through computers and other devices. The changes in technology and communication devices have reflect to consumer behavior such as more people having online access. Result from survey of ICTHS and ICTEC shows internet used was expended year by year. Percentage of individuals using internet increased by 24.2 percentage points from 57.0 per cent in 2013 to 81.2 per cent in 2018. Households access to internet also increased 28.4 percentage points from 58.6 per cent in 2013 to 87.0 per cent in 2018. Internet usage for establishment increased 11.8 percentage points from 61.5 per cent in 2015 to 73.3 per cent in 2017 (Chart 1).

Chart 1 : Internet used, 2013-2018



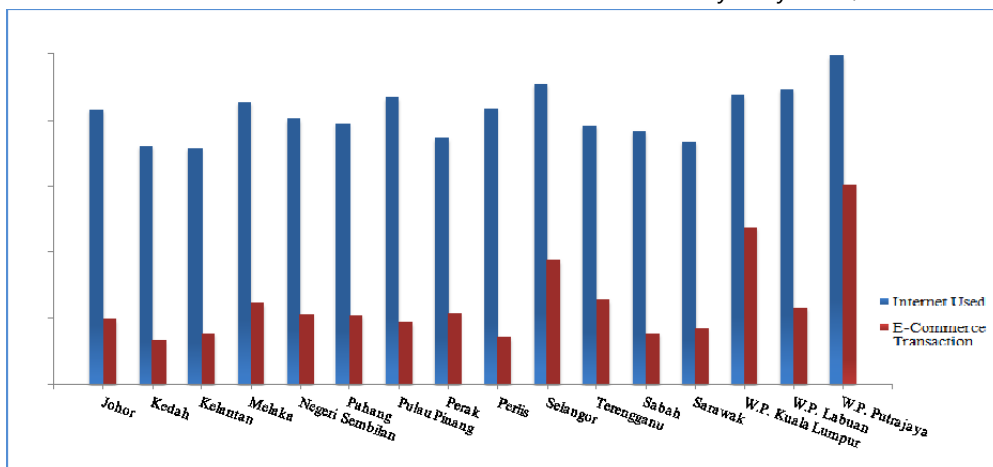
E-Commerce transactions by online purchasing or ordering goods or services by individuals in Malaysia are also moving upward year by year. According to the findings from ICTHS survey, the e-commerce individual buyer in Malaysia has reached 24.8 percent approximately 6 million buyer in 2018. Percentage of individual buyer increased by 9.5 percentage points from 15.3 per cent in 2013 to 24.8 per cent in 2018 as shown in Chart 2.

Chart 2 : E-Commerce individual buyer, 2013-2018



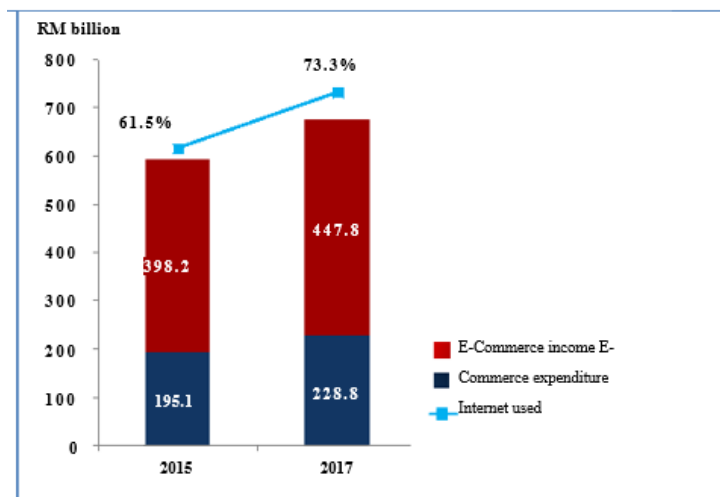
In 2018, W.P Putrajaya recorded the highest e-commerce individual buyer of 60.5 per cent, followed by W.P. Kuala Lumpur (47.5%) and Selangor (37.7%). It also shows the smooth usage of the internet in Malaysia where internet users recorded more than 70 per cent for each state (Chart 3).

Chart 3 : Internet Used and E-Commerce individual buyer by state, 2018



Increasing rate of internet used is a key drivers for the development of e-commerce in Malaysia. Findings from ICTEC 2017 survey showed 73.3 per cent establishment used internet as compared to 61.5 per cent in 2015. E-commerce income is higher than e-commerce expenditure for both year in 2015 and 2017. E-Commerce income registered RM 447.8 billion while e-commerce expenditure were 228.8 billion in 2017 (Chart 4).

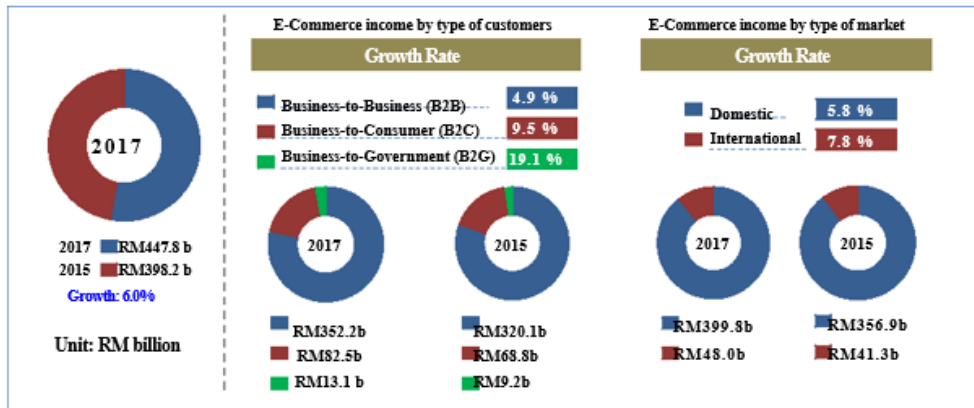
Chart 4: Internet Used and E-Commerce Transactions by Establishment, 2015-2017



The value of Malaysia e-commerce income grew 6.0 per cent in 2017, recorded RM 447.8 billion compared to RM398.2 billion in 2015. In 2017, E-commerce income was dominated by the domestic market of RM399.8 billion with a share of 89.3 per cent compared to the international market of RM48.0 billion (10.7%). The highest income from e-commerce transactions by type of customer was obtained through Business to Business (B2B) at RM352.2 billion

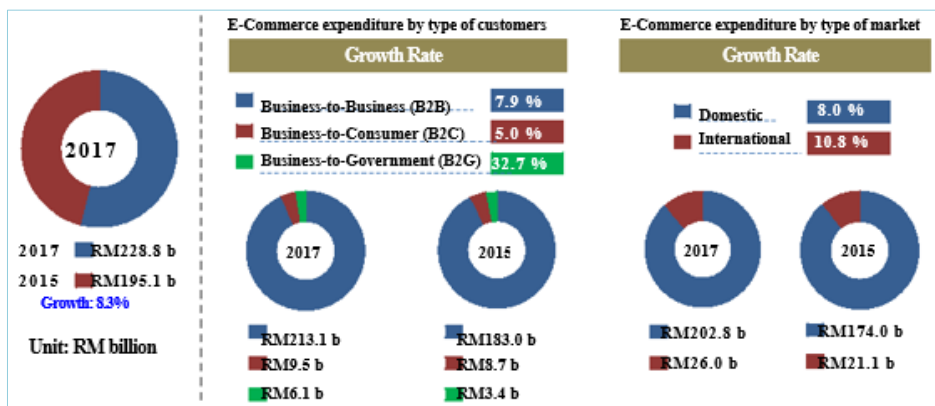
with annual growth rate of 4.9 per cent. This was followed by Business to Consumer (B2C) of RM82.5 billion (9.5%) and Business to Government (B2G) RM13.1 billion (19.1%) as shown in Chart 5.

Chart 5: Income from E-Commerce Transactions, 2015-2017



Meanwhile, e-commerce expenditure grew 8.3 per cent with a value of RM228.8 billion compared to RM195.1 billion in 2015. It was recorded domestic market of RM202.8 billion with a share of 88.7 per cent compared to the international market RM26.0 billion (11.3%). B2B recorded the highest e-commerce expenditure of RM213.1 billion with annual growth rate of 7.9 per cent, followed by B2C (RM9.5 billion; 5.0%) and B2G (RM6.1 billion; 32.7%) as shown in Chart 6.

Chart 6: Expenditure from E-Commerce Transactions, 2015-2017



4. Conclusion

Malaysia is rapidly growing region with overwhelming usage of internet access and becoming more attractive to business especially to the e-commerce activities. Modern businesses are in a race to provide the best finest

services to their consumers, with the development of ICT technology and the era of digital economy, most business has become simple, easy and up to date.

The statistics of ICT used and e-commerce are important in order to monitor the resilient infrastructure of ICT and sustainable industrialisation in Malaysia as indicated in the SDGs by 2016, the proportion of the population covered by a third generation (3G) mobile broadband network stood at 61 per cent in the Least Developed Countries (LDCs) and 84 per cent globally. Based on the result, Malaysia has improved on the ICT infrastructure especially on the access and development of information and communications technology and strives to provide universal and affordable access to the Internet. These efforts will make the e-commerce business become more competitive in order to generate employment, income, facilitate international trade and enable the efficient use of resources as targeted in goal 9 of SDGs. The fast growth of e-commerce business and ICT's technology will accelerate economic growth, connectivity, mobility and well-being of the people.

References

1. Department of Statistics Malaysia (2017). Information and Communication Technology Satellite Account 2017.
2. Department of Statistics Malaysia (2018). Annual Economic Statistics Information and Communication Services 2018.
3. Department of Statistics Malaysia (2017). ICT Use and Access by Individuals and Households Survey Report 2017.
4. Department of Statistics Malaysia (2018). Usage of ICT by Businesses and e-Commerce 2018.
5. Muhammad Jehangir, P.D.D Dominic, Naseebullah and Alamgir Khan, (2011). Towards Digital Economy: The Development of ICT and E-Commerce in Malaysia Vol 5. No 2. pp. 171-178.
6. OECD (2011). OECD Guide to Measuring the Information Society 2011, OECD Publishing. <http://dx.doi.org/10.1787/10.1787/9789264113541-en>.
7. OECD (2012). OECD Internet Economy Outlook 2012, OECD Publishing. <http://dx.doi.org/10.1787/9789264086463-en>.
8. World Bank Group (2018). Malaysia's Digital Economy: A New Driver of Development.



The GUIDE approach

Wei-Yin Loh

University of Wisconsin, Madison, WI, USA

Abstract

GUIDE is an algorithm for constructing classification and regression tree models. The talk discusses the basic problem of accuracy versus interpretability and the specific features that the GUIDE approach offers. Its unique features are demonstrated through two real examples.

Keywords

Classification and regression trees; interpretable models; missing values; prediction accuracy

1. Introduction

There is a general sense that machine learning models with high prediction accuracy are not interpretable. A large part of this belief is due to methods that are inherently uninterpretable to start with, but which can be tuned to achieve arbitrarily high levels of prediction accuracy on the same data that are used for their construction. When such prediction accuracy is based on independent test samples, studies have shown that there is typically no method with uniformly highest prediction accuracy (Lim et al., 2000).

Classification and regression tree models are inherently interpretable but are often regarded as not having as high prediction accuracy as other methods such as neural nets. This talk reviews the GUIDE (Loh, 2002, 2009, 2014) algorithm which has special features for enhancing interpretability, such as unbiased variable selection. GUIDE also stands apart from many other tree algorithms in being able to seamlessly deal with missing data, a practical difficulty that most machine learning methods are not designed for. Two examples are given below to illustrate these features.

2. Consumer expenditure data

The data come from the 2013 Consumer Expenditure Survey of the U.S. Bureau of Labor Statistics. It contains information on consumers' expenditures and incomes as well as characteristics of 2838 consumer units (CUs) on more than 600 questions. Table 1 gives the names and percents of missing values in some of the variables. Details on the survey may be found in Bureau of Labor Statistics (2016, Chap. 6). We use INTRDVX as the dependent variable; it is the amount of interest and dividend income received by the CU during the past

12 months. Each variable with missing values is associated with a “missing value flag variable” that takes values given in Table 2. Flag variables have underscores in their names; e.g, INTRDVX is the flag variable associated with INTRDVX. There are 587 neither constant nor completely missing X variables that may be used to estimate the population mean of INTRDVX. About 20% of these variables have missing values; 67 of them have more than 95% missing values. No CU has complete responses on all 587 variables.

Table 1: Variables and percents of missing values in consumer expenditure data

AGE REF	Age of reference person	0
FFTAXOWE	Weighted estimate for federal tax liabilities	0
INTRDVX	Interest or dividend received past 12 mos.	0
PERINSPQ	Personal insurance and pensions last quarter	
RENTEQVX	Monthly rent if home rented	15.6
RETSURVX	Retirement, survivor or disability pensions past 12 mos.	0
RETS RVX	Flag variable for RETSURVX	0
STATE	State (39 categories)	11.1
STOCKX	Value of directly-held stocks, bonds, mutual funds	92.0
TOTXEST	Estimated total taxes paid	0

Table 2: Codes and definitions of missing value flag variables

A	valid nonresponse: a response is not anticipated
C	“don’t know”, refusal or other type of nonresponse
D	valid data value
T	topcoding applied to value

Figure 1 shows the GUIDE piecewise-constant regression tree for estimating the mean of INTRDVX. A condition is printed on the left side of each intermediate node of the tree. A respondent goes to the left branch if and only if the condition is satisfied. The sample size and sample mean INTRDVX are printed below each terminal node. For example, at the root node, the 803 respondents who are 57 years or younger go to the left subnode which has a mean INTRDVX of \$803. The other respondents go to the right subnode. The symbol “<*” is an abbreviation for “< or missing.” For example, the right node immediately below the root node is split on STOCKX. Respondents with $STOCKX < \$191,160$ or $STOCKX = \text{missing}$ go to the left subnode. The node in black shows a special case where respondents go to the left branch if and only if $RETSURVX < \$11,342$ or with flag variable $RETS RVX = A$. See Loh et al. (2019) for a deeper analysis of the data.

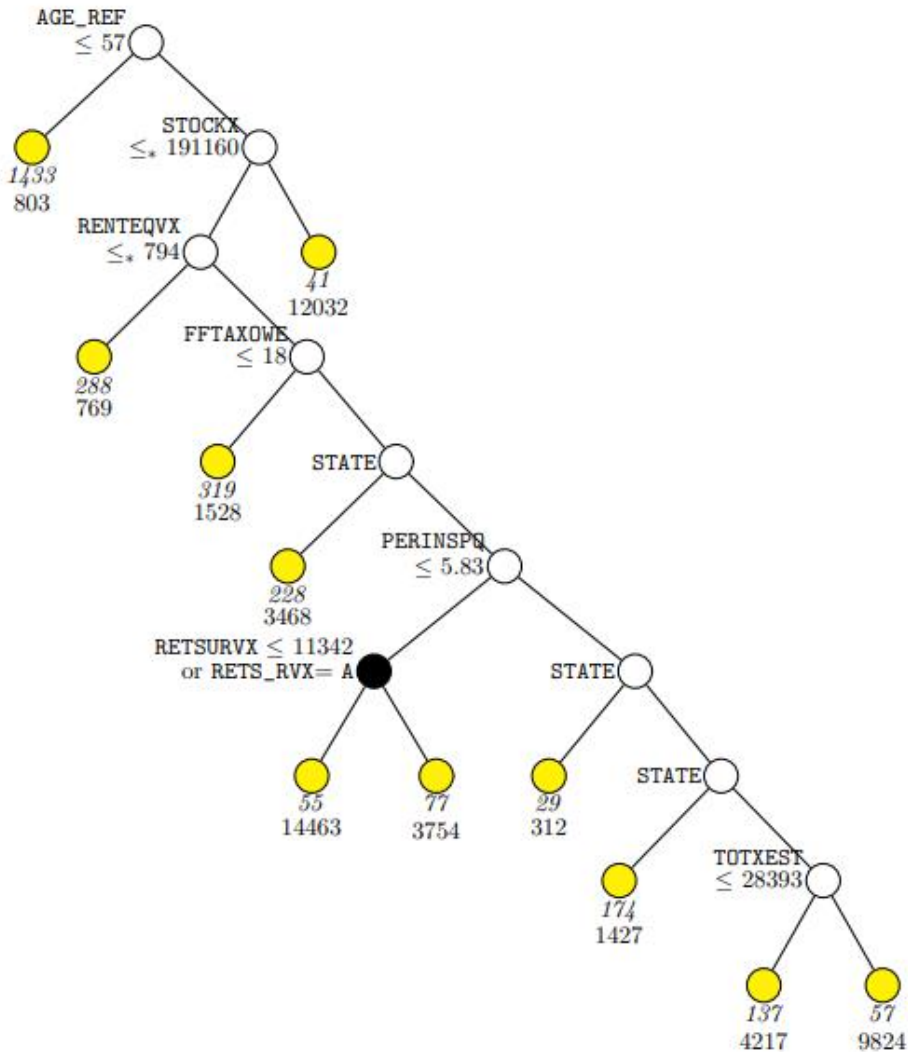


Figure 1: GUIDE piecewise constant least-squares regression tree for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '<*' stands for '< or missing'. Sample size (in italics) and mean of INTRDVX printed below nodes.

3. Crash test data

The data come from 15,941 crash tests of vehicles involving test dummies done between 1972 and 2004 by the National Highway Transportation Safety Administration (<ftp://www.nutsa.dot.gov/ges>). Each record consists of measurements from the crash of a vehicle into a fixed barrier. One variable is HIC, the amount of head injury sustained by a test dummy seated in the vehicle. For our purpose here, we define $Y = 1$ if HIC exceeds 1000, and $Y = 0$ otherwise. Thus Y indicates when severe head injury occurs. We use 109

predictor variables to fit a model to $P(Y = 1)$. Some of the variables are listed in Table 3.

Table 3: Some of the 109 predictor variables in the crash-test dataset. Angular variables PDOF, and IMPANG are measured in degrees clockwise (from -179 to 180) with 0 being front of car.

Name	Description	Variable type	Percent missing
MAKE	Vehicle manufacturer	71 categories	0
MODEL	Vehicle model	642 categories	0
YEAR	Vehicle model year	continuous	0.001
BODY	Vehicle body type	19 categories	0
ENGINE	Engine type	18 categories	0
ENGDISP	Engine displacement	continuous	0.007
TRANSM	Transmission type	9 categories	0.002
VEHTWT	Vehicle test weight	continuous	0.001
VEHWID	Vehicle width	continuous	0.027
VEHCG	Vehicle CG distance from front axle	continuous	0.024
COLMEC	Steering column collapse mechanism	9 categories	0.076
VEHSPD	Speed of vehicle before impact	continuous	0
PDOF	Principal direction of force	continuous	0.007
TKSURF	Test track surface	5 categories	0.024
TKCOND	Test track condition	6 categories	0.024
IMPANG	Impact angle	continuous	0
OCCTYP	Occupant type	13 categories	0
DUMSIZ	Dummy size percentile	8 categories	0
SEPOSN	Seat position	6 categories	0.025
BARRIG	Rigid or deformable barrier	3 categories	0
BARSHP	Barrier shape	21 categories	0

One thousand two hundred and eleven of the records are missing one or more data values. Therefore a linear logistic regression using all the variables can be fitted only to the subset of 14,730 records that have complete values. After transforming each categorical variable into a set of indicator variables, the model has 561 regression coefficients, including the constant term. All but six variables (ENGINE, VEHWID, TKCOND, IMPANG, RSTTYP, and BARRIG) are statistically significant. But the regression coefficients in the model cannot be relied upon to explain how each variable affects $p = P(Y = 1)$. For example, although VEHSPD is highly significant in this model, it is not significant in a simple linear logistic model that employs it as the only predictor. This is an example of Simpson's paradox. It occurs when a variable has an effect in the same direction within subsets of the data, but when the subsets are combined, the effect vanishes or reverses in direction.

Figure 2 shows the GUIDE logistic regression tree model, where a single predictor variable is used to fit a simple linear logistic regression model in each

node of the tree. For example, node 2 is fitted YEAR and node 6 with VEHCG. Of special interest is the split at the root node. It is split on IMPANG, which is the angle of impact. Because it takes periodic values, the split requires an angular segment as cut-point. An observation at the node goes to the left branch if IMPANG is between -135 and -73, meaning a left driver-side impact.

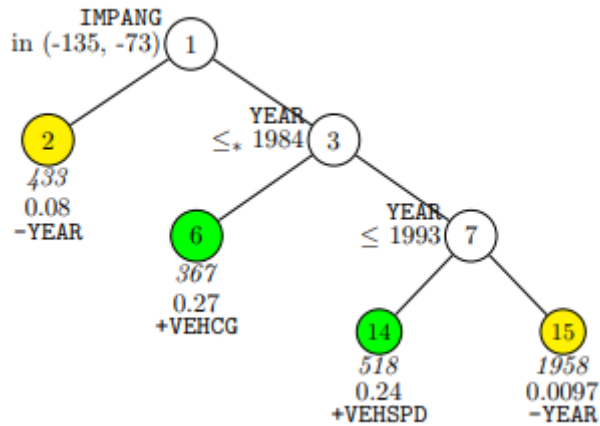


Figure 2: GUIDE simple linear logistic regression tree for estimating probability of severe head injury. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol ' \leq^* ' stands for ' \leq or missing'. Sample size (in italics), proportion of 1s, and sign and name of regressor variable printed below nodes.

Figure 3 plots the fitted logistic regression curves in each terminal node. They indicate that the probability of severe head injury decreases with YEAR and increases with vehicle speed VEHSPD and distance of front axle from the center of gravity of the car VEHCG.

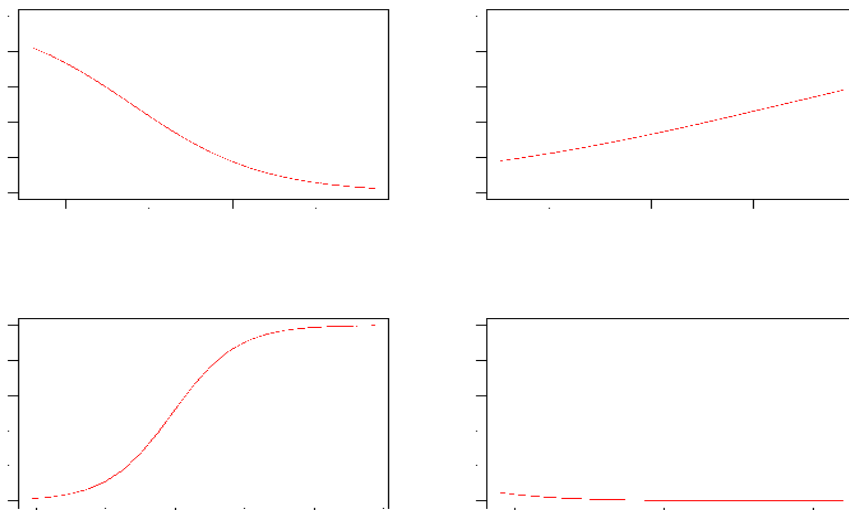


Figure 3: Estimated logistic regression functions in terminal nodes of tree for crash test data.

4. Conclusion

Additional literature for GUIDE and its software and user manual may be obtained from www.stat.wisc.edu/~loh/guide.html.

References

1. Bureau of Labor Statistics (2016). Handbook of Methods, Consumer Expenditures and Income. U.S. Department of Labor. <https://www.bls.gov/opub/hom/cex/pdf/cex.pdf>.
2. Lim, T.-S., W.-Y. Loh, and Y.-S. Shih (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning Journal* 40, 203–228.
3. Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* 12, 361–386.
4. Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics* 3, 1710–1737.
5. Loh, W.-Y. (2014). Fifty years of classification and regression trees (with discussion). *Inter-national Statistical Review* 34, 329–370.
6. Loh, W.-Y., J. Eltinge, M. J. Cho, and Y. Li (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica* 29, 431–453.



Profiling the internet economy in Singapore

Neo Soo Khee, Jeremy Tan, Choo Kit Hoong

Department of Statistics Singapore



Abstract

Recognising the need to understand the effects the internet has on the economy, the Singapore Department of Statistics has embarked on a pilot project to make use of web-based data sources to profile the internet economy in Singapore. This paper begins by defining the internet economy and discusses the work undertaken to use information from firms' websites to profile the type of internet usage for the firms. Specifically, it details the use of web scraping tools to first extract firm website addresses, and then to extract relevant features from the firms' websites. The websites will finally be classified into one of four internet categories using supervised machine learning. The paper finally presents the experiences of the pilot project and future work to expand the scope of the project.

Keywords

internet economy; web scraping; machine learning

1. Introduction

The internet permeates many aspects of our society, from the way people interact to how companies and businesses operate. Over the last few decades, the internet provided growth and start-up opportunities for many companies. Google and Temasek (2016) estimated that there would be 3.8 million new internet users every month in Southeast Asia, making it the fastest growing internet region between 2015 and 2020¹. Based on the Infocomm Media Development Authority's Business Infocomm Usage Survey, the business usage of internet for enterprises in Singapore increased from 82% in 2014 to 91% in 2018². Consequently, the internet economy in Southeast Asia is expected to grow exponentially. Increasingly, there is also a growing demand for a better understanding of the nature and effects of the internet economy.

The Singapore Department of Statistics (DOS) embarked on a pilot project to make use of web-based data sources to profile the internet economy in Singapore to better understand this emerging trend. Several national statistical offices have explored the use of web-based data sources. For instance, Statistics Netherlands studied the use of data from the web to measure the internet economy³ while the Italian National Institute of Statistics explored the use of web-based data to update the business registry⁴.

In this pilot project, enterprises were broadly classified into five categories according to their type of internet usage (Table 1). Enterprises classified under categories B1, B2, C1 and C2 make up the internet economy and the scope of the pilot project. Consumer-to-consumer economic activity was excluded from the scope.

Table 1: Categorisation of enterprises according to internet usage

Category	Definition	Examples
A	Enterprises without websites	-
B1	Enterprises which do not generate income directly from the internet and has passive internet presence	Websites with information on products/services
B2	Enterprises which do not generate income directly from the internet and has active internet presence	Websites with subscription services or social media outreach
C1	Enterprises which generate income directly online through sales of goods	Online retail stores
C2	Enterprises which generate income directly online through sales of services	Online web hosting services

2. Data Collection

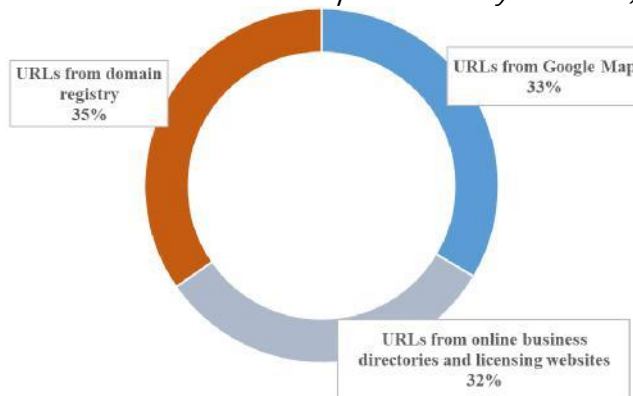
The Uniform Resource Locators (URLs) or the website addresses of the enterprises (if the enterprise has a website) were needed to classify the enterprises. The URLs of enterprises were only available for a sample of enterprises collected via traditional surveys. Hence, DOS explored using web-based sources to obtain the URLs of enterprises.

The URLs were collected based on the four steps below:

- 1) Obtained a list of enterprises from DOS's business register as the target population of this pilot project.
- 2) Purchased URLs from the Singapore Network Information Centre (SGN) For the remaining enterprises without a URL, their names and addresses were searched on Google Maps. If Google Maps recognised the enterprise and displayed its URL, the URL is then extracted and merged back to the target population.
- 3) Web scraped online business directories (e.g. Kompass and Orbis) and licensing websites (e.g. the Monetary Authority of Singapore and Travel Agents Directory) which contained URLs of enterprises. The URLs scraped were then merged back to the target population.

- 4) At the end of the data collection process, 35% of URLs were obtained from SGNIC, 33% from Google Map and 32% from online directories and licensing websites. (Figure 1)

Figure 1: Distribution of enterprise URLs by source(IC).



Once the URLs of enterprises were obtained, features were extracted from the websites to facilitate the classification process. For instance, a website that falls under 'Category C1: Enterprises which generate income directly online through sales of goods' would contain features such as a shopping cart and payment methods (e.g. visa, paypal). Feature selection was based on keywords found in a random sample of websites and further fine-tuned to add localised words to suit Singapore's context (e.g. 'SGD' and 'Singapore Dollars'). In total, 170 feature words were identified to be used for categorising the websites. The occurrences of the feature words were then counted in the website during scraping and then fed into the classification process.

3. Classification

A supervised machine learning classifier was used to classify the enterprise URLs into their corresponding internet usage categories based on the extracted features. A set of labelled websites (a total of 2,100 websites), which served as training data, was created by careful matching of enterprise URLs to their respective categories. Training data were then split into training and test datasets (80-20 split). During the testing phase, the accuracy of the classifier was determined by the percentage of URLs with the correct predicted internet category.

Different classifiers were employed and the parameters of each classifier were fine-tuned to obtain the highest possible test accuracy. The baseline test accuracy of the Naïve Bayes was 57% and a Random Forest Classifier was eventually chosen as it achieved the highest test accuracy of 79% (Table 2).

The Random Forest Classifier offered an additional ease of interpretation through its readily visible feature importance. Feature importance indicated the relative contribution of each feature to the classifier's predictions. For

instance, in the selected classifier, the word 'Shop' had a feature importance score of 0.044 which was more than seven times the average feature importance score of 0.006. This meant that the specific word 'Shop' was highly relevant in the classification as compared to a moderately important feature. This allowed a summary insight into the classifier's predictions. Feature words with notable feature importance are highlighted in Table 3.

Finally, the training dataset was used to train the Random Forest Classifier and the classifier was applied on the enterprise URLs to classify them into one of the internet usage categories (B1, B2, C1 or C2).

Table 2: Results of algorithms explored

Algorithm	Test Set Accuracy
Random Forest	79%
Gradient Boosting Machine	77%
Voting Classifier	77%
Logistic Regression	72%
Neural Network	71%
AdaBoost	70%
Support Vector Machine	68%
Naïve Bayes (Baseline)	57%

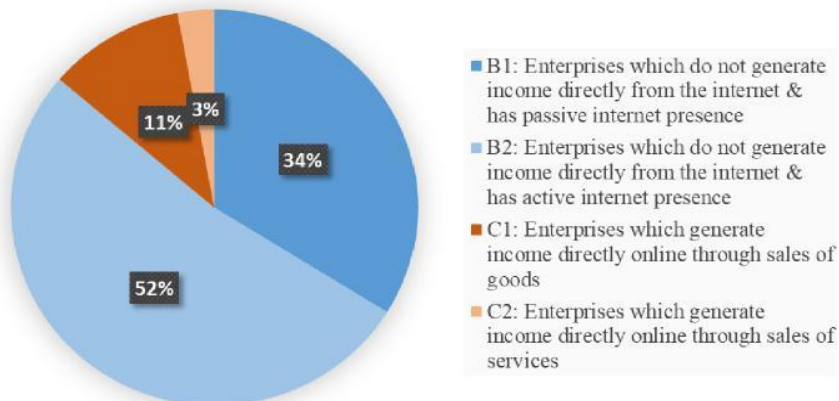
Table 3: Feature importance of selected words

Feature Words	Feature Importance
Shop	0.044
Cart	0.041
Price	0.027
Facebook	0.021

4. Results

Out of the enterprise URLs obtained to date, 14% have websites which generate income directly online (Figure 2). One caveat to note here is that the enterprises classified under 'Category C1/C2: Income generated directly online' might not generate their income wholly through online means and the online platform could be one of many different revenue streams.

Figure 2: Distribution of enterprises with websites by the internet categories



5. Discussion and Conclusion

This pilot project demonstrated the possibility of profiling and measuring the internet economy of Singapore. By using data collected from various web sources, useful insights could be derived from the internet economy.

Looking forward, available survey and administrative data could also be merged to the target population to gain a deeper understanding of the internet economy. It would also be interesting to study enterprises which do not have corporate websites but promote their services on social media or sell their goods/services through online marketplaces. (e.g. e-bay, Qoo10).

References

1. Rajan Anandan, Rohit Sipahimalani, Alap Bharadwaj, Jaideep Jhangiani, Danny Kim and Soumi Ramesh. (2016) "e-economy SEA: Unlocking the \$200B Digital Opportunity". <https://www.thinkwithgoogle.com/intl/en-apac/trends-and-insights/e-economy-sea-unlocking-200b-digital-opportunity/>
2. Infocomm Media Development Authority (2017). "Infocomm Usage-Business". <https://www.imda.gov.sg/industry-development/facts-and-figures/infocomm-usage-business>
3. Oostrom L., Walker A., Staats B., Slootbek-Van Laar M., Ortega Azurduy S. and Rooijackers B. (2016). "Measuring the internet economy in The Netherlands: a Big Data analysis". Statistics Netherlands. <https://www.cbs.nl/-/media/pdf/2016/40/measuring-the-internet-economy.pdf>
4. Bianchi G., Consalvi M., Gentili B., Pancella F., Scalfati, Summa D. (2018). "New sources for the SBR: first evaluations on the feasibility of using big data in the SBR production process". Italian National Institute of Statistics. http://www.wiesbaden2018.bfs.admin.ch/wp-content/uploads/2018/08/Paper_Bianchi.pdf



Use of web-scraping for the compilation of Consumer Price Index: Malaysia's experience



Fauzana Ismail, Fuziah Md Amin, Wan Mohd Haffiz Mohd Nasir
Department of Statistics Malaysia

Abstract

This paper describes how Malaysia, like many other countries, is inevitably involved in a phenomena called big data. Big data apparently involves a voluminous amount of data, where the data types and structures are complex and there is a speed of a new data creation and growth. Nowadays, there is an increasing amount of digital data that flooding from various sources such as from the web, email, videos and social network communication. As such, big data with its main focus on the unstructured data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale. The definition of Big Data Analytics (BDA) and its importance are also discussed in the paper. BDA is the process of examining large amounts of data to uncover hidden patterns, correlations and other insights and is very pertinent especially in the business aspect. Most National Statistical Office Big Data Analytics involve the price statistics. In Malaysia case, it is found that most data are already online. Therefore, in promoting E-Commerce in Malaysia, the traditional collection method that is conducted on the face-to-face basis is in the need of a modernisation so that it is more real time and more efficient. In addition to this, there will be a reduction especially in terms of cost and burden to respondents. Inasmuch, one of the many StatsBDA initiatives in Malaysia, as currently undergoing in the Department of Statistics Malaysia is called the Price Intelligence Module. The initiative is to create an internal portal for Price Intelligence (PI). It involves a modernisation of data collection tools for improving the quality of Consumer Price Index (CPI) in Malaysia. The modernization of data collection mainly scraping consists of the adoption of web techniques to scrape price data from related website for CPI compilation. The idea is to crawl data from hypermarkets and to be collected in the big data project. As a result of this, analysis, data visualisation, data mining, reports and dashboards as well as alert can be conducted. The Price Intelligence Module in the Department of Statistics Malaysia encompasses of Price Frequency, Distribution of Average Price by Strata Category, Trend of Average Price by Strata Category, Trend of Average Price by State, Price Distribution by State, Average Price by State and Price Mode. The paper concludes with the challenges and journey of the Price Intelligence that Malaysia had experienced

and will continue experiencing. It is an on-going process and in tandem with the data revolution that happens in the world nowadays.

Keywords

Big Data; Price Intelligence; Analytics; Consumer Price Index; Challenges

1. Introduction

Among the many importance of Big Data are reducing cost, outdo competition, better decision-making and improving products and services. The earliest record of using data to track and control businesses dated back from 7,000 years ago when accounting was introduced in Mesopotamia in order to record the growth of crops and herds. In the past few years, there has been a massive increase in Big Data startups, where all trying to deal with Big Data and helping organisations to understand Big Data. Inasmuch, more and more companies are slowly adopting and moving towards Big Data. One of the StatsBDA modules that is being developed in DOSM Big Data is Price Intelligence (PI). The modernisation of data collection mainly consists of the adoption of web scraping techniques to scrape price data from related website for CPI compilation and improving the quality of data. Another need of this Price Intelligence module is because of the growth in e-commerce nowadays. The StatsBDA in Malaysia began in August 2016 where the advertisement of the tender started until it is awarded to one of the industry players in November 2016. From that date onwards, the project will be expected to be accomplished by May 2018.

STATBDA's main goal is to make sure that no one gets left behind. Price Intelligence or PI is responsible to care for the interest of the other side of the spectrum that is the consumers. The main goal of PI is to create a price list of different goods and provide the solution for consumers on the best prices available. Via web scraping, PI extracts the product prices from the internet through a method called web crawling. The prices will then be formed into a structured data and they will be classified into different categories. The consumer can then see for themselves a list of prices from hundreds of sellers and sort out the best prices for them through what we call Price Basket Enrichment where online prices will be integrated with current CPI and the prices are made public. Price basket enrichment will not consists of all CPI basket though. It will consist only certain items which is available and suitable to use.

At the same time, the DOSM Transformation Plan 2015-2020 states that the main priority in strengthening the role of the department is to benefit the data evolution through big data. Strategic Core 1 is the creation of product and statistical service integrity and reliable while Strategy 3 is to enhance the utilisation of secondary data sources. Meanwhile Program 2 is the initiative in

the use of big data. The objective is to provide a better view in monitoring and analysing consumer prices. It is also to create a price analysis of new basket which will be used as the value added for the Consumer Price Index in Malaysia.

2. Methodology

The initiative is to create an internal portal for Price Intelligence (PI). It involves modernisation of data collection tools for improving the quality of Consumer Price Index (CPI) in Malaysia. The modernization of data collection consists of the adoption of web techniques to scrape price data from related websites for the CPI compilation. The idea is to crawl data from hypermarkets and to be collected in the big data project. From there, can be done analysis, data visualisation, data mining, reports, dashboards and alert. Discussions and consultations pertaining to this Price Intelligence Module are conducted from time to time among respected parties who involve in this project. The meetings to discuss the progress status of this project are also conducted on weekly basis. In this respect, for Price Intelligence Module, among the parties involved in the Department of Statistics Malaysia are the Prices, Income and Expenditure Statistics Division, the Methodology and Research Division and the Information Management Division. Among the challenges at the initial stage of the Price Intelligence Module, at least on the industry's personnel part, is to understand the nature and scope of work of the Department of Statistics Malaysia which involves the codes and the classifications used, the items, price statistics and the very definition of the Consumer Price Index itself. In Price Intelligence Module, there is a Data Management which objective is to classify raw online data to its corresponding Classification of Individual Consumption According to Purpose (COICOP) and to provide a working platform for managing PI Data Management. Data management involves in the process of matching data with the Consumption of Individual According to Purpose (COICOP). There is also a crawling process which involves the monitoring and alert regarding the crawling process from the selected websites. As for Price Lake, it involves the data generator, Big Data concept and monitoring storage data. In PI Module, there is Analytics & Visualisation which involves some analysis and visualisation using R programming and Tableau. As for PI Data Processing, since the crawling of the data is conducted all the time, data processing cannot be conducted on a time-base manner. This is to avoid the FTP folder from crammed. Having said that, as PI module in Malaysia made its maiden journey, there are lots of challenges and issues related to it. Beside the prices data that are crawled everyday keep changing, some of the issues that are inevitable to be encountered are the price phishing, the prices that become too broad as well as other issues.

From Price Acquisition to Price Lake, one of the many challenges pertaining to Big Data is that to make the data structured so that from there, analysis can be done. Data scraped online from the internet involves some online sources. The challenge is that some online sources provide data at the national level and each and every sources has different item category. Another source of data is from data processing system of the Consumer Price Index in DOSM. This type of data can provide data up to the state level as well as the location level. The latter means that data from the survey that was conducted, put into ftp and then into the Price Lake. At the same time, data from online (internet) also will be into the Price Lake. This is what it means with the acquisition of the data.

Online Price Data Acquisition

Instead of traditional way of collecting price data, web crawling and scraping approach are now becoming popular way of price data acquisition but when it comes to technology, we always know that there is always a price for it. Not only in terms of the platform and the robot aspects, but the challenges in the dynamism of the online sources itself.

Online Price Data Acquisition Architecture

In the general perspective architecture, DOSM's data acquisition components are described in diagram below. The sources to the internet are from both online government agencies and online retailers.

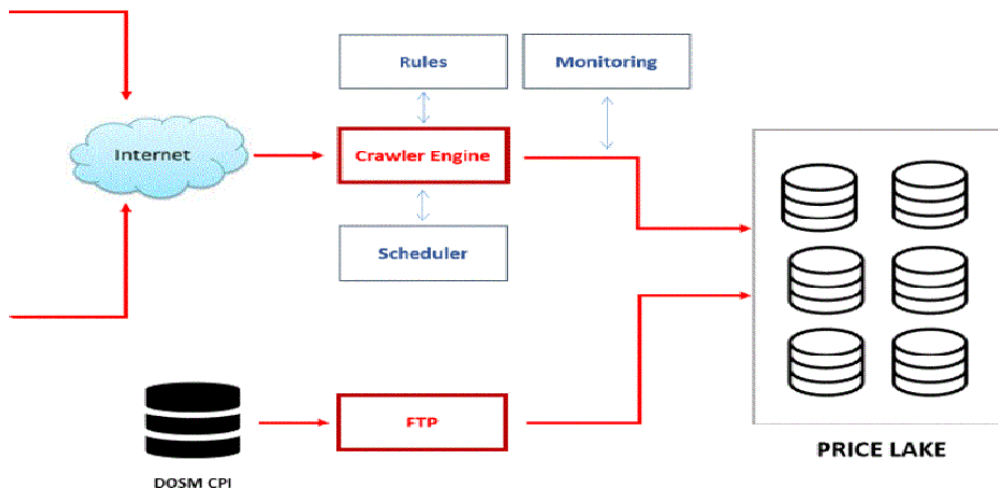


Figure 1: Online Price Data Acquisition Architecture

The main function of this data acquisition is to collect and extract data from various sources. DOSM are now crawling more than 50 websites with 22 of them are from e-commerce websites and are the biggest web scrapper among

NSOs with 300 to 500 GB data have been loaded everyday into our Price Data Lake. Price Data Lake is a Hadoop Environment under Big Data Analytics initiative. There are two sources of external data that will be captured through internet using crawling process. Online retailers as one the sources are e-commerce kind of source and another one is from online government agencies in the format of prices, index, and rates. Crawler engine works based on rules defined for each website to determine the objects required to be captured. Rules are developed in the format of programming scripts to parse HTML code of the website pages. There is a scheduler component in the crawler engine to define the time of the engine to work. The scheduler will be based on each website that working on some algorithm to avoid blocking action by the sources. The crawler also has component for monitoring to ensure the data captured properly based on the rules and schedule defined. Data that have been captured by crawler engine will be sent and stored into Price Lake. Data from internal sources stored in existing DOSM CPI system will be transferred to the Price Lake through FTP approach.

Web Crawling and Scraping Approach

External Data Sources are sources of data located outside DOSM environment. Mostly they are in the form of website and portals and connectivity to those data will be using internet connection. Online retailers are portal or website that provide list of prices of daily goods. DOSM crawl and scrap most of the price items varies from consumer electronics and electrical product, household goods, toys, sport equipment, hand phone, camera, groceries, house rental, furniture, fashion product, etc. Each of the website provides various content categories and most of the contents are on the title, picture, price, seller, description, etc.

Crawling is a process of capturing the HTML pages of the website and store them into local repository. HTML pages contains various codes that constructing the web page. A scrapping process is required to extract object interest within the HTML codes. Because of dynamism most of the webpages, a native programming by using Python was used to develop web crawling and scraping program. Crawlers works based on program developed to parse the HTML files and to define codes that represent the interested objects. The program will result in the format of rules and the rules have high dependency with the layout of target website. Once the layout of targeted website changes, rules need to be change in order for the crawler engine to capture the data. Hence, a continuous monitoring is required. Crawler works as robots that simulate human visiting the website and collect the content. Several library are needed for programming the crawling script in Python. Those of the libraries are "Request", "Beautifulsoap", "re", "os" and "time". "Request" is a library in Python that being used for collecting the HTML in a website while "Beautiful

Soap” was a library being used for taking the data that we need in HTML. “re” was used for taking the string in the string and “os” being used for to make a directory automatically. “time” library was used for make the program sleep whenever possible. First of all, we create one function in a python program to get all the category link in the website. Then we create one function to initiate the variable and the constant. We parsing the category link by using “Beautifulsoap” library that we mentioned above in the form of “bs4”. The “bs4” was always in the form list, thus “for” was using for looping of the data. The selector in HTML needs to be identified before we can perform parsing the HTML. The selector are referring to “tag” in HTML, as a “class” or “id” in CSS. The selector in HTML can be identified by utilizing CSS and Xpath selector. After we have parsing the category link with data looping and put it on the variable result, we created one function code to request URL category link to get data in the HTML and we store it in URL index 1. Again we parsing the HTML data to get URL items link that have one parameter that also being put on the variable result as a list. The next function was to get the data and put it on two parameters with one for index name and another one for data items link. The “os” library being used for create the folder automatically and the data being store based on their indexes. Then we created function to get data until the last pages of the website to get all indexes in the category link. The last page will be looping to get all the index link before we created the last function for request all the index link. There are different approach and techniques for a different website to develop the crawling program in python. In some cases, the website are using AJAX and there is no HTML content revealed. This needs to be treated differently. Asynchronous JavaScript and XMLHttpRequest, or abbreviated AJAX, is a web-based programming technique for creating interactive web applications. The goal is to move most of the interactions on the web surfer's computer, to exchange data with the server behind the scenes, so that the web page does not have to be re-read in its entirety every time a user changes. First, we need to check the AJAX by inspecting in a website, then we can click on the network button and select XHR. The analysis needs to be made on AJAX base link, id of AJAX and the end of AJAX link. The category of URL AJAX in a website needs to be collected by searching the ID in URL AJAX that have been always changed in the content element. In this case, Selenium library must be used because the element cannot be requested. Figure 3 are one of the example of parsing html result that had been stored in Hbase after the robot successfully save the page description including the indexes.

How crawlers work is not much different on how spam agent works. Sometimes the target websites implement security to detect robots accessing the website, and block them to reduce the target website's bandwidth. In order to avoid detection as robots, crawler must be smart enough to make the target website not to detect crawler is a robot. DOSM have been implemented several

strategies to cater this matters. The Robot will be simulating to be declared as a browser and being configured to adopt Security Socket Layer (SSL). The characteristics of the website have been detected whether it use a captcha, redirection, etc. The website that had been crawled will be randomly accessed using multiple public Internet Protocol (IP). The crawling period also being adjusted accordingly. On top of that, DOSM also implementing The Onion Ring (TOR) based on Deep Web approach and Proxy implementation. All those strategies will be implemented as integrated algorithm in the Crawler scheduler.

Since there are risks resulted from target website layout changes and also security deployed in the target website, a monitoring agent was used to ensure that all components in the crawler engine runs well. Monitoring agent will send alert and notification to us when there are issues such as no data captured from specific website. Our developer has to check and analyses the cause of the issue and doing problem solving as soon as possible. It was highly noted that the data crawling and scraping are now running heavily for all the 50 sources. There is no doubt that by having native programming to performing all the task, a highly maintenance are needed by skillful and experience technical person. DOSM's Technical Person with collaboration from industry's personnel are monitoring and maintaining crawling and scraping processes with 10 person are working at one shift time.

3. Results

All processes above resulted in the development of Price Intelligence module in DOSM. There are a number of 100 selected items in the visualisation of this module. The items among the most highly consumed by the households in the country. Among the items that take into consideration are of fruits, vegetables, fish, seafood and food away from home. Among the expected output from the Price Intelligence are price and changes comparison between online and existing data, average price by Malaysia, state, strata, most expensive/cheapest price by item, by location. The suggestion of visualisation for data survey are Price frequency, Product price distribution by CPI category, Trend of average price by category, Trend of average price by state, Price distribution by data type, Average price by state and Price mode. The suggestion of visualisation for online data are Price frequency from item at six digit Classification of Individual Consumption According to Purpose (COICOP) code, Trend of average price by outlet category, Distribution of average price, Trend of Average Price by area, Price Distribution by state, Average price by state and Price mode. In the data visualisation, data can be filtered into 2, 4 and 6 digits of CPI. The value creation of this Price Intelligence platform is to obtain a holistic information pertaining to online prices as well as offline prices.

Enabling monitoring and forecasting the price trend in future as well as a useful input for the government to decide on the price control.

4. Discussion and Conclusion

The session of transfer of knowledge and technology are also conducted from time to time. The main objective of the session is to transfer knowledge and skills to the Department of Statistics manpower to ensure DOSM has all the necessary skill set to manage and support all modules related to the StatsBDA project. Knowledge is transferred in the form of formal classroom training on the tools used to develop the system, system administration training and application usage training as well as technical documentation covering the system architecture, design and user. In respect of Price Intelligence Module, one of the way forward is to keep updating the Price Intelligence dictionary. The Price Intelligence dictionary involved the adoption of the COICOP. In terms of QA and QC data dictionary, five permanent manpower and four daily short-term manpower were allocated to update the existing dictionary into new dictionary. They are from consumer price index unit and Price Intelligence Unit. The dictionary is already developed with one hundred and fifty plus items with COICOP Malaysia at seven digits level where altogether there were 800 item specifications in Consumer Price Index basket. The proposition is to complete the balance of the item according to the main groups. The priorities were given to food item (groceries), clothing and footwear, household equipment items and electrical goods as well as personal care items.

Another way forward is to increase the number of e-commerce websites and the number of item covered. Apart from that, is to study the effectiveness of online prices as input to the CPI calculation and as an alternatives to the manual survey of DOSM, besides saving cost and time. Strategic collaboration with agency that deals with national price council in maximising price data in price lake is also conducted. In the meantime, the price lake itself is also needs to be enhanced as a way forward. Way Forward is also in extending this Price Intelligence Module to be accessed by the provider of our online data which is also known as the price catcher. This is in line with the facing challenges nowadays in terms of cost of living aspects. Access to the Price Intelligence Visualisation for data namely Price Catcher, Tesco and Manual Survey DOSM where the average price of one hundred items are already published. These one hundred items are selected based on the most significant consumption of users in Malaysia. For the Price Intelligence Module in relation to human resources, for first level support, personnels from Management of Information Division will be responsible for the maintenance and monitoring the data crawling. For the second and third level support, it will be implemented by the

contractor during the maintenance period since the maintenance and repair process are in need of a large number of manpower with high skills.

In conclusion, the module of Price Intelligence is further expected to reduce the burden in DOSM states which involve in the consumer price collection that is conducted on the monthly or weekly basis. This is because Price Intelligence can provide us with the data from the online source. Web scraping from the internet is an efficient way of collecting the data needed in the CPI. Having said that, few methodological issues as well as legal consideration are needed to be resolved before such an approach can be implemented in the official CPI. In terms of legal aspect, although acquiring data automated from the online sources are now becoming common things especially on dealing with the statistical data collection, most of the online sources that crawled did not mention any clear restriction on automatic data acquisition. Since DOSM is the official statistics agency, there is act that permit DOSM to do data collection. However this need to be made up clearly and should be made officially by make amendment to the current statistical act accordingly.

At the same time, some literatures and researches based on the practices of other countries in dealing with the Big Data are also conducted. Inasmuch, the price is indeed intelligence in itself but the human intellectuality in making choices based on the prices information must unfailingly takes precedence.

Reference

1. StatsBDA Bullet Volume 2 (Editorial Board, STATSBDA Team c/o Methodology and Research Team) Reference:
<http://bigdata.black/infrastructure/platforms/history-of-big-data/>



Research on deepening the application of big data in government statistics



Li Jijiang

Beijing Municipal Bureau of Statistics, Beijing, China

Abstract

With the advent of the era of big data, new opportunities and challenges are brought to government statistics. In terms of the external environment of government statistics, the big data are considered as the new competitive resources to be raised into the national strategic level, so various countries around the world are scrambling to enact relevant strategic policies to promote the research and development of big data. In terms of the internal environment of government statistics, the big data have such distinct characteristics as a broad variety of sources, diverse forms and non-standardization to exert a great impact on traditional data collection, analysis and publishing, etc. The government statistics is no longer the only data producer and publisher, and faces many problems like prominent phenomenon of data coming from various sources, increasingly obvious time lag of official data, and poor utilization of data. In such internal and external environments, the application of big data is still in the preliminary exploration stage for the government statistics, so it is especially necessary to explore how to deepen the application of big data in government statistics. By elaborating the concept and characteristics of big data, analyzing the thinking contained in big data, and revealing the difference in thinking between the big data and government statistics, this paper, on the basis of summarizing the characteristics of big data applications in enterprises, explores how to draw on the thinking and application experience of big data, and finds out the point of convergence between big data and government statistics, thus achieving complementary effects. This paper starts from the statistical business process, divides the government statistical process into three parts such as data collection, data analysis and data release, analyzes the applicability of big data in each part, and then makes targeted suggestions, so as to improve the data quality of government statistics, reflect the economic and social dynamics of the new era in a timely and accurate manner, and provide a better reference for government decision-making.

Keywords

big data thinking; statistical process; application space

1. Introduction

In recent years, the research on application of big data has become a hot spot. The relevant governmental authorities are required to establish a big data platform to integrate the information sources, realize one-stop online government services, and improve the urban management level. However, the research on big data and governmental statistics is relatively scarce and mainly focuses on exploring the application of big data technology in government statistics, which is still in the preliminary exploration stage. Actually, the emergence of big data has brought new opportunities and challenges to traditional government statistics. This paper aims to study how to deepen the application of big data in government statistics, improve the level of government statistics, better grasp the trend of social and economic development, and make the governmental statistics really become a “think tank” for government decision-making.

This paper includes five parts: the first is the research background, which states the opportunities and challenges that the government statistics face in the era of big data as well as the necessity and importance of subject selection based on the current internal and external environments; the second is the rational understanding of “big data”, which pays more attention to the data processing thinking contained in the concept of big data and brings a deep understanding of the differences between big data and government statistics to lay a foundation for deepening the application of big data; the third is the summarization of application characteristics of big data in enterprises, which provides a new thinking for the application of big data in government statistics by drawing on experience; the fourth is the analysis of application space of big data in government statistics, which respectively analyzes the current situations and improvement possibility from such three aspects as data collection, data analysis and data issuing, by taking the business process of government statistics as an entry point; and the fifth is the provision of guarantee for the application of big data, which puts forward measures from legal system, personnel and technology to ensure that the application of big data is deepened in government statistics.

2. Methodology

The literature research method and qualitative analysis method are adopted in this paper. By searching and consulting the references about the application of big data, deeply understanding the concept of big data and knowing the current application situations of big data in enterprises and governments, it is found that the application of big data in enterprises is extensive and mature but the application of big data in government statistics is still in the preliminary mining stage. Therefore, the research area of this paper is put forward. In other words, from the aspect of government statistics,

this paper aims to explore how to deepen the application of big data in government statistics.

In the qualitative analysis method, it is required to state the thinking of big data and the differences between it and the government statistics on the basis of recognizing the concept of big data; to summarize the application characteristics of big data in enterprises so as to the thinking for the application method of big data for government statistics; to qualitatively the available room for improvement in each link of government statistics process, including data collection, data analysis and data issuing, so as to pertinently put forward how to use the big data to improve the current situation.

3. Result and conclusion

(I) Stating the differences between the thinking contained in big data and the government statistics

Besides the concept and characteristics of big data, the differences between the thinking contained in big data and the government statistics should be also disclosed. The big data represents the data-driven thinking instead of any preset hypothesis. The data should speak for itself. In other words, after the data size reaches a certain extent, the regular and trend-oriented information will appear, and such information is always unexpected. The government statistics represents a thinking of hypothesis verification. In other words, the sample data is collected and analyzed to infer the overall situations or verify the original hypothesis. The differences between them are shown as follows: (1) the thinking of big data has higher tolerance for imprecision, in which the precision requirement for individual data is relatively low but the trend manifested after all the data is gathered together is emphasized; (2) the thinking of big data expands the application range of simple algorithm, and the big data has a lower dependence on algorithm if compared with the government statistics; (3) different processing method for "useless" data, the government statistics is used to eliminate or reduce the general impact of "useless data" on inference by replacing or expanding the sample data and optimizing the algorithmic routine, but it is thought in the thinking of big data that each kind of data is provided with valuable information and the "useless" data is utilized by means of reverse thinking.

(II) Summarizing the application characteristics of big data in enterprises

The application characteristics of big data in enterprises are shown as follows:

(1) mining and analysis of massive data (i.e., "sample=population), indicating that new information may appear after the data size reaches a

certain extent; compared with the current era of big data, the past can be called the era of small data, in which some results with a tendency could not be manifested and many conclusions were hard to be reached due to the limitation of data size. With an unprecedented increase of data acquired, the information contained in the data has also seen explosive growth. In the analysis of massive data, the meaning of single data seems to be less important, but the full data aggregated will show an effect of “1+1>2” and then generate the other new information outside this area.

(2) Reuse of historical data, meaning that the potential value of data is continuously mined with each mining as the data that is used only once will not become worthless; in the traditional thinking, the data that is used only once will be stored and filed after reaching the specific purpose, but the era of big data tells us that the data can be reused, and the potential value of data can be continuously mined with each mining to bring more useful information.

(3) Reciprocal recombination of data in different areas, meaning that the use of data is not limited to one area and new valuable information may be generated if the data in this area is used into another area; for example, through the exchange about flight information and weather information on the website of FlyOnTime and the analysis of historical big data, it was concluded that the delay time due to heavy fog is twice as much as that due to snow for the flight from Boston to Laganrdia Airport, New York. Therefore, the possibility of flight delay and the delay time can be more accurately speculated through the application of data exchange.

(4) Reverse use of “useless” data, meaning that the “useless” data is used by means of reverse thinking so as to make up the forward use. The spelling checker of Google is known perfect in the world. Its strong background database contains the spelling mistakes entered into the search box from 3 billion queries processed every day and then informs the system of the content actually entered by users through feedback loop so as to display the related correct spelling results¹. The wrong spelling that seems “useless” is highly related to the correct spelling actually. The database is updated in real time through reciprocating feedback loop.

We can take inspiration from the application of big data in enterprises that the big data is three-dimensional and thus different information can be mined from different dimensions. In terms of the data size, the tendency information can be shown through the aggregation of massive data; and in terms of the time dimension, the reuse of the past historical data will provide some reference for future prediction; in terms of the field interaction, more accurate

¹ Big Data Era: Big Changes In Life, Work and Thinking: 144-145, Zhejiang People's Publishing House, 2017.12

and appropriate conclusions can be drawn through the reciprocal recombination of data between different fields; in terms of the way of thinking, the perception of useless data should be changed as any data in the big data contains the "information" and the information that seems "wrong" is just the evidence for "correct" information. The value of big data requires frequent mining and use from different angles, so the potential value will be released continuously.

(III) Analyzing the application space of big data in the business process of government statistics

Currently, the application of big data in government statistics mainly focuses on specific fields, and the government statistical methods are perfected by technical means. For example, in the population statistics, the big data from mobile communications is used to realize the dynamic monitoring of population; in the agricultural statistics, the remote sensing technology is applied to obtain the information about crop planting area, floor area of facility agriculture and yield prediction of crops, etc.; in the real estate price statistics, the data endorsed on the internet is used in most of the large and medium-sized cities to calculate the price index of new houses and second-hand houses; in the transportation statistics, the big data generated from the highway network monitoring system is used to reckon the highway freight information.

A complete set of business processes has been formed in practice for government statistics over the years, which is mainly divided into three links, namely data collection, data analysis and data issuing. The whole business system is quite mature and basically meets the social statistical needs. Therefore, it is thought in this paper that the application of big data in government statistics should not be limited to the level of technical means and also the specific fields of statistics under the current business process system of statistics and a deeper application is that the thinking of big data should run through the government statistics.

By taking the statistical business process as an entry point, the experience and methods are drawn from the application of big data in enterprises to realize the perfection of data collection, data analysis and data issuing based on the analysis on expandable application space of big data in each link, improve the data quality for government statistics, deepen the data analysis and enhance the foresight of trend analysis.

In the data collection process, a large amount of source data is manually collected, the data quality is interfered by human factors, and most of data is collected after the event. Therefore, such data has poor timeliness to reflect the hot spots and emergencies and low predictability for future trend. It is hereby proposed to optimize the data collection methods by means of big

data and improve the quality of source data. Through the collection of unconventional data such as hot spot and emergency, this paper attempts to use big data to realize the transformation from post statistics to prior statistics. The conventional data is collected through the integration of big data instead of manual collection so as to improve the quality of statistical source data.

In the data analysis process, the quantitative analysis for a specific industry is manifested as comparison of major indicators on a year-on-year basis and comparison of data on a month-on-month basis, and the qualitative analysis on overall development depth of the industry is absent. The analysis of each industry by government statistics shows the modularized characteristic, and all the professional data lacks integration and connection. The analysis on overall economic operation is manifested as the listing of various kinds of industry analysis, and the correlative impact between the industries is not taken into consideration. Therefore, the data analysis is innovated by using the thinking of big data. For this purpose, it is required to explore the historical data and promote the longitudinal deepening of data analysis, strengthen the data exchange between different areas and break the modularized analysis, and attempt to make "full data" analysis by using the big data technology.

In the data issuing process, the government statistics lays emphasis on the information disclosure form, and the data issued is mostly the aggregated data processed, so this data cannot be directly used by enterprises and the public, and the data value cannot be fully mined. Therefore, the data issuing method is innovated by means of big data, and attention should be paid to the reuse of data. For this purpose, it is required to introduce the real-time dynamic issuing method to enhance the visibility of data, and perform the standardized processing of issued data to facilitate the reuse of data by enterprises and the public and realize data sharing and efficient utilization.

(IV) Putting forward the guarantee for deepening the application of big data

The guarantee measures are taken from legal system, personnel and technology:

First, strengthen the construction of statistical legal system and perfect the environmental guarantee. The government statistics shall be used to strengthen the top-level design, standardize the behavior of each subject in the application of big data from the aspect of legal mechanism, stipulate the responsibilities and obligations of data provider and user in detail, specify the scope of data resource sharing between the enterprises, between the enterprise and the government and between the government departments, and actively promote the open sharing of big data among various subjects so as to provide guarantee and support for deepening the application of big data in government statistics and protecting the personal privacy and trade secret.

Second, improve the quality of statistical personnel and strengthen the personnel guarantee. The government statistics should focus on cultivating the big data analysis technology and thinking of statistical personnel to improve their quality. Special big data collection team, big data analysis team and big data issuing team should be established to professionally cultivate the statistical personnel for big data. A communication mechanism with universities and enterprises should be established to periodically learn and share the latest development trend of big data and promptly update the theoretical knowledge and practical experience on big data. The cultivation of inter-disciplinary personnel should be strengthened in a professional and cooperative way, and the integration of big data technology and thinking into government statistics should be also accelerated.

Third, promote information construction and provide technical guarantee. Firstly, research and develop the big data analysis technology to realize the selection and integration of data with multiple sources and indicators, and create a networked and easy-to-use analysis tool set to analyze the data from multiple dimensions such as time, space and professional field. Secondly, enhance the technical support of the platform for data mobility, realize the direct import of data into each port and platform under the unified data flow format and standard, and truly realize data sharing between the divisions and between the departments. Thirdly, increase the input on research of automatic data capture technology, and realize the interconnection between the platforms and between the ports. The statistics collection platform can automatically identify the fields accurately from the enterprise port, and capture and integrate the corresponding data in a real-time manner to reduce the influence of human factors on data quality.

References

1. ZouYing(2018). Research on thinking reform of government statistics in the era of big data [J].Modern Economic Information.
2. LiuYangjian(2018). Research on reform of government statistics under the background of big data [D].Journal of Nanchang University.
3. Zhang Yong(2018). Research on the influence of big data on statistical work of grass-roots government and countermeasures[J].China Market.
4. SongRuojin(2018). Research on the issues involving the statistical work of Liaoning Municipal Government and strategies under the background of big data [D].Journal of Shenyang Normal University.
5. YuHuiyong(2018). Application of big data in government statistics [J].Journal for Party and Administrative Cadres.
6. LvRuilin(2018). Analysis on transformation of government statistics in the era of big data [J].Science & Technology Economy Market.

7. WangYiting(2017). Research on the reform of China's government statistics under the background of big data [D].Journal of Jilin University of Finance and Economics.
8. Big Data Statecraft Strategy Research Group(2017). Big Data Reading for Leading Cadres.People's Publishing House.
9. Victor Mayer-Schonberg, Kenneth Cukier(2017). Era of big data: revolution of life, work and thinking (translated by Sheng Yangyan and Zhou Tao). Zhejiang People's Publishing House.
10. WangZhenjie(2017). Application of big data in government statistics [J].China Management Informationization.
11. WangWenpeng(2017). Brief discussion on application of big data in government statistics [J].Statistics & Consultation.
12. ZhaoYingshu, Dai Mingfeng(2017). Transformation of government statistics in the era of big data and its thinking [J].China Statistics.
13. ZhangQian(2016). Research on statistical function transformation of local government under the background of big data[D].Journal of Minzu University of China.
14. ZhangHongjian(2016). Thinking about government statistics in the environment of big data [J].China Statistics.
15. LiPu(2016). Research on government statistics under the background of big data and countermeasures [D].Journal of Shenyang Normal University.
16. CuiPengda(2015). Reform of government statistics in the era of big data [D].Journal of Jilin University.
17. Li Fang(2015). Research on the application of big data in government statistics [D].Journal of Yunnan University.
18. MaJiantang(2015). Big data: new opportunities for government statistics. China Statistics Press.
- 19.XiaoGan(2014). Research on the impact of big data on government statistics and its countermeasures [D].Journal of Nanjing Tech University.
19. LinChenqi(2014). Discussion on the application of big data in government statistics [D].Journal of Fuzhou University.



On the use of surrogate models to speed the ABC inference for epidemic models in networks



L. Leticia Ramírez-Ramírez¹, Rocío Maribel Ávila-Ayala², Arturo Márquez-Cerda³

¹Centro de Investigación en Matemáticas, Guanajuato, Gto. Mexico

²Universidad Nacional Autónoma de México, Mexico City, Mexico

³Universidad de Guanajuato, Guanajuato, Gto, Mexico

Abstract

In order to explain outbreak evolutions that largely deviate from the results provided by epidemic models based on the law of mass action, the epidemic models in a network of contacts have been introduced to generalize them. These models directly incorporate a population contact pattern that dictates the potential infections between individuals and can be modeled based on geographical distance and some other social patterns in the population. The epidemic model increases its complexity with this pairwise contact network structure and in most cases, obtaining its likelihood function is extremely expensive or impossible. In this work, we present a statistical inference of epidemic compartmental models based on the Approximate Bayesian Computation (ABC) that is likelihood-free and we propose some surrogate versions to reduce the required computing time for inference. We illustrate this proposal with computational experiments of epidemic outbreaks in simulated networks.

Keywords

Epidemic models; Networks of contacts; Likelihood-free method; MCMC; Surrogate models; Recurrent Neural Networks.

1. Introduction

The SIR epidemic model considers that an individual can transit through different status when infected during an outbreak. Let $S(t)$, $I(t)$ and $R(t)$ be the number of susceptible, infected and removed individuals at time t , respectively. In a closed population, the population size remains constant ($N \equiv N(t), t > t_0$) and the state space $\{S(t); I(t); R(t)\}$ can be monitored using only $\{S(t); I(t)\}$. Let $p_{(s,i)}(t)$ be its corresponding joint probability function $p_{(s,i)}(t) = P(S(t) = s, I(t) = i)$, where (s, i) is a vector of possible $(S(t), I(t))$ states.

If the population is well-mixed, we can assume that the interactions between infective and susceptible individuals are fully dictated by their number, and the transmission and removal (recovery) rates β and γ .

Using the (forward) Chapman-Kolmogorov equation we have the (Chemical) Master equation:

$$\frac{dp_{(s,i)}(t)}{dt} = \frac{\beta}{N}(s+1)(i-1)p_{(s+1,i-1)} + \gamma(i+1)p_{(s,i+1)} - \left[\frac{\beta}{N}si + \gamma i\right]p_{(s,i)} \quad (1)$$

The reproductive number, R corresponds to the expected total of new cases originated by a typical infected case. Based on model (1) it is equal to β/γ . This number is paramount for determining if the outbreak will infect few individuals or if it can develop into an epidemic.

Understanding the transmission process and its parameters can help us to identify outbreak interventions that can reduce R below the threshold value. In this work, we assume the SIR compartmental model and evolving in a network of contacts that is fully known. The objective is estimating the infectious agent SIR parameters based on surveillance-alike information. That is, we only observe the time-aggregated counts of new cases entering status I .

We also propose the introduction of surrogate epidemic models that can reduce the computational burden of the original ABC. This is compared to the regular ABC using simulated data on random networks.

2. Methodology

2.1 Epidemics in Networks

In the considered SIR epidemic model in a simple network $G(V, E)$, the infectious agent is independently transmitted between pairs of connected individuals. This transmission event occurs from an infective to a connected susceptible individual with rate $\delta > 0$. On the other hand, infective individuals recover independently after a time that is exponentially distributed with parameter $\lambda > 0$.

In this model, the reproductive number is $R_{net} = \nu E(K_1)$, where K_1 is the excess degree (Newman, 2002) and $E(K_1) = E(K^2)/E(K) - 1$ under the configuration model, where K is the degree of a randomly selected vertex. The parameter ν corresponds to the probability of transmission between a susceptible and an infective individual, throughout the infectious period of the latter. That is

$$\nu = \int_0^\infty (1 - e^{-ru})dF_1(u),$$

where $F_1(\cdot)$ is the infectious period. Then

$$R_{net} = \frac{r}{r+\lambda} E(K_1).$$

(2)

2.2 ABC-MCMC

The ABC (Del Moral, et, al. 2012) is a class of methods that provide an alternative to the likelihood computation. It is a rejection sampler for the posterior distribution $f(\theta|\mathbf{y})$ of the parameters θ based on the observed data \mathbf{y} . In this work we focus in the ABC-MCMC (Marjoram, et al. 2003) that can be implemented without specifying likelihood function but when it is easy to sample $x \in D$ from $f(\cdot|\theta)$, for any θ in the parametric space Θ . Since this method is "likelihood free", it is gaining a lot of attention in a wide range of scientific disciplines where their complex models have intractable likelihoods, or they are too expensive to calculate.

The ABC methods use the modified posterior density on $\Theta \times D$

$$\pi_\epsilon(\theta, \mathbf{x}|\mathbf{y}) \propto \pi(\theta)f(\mathbf{x}|\theta)I_{A(\epsilon, \mathbf{y})}, \quad (3)$$

where $I_{A(\epsilon, \mathbf{y})}$ are the pseudo-observations x that are "closer up to ϵ " to the true observations y (Marin, et, al. 2012). Formally $A(\epsilon, \mathbf{y}) = \{z \in D: \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \epsilon\}$, where $\eta(\cdot)$ summarizes the observations, and ρ is a distance function.

A modification to (3) replaces the indicator by a kernel function K_h evaluated on a joint statistic for x and y . That is

$$\pi_h(\theta, \mathbf{x}|\mathbf{y}) \propto \pi(\theta)f(\mathbf{x}|\theta)K_h(t(\mathbf{x}, \mathbf{y})),$$

where h is the kernel bandwidth and $t(x, y)$ is a discrepancy measure between observations. Then the Metropolis-Hasting can be modified to perform ABC-MCMC inference, considering its acceptance probability for candidate $\tau \in \Theta$ equal to

$$\alpha(\theta|\mathbf{y}) = \min \left\{ \frac{\pi_h(\tau|\mathbf{y})q(\theta^{(t)}|\tau)}{\pi_h(\theta^{(t)}|\mathbf{y})q(\tau|\theta^{(t)})}, 1 \right\} = \min \left\{ \frac{\pi(\tau)K_h(t(\mathbf{x}_\tau, \mathbf{y}))q(\theta^{(t)}|\tau)}{\pi(\theta)K_h(t(\mathbf{x}_{\theta^{(t)}}, \mathbf{y}))q(\tau|\theta^{(t)})}, 1 \right\},$$

where $\mathbf{x}_{\theta^{(t)}}$, \mathbf{x}_τ are pseudo observations from $f(\cdot|\theta^{(t)})$ and $f(\cdot|\tau)$, respectively, and q is the proposal distribution.

In the SIR epidemic model in networks, the parameter of interest is $\theta = (r, \lambda)$. For arbitrary networks, the likelihood is intractable but using the agent-based and event-driven algorithm used in Ramírez- Ramírez, et al. (2013), we can obtain pseudo observations for any $r, \lambda > 0$.

2.3 Recurrent Neural Networks

Deep Neural Networks are gaining popularity because they have shown good results in diverse problems like in the areas of object and speech recognition (see Krizhevsky et al.2012 and Dahl et al., 2012 for examples). They are machine learning technique oriented to learn high-level abstractions by using hierarchical architectures. This architecture is based upon the classic approximation theorem by Cybenko that states that perceptrons with a hidden layer of finite size and sigmoid activation functions can approximate complex

continuous functions and was implemented using back-propagation algorithm in multilayer perceptrons to overcome the classic XOR problem.

Recurrent Neural Networks (RNNs) have been successfully used in prediction of time ordered data (see da Silva et al., 2017). Cho et al. 2014 defines a Recurrent Neural Network (RNN) as a neural network with a hidden state that operates on a variable length sequence. Here, the hidden state is updated using its current value and a new input sequence. If we segment the sequence in fixed length subsequences, and feed them to the network, the hidden state will depend on the past observed sequence values. This structure then allows training the stimuli on previous sequence values, which is why these networks are referred as having *memory*.

We use neural networks to find an efficient forecaster for an outbreak given the network of contacts and the infectious agent parameters. To train the RNN, we generate random networks using the configuration model on degree sequences sampled from probability distributions (1) Poisson, (2) Power Law and (3) Polylogarithmic (Newman, 2002); and graph generation models: (4) Watts- Strogatz and (5) Barabasi-Albert. These were selected with parameters such that we originate diverse topological network features. On each network, we simulate a SIR epidemic outbreak with parameters varying in a range of values reported in diverse literature for influenza (Cori et al., 2012 and Biggerstaff et al., 2014). From the observed outbreak we recreate the surveillance information, where the individuals entering state *I* are aggregated at regular time intervals.

We employ Gated Recurrent Unit (GRU) layers Cho, et al. (2014) in order to prevent the vanishing/exploding gradient problem that affects standard neural networks. We use the Mean Squared Error (MSE) as error metric for training and evaluation of our models. The model is fed with information on *k* consecutive time series points of the number of individuals in each compartment, and some network topological features (Mean, Variance and Asymmetry of its degree sequence; order and size of the network, edge density, clustering coefficient, assortativity and variance of its eigencentality coefficients).

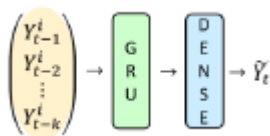


Figure 1a: Initial RNN architecture.

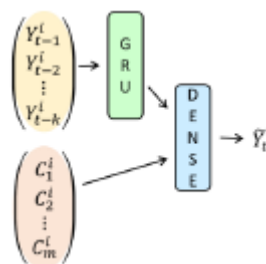


Figure 1a: Initial RNN architecture.

The recurrent layer transforms the input sequence into a fixed length vector $(Y_{t-1}^i, \dots, Y_{t-k}^i)$ that encodes relevant features of the time series and the “memory” of its hidden state about previously seen sequences (Figure 1a). These features are combined with the structural properties of the pertaining graph (C_1^i, \dots, C_m^i) and feed into a dense layer that generates the model’s prediction (Figure 1a). is modified to consider the contact network structure or network topology, as constant variables (Figure 1b).

3. Result

3.1 Exact simulations

To assess the ABC and the proposed modifications, we recreate two set of synthetic data on epidemics evolving in two different random networks (each with 500 vertices) with degree distribution: (A) Poisson (2.42) (B) Polylogarithmic (0.1, 2). For each network we simulate the outbreak surveillance information using parameter values $r = 0.03$ and $\lambda = 0.01$, and initial states $(S(0), I(0), R(0)) = (N - 2, 2, 0)$. The generated data is then set “real data”. As Dutta, et al. (2018), we assume the knowledge on the number of initial cases in I and R but we do not specify the individuals in these states. In contrast with Dutta, et al. (2018), we perform the statistical inference only from the surveillance-like reports.

The ABC-MCMC described in Section 2.2 is implemented with discrepancy measure

$$t(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{\frac{(x_i - y_i)^2}{y_i I(y_i > 0) + \varepsilon I(y_i = 0)}}$$

where $\varepsilon > 0$ is a small constant introduced to define $t(\mathbf{x}, \mathbf{y})$ beyond the observed outbreak span.

We use a Gaussian kernel function and the proposal distribution q corresponds to a mixture of Gaussian densities. This q allows to include two distributions with different standard deviations (in our case, 0.005 and 0.1) to improve the parametric space exploration. For these experiments, we consider that r and λ have independent prior exponential distribution with parameters 3, that is equivalent to be approximately 95% confident that each of the real values are between 0 and 1. After removing the burn-in and thinning the series, the parameters sampled from the posterior produce the statistics in Table 1. We observe that for the two networks, the 95% posterior intervals contain the true parameter values.

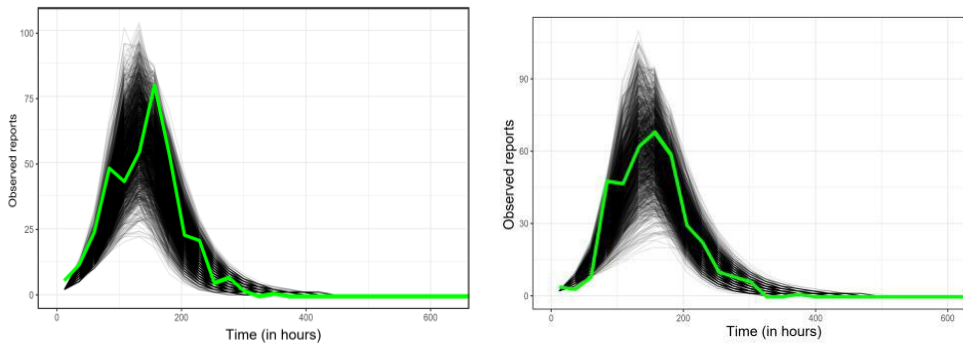
	A: Poisson		B: Polylogarithmic	
Quantile	r	λ	r	λ
2.5 %	0.0250	0.0015	0.0176	0.0008
50 %	0.0333	0.0127	0.0249	0.0085

97.5 %	0.0622	0.0341	0.0411	0.0269
---------------	--------	--------	--------	--------

Table 1 Statistics of posterior ABC samples.

3.2 Surrogate model A: “Mass action”

Since the replacement number is related to how the number of cases increase in a very large population, it seems reasonable to try to approximate the network epidemic model by a deterministic model (1) with parameters that reproduce this number. If we select the parameters β and γ in (1) with values $\beta = rE(K_1)$ and $\gamma = r + \lambda$, then we have $R = R_{net}$. Then at each step of the ABC-MCMC, we replace the agent-based simulation for the proposed parameter $\tau = (\tau_r, \tau_\lambda)$ by the numerical solution to the deterministic model with parameters $(\beta = \tau_r \overline{K_1}, \gamma = \tau_r + \tau_\lambda)$. The aggregated new infective cases based on the solution to the deterministic model with accepted ABC-MCMC parameters are depicted in Figure 2. The green lines represent the synthetic data y .



Network A: Poisson.

Network B: Polylogarithmic.

Figure. 2: Simulated reports from outbreaks with sampled approximated ABC-MCMC parameter values, using the surrogate model (1).

The resulting statistics for the accepted posterior parameters are presented in Table 2. We can observe that the intervals can be slightly biased and wider, and except for r in the Poisson network, they include the true parameters.

	A: Poisson		B: Polylogarithmic	
Quantile	r	λ	r	λ
2.5 %	0.0306	0.0040	0.0216	0.0032
50 %	0.0413	0.0218	0.0290	0.0220
97.5 %	0.0616	0.0521	0.0385	0.0423

Table 2 Statistics ABC samples with of surrogate model “mass action”.

Regarding the computational cost, the direct ABC-MCMC required approximately 144 minutes for each of the two synthetic data sets, while the

surrogate-ABC-MCMC took less than 2 minutes to obtain the posterior sample.

This ABC-MCMC modification becomes highly attractive not only because it is fast to compute, but because it only requires partial knowledge of the network, that is its first and second moment of its degree.

3.3 Assessing surrogate model B: “RNN”

The proposed RNN architecture allows us to predict the dynamics of epidemic outbreaks on contact networks with similar degree distributions (in the experiments depicted in Figure 3 we used networks with Poisson generated degree sequences). Although this model requires a considerable amount of time to train (approximately 5 minutes for two sets of simulations), it offers fast prediction as it took an average of 38.5ms per simulation on a i7-8650U processor.

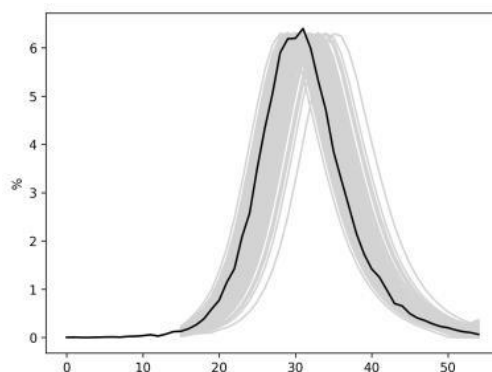


Figure 3: Predicted outbreaks with our augmented RNN architecture.

4. Discussion and Conclusion

We base the likelihood-free inference ABC-MCMC on more realistic surveillance-like information, that report only the new infective in intervals of time. We emphasize on the inference for the SIR parameters r and λ , but using this same ABC methods, we can do the inference of any epidemic model from which we can obtain pseudo-observations. The ABC-MCMC can be very computer expensive if it uses the agent-based simulation, but in this case, many other characteristics can be inferred, such as the degree of the initial infectious cases, the most exposed community, etc. This approach also allows doing inference for non-Markovian epidemic models.

We explore the use of surrogate models to run more efficient ABC-MCMC. For the specific case of SIR (or SEIR) models, we propose harnessing model (1) but other alternatives must be used when the model has infectious (latent) periods are not exponentially distributed. In these more general scenarios, we propose some models that allows for fast forecasting, while they can incorporate the network topological features that are relevant for the outbreak evolution. With this in mind, we propose using RNN that can be trained on

simulated outbreaks with a variety of networks of contacts and epidemic parameters. Once trained, the predictions are very efficient to obtain.

The motivation of the use of surrogate models is clear and the results are promising, however it is extremely important to study the error that arise from their introduction, in order to control it to our needs.

References

1. Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., and Finelli, L. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious diseases*, 14(1): 480.
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
3. Cori, A., Valleron, A., Carrat, F., Tomba, G. S., Thomas, G., and Boëlle, P. (2012). Estimating influenza latency and infectious period durations using viral excretion data. *Epidemics*, 4(3):132–138.
4. da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., and dos Reis Alves, S. F. (2017). *Artificial Neural Networks: A Practical Course*. Springer.
5. Del Moral P, Doucet A, Jasra A (2012) An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing* 22(5):1009–1020.
6. Dutta R, Mira A, Onnela JP (2018) Bayesian inference of spreading processes on networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 474(2215):20180,129
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS'2012)*, pp 1097-1105.
8. Marin, J.-M., Pudlo, P., Robert, C. P., y Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.
9. Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26), 15324-15328.
10. Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, 66(1), 016128.
11. Ramírez-Ramírez L.L., Gel Y.R., Thompson M, de Villa E., McPherson M. (2013) A new surveillance and spatio-temporal visualization tool SIMID: Simulation of infectious diseases using random networks and GIS. *Computer methods and programs in biomedicine* 110(3):455–470.

12. Tang, L., Wang, X., and Liu, H. (2009). Uncovering groups via heterogeneous interaction analysis. In ICDM '09: Proceedings of IEEE International Conference on Data Mining, 503–512.



Adding statistical value in a rapidly changing government data “Eco-system”



Vince Galvin, Charlie Russell, Joe Winton
Statistics NZ, Wellington, New Zealand

Abstract

New data sources, new data analysis methods and new tools provide considerable opportunity to improve the design and implementation of public policy programmes. However, the technical challenges of making improvements have to be put in the context of all the institutional arrangements that have always made realising the potential of using data challenging. This paper will describe the experiences of Statistics New Zealand in working with other Government agencies to find a mix of principle statements, standards, creating networks and developing capabilities. The central themes will be that of trying to come to grips with the complexity of policy environments and trying to find practical ways ahead.

1. Introduction

This paper discusses Stats NZ’s journey so far as a National Statistical Office (NSO) as we broaden our role in New Zealand’s data system. It aims to look at this experience through a technical statistical lens. It examines the role that Stats NZ has played as the wider New Zealand Government works through how to realise the potential of new data sources, methods and tools. It highlights the criticality of some points that have been important to what has happened in New Zealand:

- the acquisition of a mandate for Stats NZ to take on a wider role in leading the wider Government.
- the “game changing” nature of the having a centerpiece development like the Integrated Data Infrastructure (IDI).
- the opportunities to participate and influence large scale data analysis projects and initiatives.

This paper is a discussion of our experiences to date, followed by some reflections on what we have learnt from the activities we have undertaken.

2. Context of Statistics New Zealand’s role in the wider data System

Stats NZ received an explicit mandate to be a leader of the NZ Government wide data system. Since the establishment of the Government Chief Data Steward (GCDS) role in 2017, Stats NZ has worked in partnership with government agencies to:

- set the strategic direction for government's management of data with the development of a Data Strategy and Roadmap focused on unlocking the value of data for all New Zealanders¹
- lead the state sector's response to new and emerging data issues, most recently by fast-tracking work to strengthen transparency of, and accountability for, government's use of algorithms.

In September 2018, the government (through a cabinet mandate) agreed to empower the GCDS to:

- set mandatory standards and guidelines for the collection, management and use of data by government agencies; and
- direct agencies to adopt common data capabilities, such as data tools, linking infrastructure, or sharing platforms (subject to an opt-out process)².

These new powers are intended to establish the foundations necessary for agencies to manage and share data so that it provides a useful base to inform decisions, within existing privacy and security settings.

In NZ there are already pockets of significant progress in the way that data is managed and used across government. However, the system as a whole has not been designed and managed with all-of-government needs in mind. Often agencies focus on supporting their own operations as effectively and efficiently as possible. At the same time, increased use of sophisticated data analytics to drive decision making (e.g. artificial intelligence and algorithms) within government elevates the need for strong common data practices to maintain trust and confidence, ensure privacy is protected, and to foster the ethical use of data.

3. Explicit Mandate to Conduct a Review – Algorithmic Transparency

A good example of Stats NZ using its new role and its acknowledged competencies in data system issues was undertaking a review of the use of algorithms in programme delivery.³

How did this happen?

Concern about data use led the New Zealand Government to want to 'take stock' of the use of algorithms by government agencies.⁴ The Minister for Government Digital Services and the Minister of Statistics commissioned an

¹ Data Strategy and Roadmap <https://www.data.govt.nz/about/data-strategy-and-roadmap-setting-the-direction-for-new-zealands-data/>

² Cabinet paper – Strengthening Data Leadership <https://www.stats.govt.nz/corporate/cabinet-paper-strengthening-data-leadership-across-government-to-enable-more-effective-public-services>

³ Government Algorithm Transparency <https://www.data.govt.nz/use-data/analyse-data/government-algorithm-transparency/>

⁴ Panel Presentation by the Privacy Commissioner, John Edwards, to the Data Analytics Forum on Ethics and Algorithmic Transparency on 18 July 2018 in Wellington <https://www.privacy.org.nz/assets/Files/Speeches-presentations/2018-07-18-Data-Analytics-Forum-Panel-presentation.pdf>

assessment of how government agencies use algorithms to analyse people's data, to ensure transparency and fairness in decisions that affect citizens.

The focus of the review was primarily social agencies using algorithms that directly impacted people. The review consequently covered 14 agencies and assessed algorithms against the GCDS' and Privacy Commissioner's Principles for the safe and effective use of data and analytics.⁵

What did the review find?

The review found that agencies use a range of algorithms in their day to day operations and that all of the algorithms described by agencies are embedded in policies that are intended to deliver public benefit. These include: improved efficiency, streamlining processes, proactively targeting specific support to individuals, supporting decisions, protecting New Zealand from risks and threats, and providing assessment or forecasting for policy development.

It also found there are opportunities to increase collaboration and sharing of good practice across government to ensure that all of the information that is published explains, in clear and simple terms, how algorithms are informing decisions that affect people in significant ways.

In broad terms the nature of the conclusions was that agencies were acting with the right sort of issues in mind but that they were acting separately. This highlighted the risks of lines being drawn about ethical behaviour separately across Government and the potential inefficiencies of agencies separately re-inventing the wheel. In general, it reinforced the message of a competent but siloed approach to data and analytics issues.

The review was an example of Stats NZ operating outside our usual mandate as a producer of statistics. It demonstrates that there are direct connections between the foundations elements of good statistical practice and a broader role of helping the New Zealand Government develop a well-functioning data system.

4. The Centrality of the Integrated Data Infrastructure (IDI)

With cross government data initiatives, the issue generally arises how the National Statistical Office can obtain leverage as it tries to engage agencies in data issues. In New Zealand, apart from the explicit mandate provided by the GCDS role, a critical aspect of acquiring leverage comes from how central the IDI has become to how Government makes investments in its social programmes.

⁵ Principles for the safe and effective use of and analytics <https://www.privacy.org.nz/news-and-publications/guidance-resources/principles-for-the-safe-and-effective-use-of-data-and-analytics-guidance/>

The scope of the IDI

The IDI contains deidentified person-centred microdata from a range of government agencies, Stats NZ surveys (including the 2013 Census), and non-government organisations. To integrate datasets, Stats NZ links information together using identifiable data, including first and last name, date of birth, age, sex, and country of birth. There are eight broad categories of data in the IDI, covering Health, Education and training, Benefits and social services, Justice, People and communities, Population, Income and work and Housing data.

Key aspects of providing the IDI as a service.

- We gave priority to data expansion over system usability investment

A unique combination of circumstances arose that provided Stats NZ with the opportunity to be entrusted with extensive data holdings from across Government. The wider political environment meant that for a period of time agencies across the New Zealand Government felt some compulsion to get their data into the IDI. Stats NZ took the view that we would be better placed to take advantage of this circumstance than focus on creating a perfect data base. The consequence was that the IDI took in data from 40 agencies, rather than the 10 that had originally been planned but there have been some struggles to optimise the performance of the IDI software systems. Without taking this opportunity, Stats NZ may still be arguing over the design for an integrated data system.

- Establishing a sense of collective ownership

The IDI team within Stats NZ is embedded at nearly every step of the production process, bearing the bulk of the workload rather than spreading it across the system. Fewer than 20 of the almost 300 currently active IDI projects (April 2019) are undertaken by Stats and along with stakeholders and partners Stats NZ views the IDI as a system asset. Partnerships, both within and outside Government are essential, but are currently limited to prioritising new data sources and some sharing of findings.

More active partnerships will have two main benefits. Firstly they will ensure that the maintenance of the system, and continued innovation, is a system-wide undertaking and is not compromised by one ministry or sector's priorities; and, secondly, they will help increase trust and buy-in from the people represented in the data, to ensure research that flows from the IDI is sustained and trusted.

- Stimulating Community building among researchers

The group of people who have become users of the IDI have the sort of diversity that might be expected. This community can be broadly described as IDI Experts, Subject Matter Experts, Methodological researchers and production Modellers. It became clear early that it was important to create

“virtuous cycles” of information sharing. As agencies make more use of their own data in the context of it being linked to other agencies data, the drive to standardise practices arises out of the need to solve their own business problems

In New Zealand it has become clear that individual researchers who are more active in maintaining personal networks are doing better in making good use of linked data. They are able to overcome dependency on agencies and individuals within agencies to get enough metadata and general intelligence about using data files to produce work that crosses agencies and sectors.

The Impact of the IDI and our biggest Lessons

The IDI has become a corner stone of the work of the social sector in Government in New Zealand. Social agencies compiling business cases are required to have an evidence base of the wider aspects of their proposals and this inevitably requires them to use the IDI. The biggest impact has been to raise the bar for what constitutes adequate evidence in a business case.

From the perspective of the provider of the service the experience has highlighted the enormous value of getting started and learning as you go. Taking the opportunity to build up extensive data resources has been at the core of the value of the IDI and had created a sense of “potential that needs to be realised” in the wider government.

5. Understanding the work being done by the analyst community

With Government acting so separately, no small effort is required to keep track of what is happening in the analytics community around us. Stats NZ looks to expand its ability through participating in the local analytics community as activity as we can;

- By providing leadership to attempts to build communities of analysts

This has taken the form of trying to help establish a Government Analytics Network (GAN), taking a leadership role in the Artificial Intelligence forum and supporting a wider community of interest known as the NZ Data Analytics Forum. The experience of participating in these networks has been very mixed, with it usually being the case that information sharing is much easier to achieve than collaboration unless an imperative to resolve a pressing issue exists.

- Participating in the market as a provider and purchaser of services

Stats NZ has established its own Data Ventures unit⁶, to look at how we can explore commercial opportunities to acquire data and develop products around the data. We are also investing in developing our consulting function. In the course of executing our own work we have also spent considerable time talking to Analytics consulting firms about what data sources, methods and

tools they use. This is very valuable for getting a sense of what new possibilities are being considered and deployed.

- Technical assistance committees

As a range of agencies have looked to invest in large scale microsimulation models based on the IDI they have established technical advisory committees alongside the project team. Stats NZ is invited onto these committees and they provide important insights into the “cutting edge” attempts to deliver value to the policy development process. *The main lessons highlighted from this work were:*

- Agencies are considering doing their work very differently and unless Stats NZ understands what is being considered then we run the risk of being irrelevant to the attempts of agencies that are key clients to achieve their goals.
- The appetite to obtain insight into what practitioners are doing and what it actually achieves is nearly insatiable. Interestingly very little of the enthusiasm to come and listen to others speak has translated into functional arrangements for collaborating.
- This experience echo’s many of the issues that were referred to in the Algorithm stocktake report. Very similar challenges are being faced in each agency and they are being tackled separately.
- Despite many acknowledging the benefits of collaboration and consistency in building these models, there are not yet incentives to do so meaningfully which outweigh the needs of individual agencies.

6. Deploying Analytical Capability in Public Policy

Population Segmentation has been the Success Story

Significant value has been achieved by simply describing how a wider system of public policy works. Describing who goes through which process, what happens to them in the system and how they fare with their health or labour market outcomes, continues to provide the best possible basis for examining the need for change. Examples of the types of insights researchers are gaining include:

- Exploring the gender pay gap, and the impact of parenthood on personal outcomes
- Using integrated data to gain a better understanding of the long-term, system-wide causes and impacts of mental health
- Analysis into cardiovascular disease such as defining who is most at risk; and how housing damage from the Christchurch earthquakes impacted cardiovascular events.

- In the first research of its kind, researchers have been able to describe the whole story about what it means for NZ babies born prematurely. and their long term outcomes.
- Oranga Tamariki are gaining insights from the IDI that will help improve the well-being of children and young people.

Trying to use large scale models to help develop Policy is still a work in progress

The most striking lessons have arisen from participating in the attempts by major policy agencies to realise value from making substantial investments in microsimulation. The approach of the Justice Sector, for example, has been shaped by the view that social outcomes are largely emergent, in the sense that the causal mechanisms cannot be fully specified. This thinking is influenced by theories which found that integrated theories tend to explain less variance than individual theories. Consequently, their microsimulation model is quite simple and designed not to explain causal mechanisms but to describe baseline trajectories of offending. The IDI is primarily useful in terms of understanding the distribution of risk across the population, and patterns of service use and co-occurrence of problems. This helps understand who the Justice Sector might want to focus on and which government services they're already interacting with.

In broadly similar vein, the Vulnerable Children's model aims to understand demand for support services, the benefits and costs of providing services, and the key drivers of people transitioning into, through and out of the support systems. The idea is broadly that in the current timeframe data can be collected on subjective measures of well-being and detailed segmentation analysis of the operation of the current system can be done, while a longer term model is used to understand whether there is a relationship between particular indicators of wellbeing and future outcomes.

Both these approaches are responses to the general problem of trying to predict how people will respond to different incentives and constraints. This is more feasible in the short term and for small changes in programmes but inherently difficult for significant life changes over an extended period of time.

The problem is trying to find practical ways of getting a sense of where these limits of what can usefully be predicted turn out to be in practice. Short term projections of service usage can be predicted from data about upstream pressure within the system, but it can be surprisingly hard to make accurate predictions about major service usage over a 4 year time horizon.

7. Reflections

- The wider environment matters, but demonstrating value matters more

The explicit system mandate is useful in our discussions but having examples and successes to point to helped demonstrate the benefits of our potential contributions. Activities like the algorithmic transparency review are very valuable for helping focus analysts in other agencies on the obvious need for collaboration.

The early focus of the IDI on adding large amounts of data over building a perfect infrastructure has led to a huge growth in the use of integrated administrative data in government policy. This would not have been possible without jumping on opportunities provided at the time. All this work progresses best when there is an imperative to act and so we took a deliberate decision to be opportunistic.

As a result the IDI became central to the process for bidding for new money in the social sector. Evidence from analysing IDI challenged critical assumptions that had historically underpinned public policy. People working in social agencies now have both a sense of considerable untapped potential and wanting to contribute to helping the IDI develop further.

- NSO's need to be a part of the user community, and significant technical challenges need community solutions

With so much changing about how Government agencies intend to go about their work, there is a question of how perfectly Stats NZ needs to be able to answer questions around what is possible and what is needed, before it can start to add value. There used to be a lot of dialogue about understanding user need in NSOs. It seems these days "what is needed" isn't sitting there to be discovered – user need is being continuously created, and NSOs need to figure out how to be part of this continuously evolving conversation.

In our work with agencies trying to predict the impact of long-term policy changes and in working with communities to help illuminate cause and effect playing out in very local environments we have encountered limits to what data can explain. Which of the limits are inherent and which can be overcome with different approaches will need well focussed collaboration to understand. Everyone is more inclined to document great success than efforts that didn't work but it seems that these limits need to be found by trial and error. This will be a better process when everyone can see what has worked and what hasn't.

8. Concluding Remarks

The famous physicist Heisenberg said that "Whenever we proceed from the known into the unknown we may hope to understand, but we may have to learn at the same time a new meaning of the word "understanding"".

We have the same sense of looking for the best ways ahead in bring statistical expertise to improvements in the New Zealand data system leadership. There are a very number of sensible initiatives in the broad categories of providing infrastructure, facilitating standardisation and facilitating the identification and utilisation of best practice that could be suggested and progressed.

It is undoubtedly critical to have these frameworks and plans for their implementation. However, it seems that sometimes value resides in unlikely places. It is essential to develop mechanisms for understanding when two ideas that seem equally sensible have had very different impacts and following these opportunities to see where they can take you. Providing solutions is the key to developing the picture of what shape our future role will take.



Challenges in implementing a new imputation method into production in the 2017 Economic Census or what to do when the research approach oversimplifies the problem



Katherine Jenny Thompson, William C. Davie Jr.
Economic Statistical Methods Division, U.S. Census Bureau.

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY19-ESMD-B00002)

Abstract

Beginning in 2017, the Economic Census (EC) will use NAPCS to produce economy-wide tabulations of product sales. This change facilitated the development of a statistically defensible and operationally feasible imputation approach for these items, along with associated methodology for estimating variances. Research teams were established to recommend methods for both. To avoid confounding treatment effects with respondent size effects, both studies restricted the analysis variables to a limited set of products within industry. This greatly simplified the evaluations, but left potential implementation challenges uninvestigated. This paper focuses on the implementation of the recommended imputation methods in the 2017 EC, discussing the modifications and enhancements needed to accommodate the complete set of variables along with the implementation issues that motivated each. We conclude by sharing our “lessons learned,” along with a laundry list of topics for further investigation before the 2022 EC.

Keywords

Hot deck imputation; multiple imputation; production implementation

1. Introduction

Improvements to official statistics programs may require complicated changes to existing methods or procedures to address new – emerging – requirements or to accommodate new requests. Such changes are always constrained. Budget constraints could force an overall reduction in sample size; methodologists would need to revise the sampling design while maintaining predetermined reliability levels on key statistics. Project sponsors might commission the collection of additional data items, request preliminary tabulations (and publication) of survey estimates, or desire subdomain

estimates not considered in the initial design. Again, these applications would be constrained by reliability and confidentiality mandates. In any case, the appropriate solution is rarely obvious; research is required. Of course, time and resource constraints can be prohibitive, so the research problem may be simplified to allow for a transparent and repeatable solution, thus delaying the unaddressed details or unforeseen nuances until the implementation stage, leaving little time or resources for additional research.

Efficiency frequently dictates the research and implementation processes. In the Economic Directorate of the U.S. Census Bureau, it is a common practice to establish a “dedicated” research team with a fixed duration comprising representatives from the relevant job series with differing experience levels, perhaps utilizing matrix management. Team responsibilities encompass defining and scoping the research problem, obtaining data, designing and conducting the research, writing and testing programs needed to carry out the research, documenting the findings, and presenting the “data-driven” recommendation. Assuming that the recommendation is endorsed, a subsequent implementation team is established. This team’s composition can differ greatly from the research team, as expert staff are required – and production programmers must be included in the discussions – although some overlap in membership between the research and implementation teams is desirable. As with the research team, the implementation team usually operates under a fixed deadline. Team responsibilities include writing specifications that implement the research recommendation while addressing the issues that were *deliberately* unaccounted for in the research. Logistical issues such as coding, testing, and validation are likewise included. An education component is not unusual, as the implementation team members may be unfamiliar with the methods under consideration.

The Economic Census (EC) is the U.S. Government's official five-year measure of American business and the economy. The 2017 EC leadership team endorsed a number of innovative updates, each introducing a new set of methodological and production implementation challenges. This paper focuses on the introduction of the North American Product Classification System (NAPCS) (see <https://www.census.gov/eos/www/napcs/more.html>). The EC collects a core set of data items from each establishment called general statistics items: examples include total receipts or value of shipments, annual payroll, and number of employees in the first quarter. In addition, the EC collects information on the revenue obtained from the sales of good and services (hereafter referred to as “products”). Although data on over 8,000 different NAPCS products are requested, evidence from prior economic censuses indicates that many products are rarely reported. Legitimate zero values are expected for the many products in an industry, at both the individual establishment and total industry levels. Respondents can report

data from a long, pre-specified list of potential products in a given industry – some lists contain more than 50 potential products – and can write in descriptions of other products that were not pre-specified. Product lists can differ by industry within a sector. Furthermore, some product descriptions are quite detailed, and some products are mutually exclusive. Consequently, some establishments choose not to report any product data (complete product nonresponse). Within an industry, reporting units tend to provide the same small set of common products (Fink, Beck, and Willimack 2015).

The introduction of NAPCS marks a major departure from the prior collections which explicitly linked product codes to industry, allowing for different missing-data treatments for products by sector. Implementing a NAPCS-based collection necessitated the development of a single imputation approach for all EC products to allow production of cross-sector tabulations. This paper describes the research process used to determine the product imputation method and the process for implementing the research recommendations into the 2017 EC production systems. Research and implementation were accomplished by two different teams, with a small fraction of membership overlap. Section 2 summarizes the research approach and resultant recommendations. In Section 3, we discuss the implementation of the recommended methods into a production system, specifically focusing on some of the unaddressed or unforeseen – but important – details that were excluded from the research study. We conclude in Section 4 with a few general observations about implementing research-based results in a production system.

2. Methodology

Thompson and Liu (2015) give an overview of the large scale research project conducted to determine a single, unified imputation method for EC broad products under NAPCS. The research was undertaken by a commissioned team whose members included methodologists, subject matter experts, and classification experts. The latter two groups developed the test data used for all analyses and provided expertise on the 2012 EC procedures; the 2017 collection procedures and NAPCS-based collection structure were under development during the time of the research. The methodologists' familiarity with the subject matter and expertise on the current procedures ranged from completely novice (the majority) to extremely knowledgeable about a selected subset of industry-specific procedures. Both team leads were methodologists who were familiar with EC processing procedures and methods in general but had little or no experience with the specific procedures used in product processing.

The team divided the project into the three separate components listed in Table 1, each lasting approximately two to three months. The project started

slowly, with the acquisition of a processing environment and historical data sets. Subject matter experts extracted the EC test data from industries provided by the classification experts. They also provided classification rules for donor records (whose values can be used for imputation) and recipient records (need an imputed value), thus ensuring that industry-specific “must-product” rules would be enforced by any imputation method. The classification experts provided industries whose product distributions were expected to remain largely the same under NAPCS. Even so, the historical product data were not expected to be perfect predictors of the 2017 product data due to numerous collection changes from the 2012 EC.

Table 1: Research Components

Component	Purpose	Leaders
Test Data Preparation and Knowledge Sharing	<ul style="list-style-type: none"> • Find test data with comparable products under 2012 EC and NAPCS • Define donors/recipients • Bring staff “up to speed” on data collections 	Subject Matter and Classification Experts
Exploratory Data Analysis (Empirical Data)	<ul style="list-style-type: none"> • Understand the “nature” of reported data to assess potential imputation methods • Understand the “nature” of missing data to assess potential imputation cells and to develop response propensity models 	Methodologists
Evaluation Study	<ul style="list-style-type: none"> • Evaluate the performance of considered imputation methods over repeated samples 	Methodologists

The team agreed to study only broad products and to limit the analyses to national-level industry estimates. Broad products can be collected in different industries, although many industry classification procedures rely on specific product categories. Detailed products are industry-specific breakdowns of these products and are not necessarily requested for all broad products. Broad and detailed products comprise nested one-dimensional balance complexes. The broad product values within a given establishment are expected to sum to the total receipts value reported earlier in the questionnaire. Detailed product values are expected to sum to their associated broad product value. Additionally, a particular detailed product is associated with only one broad product. Missingness tends to be higher with detailed products than broad products.

It is not easy to develop viable imputation models for products. Auxiliary product data are not readily available. Moreover, other predictors such as total receipts are often weakly related. In most industries, the frequently reported products are highly correlated with total receipts and generally make up the majority of the total value of receipts, whereas the remaining products are not. Thus, the best predictors of an establishment's products are the industry assigned to the establishment from the sampling frame, which may change after collection, and the total receipts value (Ellis and Thompson 2015). Given the lack of predictors and the concerns about the consistency of the 2012 and 2017 data collections, the team considered four candidate imputation methods:

- Ratio imputation
- Sequential Regression Multivariate Imputation (SRMI) as described in Raghunathan et al (2001)
- Two variations of hot deck imputation (random and nearest neighbor), both which imputed the multivariate distribution of products from donor establishments.

The team decided on a simulation approach to create industry "populations" from historical sample data in 39 industries by applying each candidate imputation method to replace the missing data as suggested by Dr. Trivellore Raghunathan (University of Michigan). Product nonresponse was induced in 50 independent replicates in each completed population, and all four candidate methods were used to "complete" the datasets, using multiple imputation to obtain the imputed estimates, standard errors, and evaluation statistics (imputation error and fraction of missing information).

By design and necessity, the simulation study made some simplifying assumptions beyond those already mentioned. Small sample size effects were controlled by choice of estimate level (national) and the selection of study industries. These choices sidestepped issues that would arise from small respondent sample sizes in imputation cells. The evaluation was restricted to the two best-reported broad products in each studied industry in terms of number of establishments that reported the product. Rescaling the size of the problem reduced computation time and increased available time for analysis, although it did impact the study's "representativeness." Lastly, the evaluation used rank-based tests within industry to compare the procedures, so that substantive improvements or deficiencies in specific situations were largely ignored. The evaluation procedures found common patterns among the methods on each evaluation criterion on the *best-reported products* instead of using statistics for every product reported in an industry. The team recommended using hot deck imputation for broad products in all industries, allowing different hot deck variations by industries. This recommendation was endorsed by the project stakeholders. That said, the recommendation was

incomplete. No guidance was provided in terms of optimal imputation cells, minimum cell size (or collapsing rules), backup imputation methods (in the event of no donors), or dollar value and additivity requirements for final imputed data (no imputed values of less than \$500 were allowed, but all rounded values were required to add to the associated value of total receipts). Imputation of detailed products was not addressed in the research study, nor was calibration to industry totals.

3. Results

A new team was established for implementation. Leadership was provided by project management experts with extensive familiarity with the subject matter and with the planned EC processes. A large representation of (industry) subject matter experts were included, as were production programmers. Four methodologists from the research team were retained as consultants, with a production methodologist recruited to develop the final specifications. [Note: This methodologist directly supervised a research team member and was not unfamiliar with the research project, even without working directly on the team.] The team lead and one subject matter expert had participated on the research team, but the remaining team members were recruited from other EC projects and were not familiar with the earlier research or methods.

As with the research team, the project began slowly with an educational component. Team members each had their own set of implementation issues that needed to be addressed. The production programmers were concerned about the computing resource demands of fully implementing hot deck imputation; there were also staffing concerns as team members had to juggle working on this project with other high priority projects. Fortunately, the methodologists were able to provide concrete examples of efficient hot deck systems, which partially alleviated these concerns.

Presentation of the hot deck methods was generally met with acceptance. However, the subject matter experts balked at cell-collapsing procedures for two reasons. First, they were not convinced that imputation cells with sufficient observations should be combined with imputation cells with insufficient observations. Instead, they insisted on using the original imputation cells in the former case, reserving the collapsed imputation cells only for the latter case. Second, they did not have resources to research alternative cell collapsing procedures. Eventually, the implementation team compromised under strong protest from several methodologists, agreeing to the unconventional preferred cell-collapsing procedure but using a minimum cell size of one establishment. Although the specifications called for parameterized imputation cell definitions with three levels of collapsing, the imputation cell definitions that were coded into the production program varied little by industry.

The subject matter experts were emphatic on one point: they wanted to maximize the use of unit-level reported data in the imputation procedures whenever possible. However, businesses are more likely to report broad product categories than detailed categories. Restricting donor records to establishments that provided usable values for broad and detailed products was too restrictive for many industries and would likely lead to inefficient estimates. An inspection of 2012 EC counts of reported broad and detailed products within the most restrictive imputation cell definitions (industry by state by unit type) confirmed this suspicion. The majority of imputation cells contained at least five establishments that reported usable broad products; this was not the case with the detailed products.

Moreover, the research team had not studied imputation methods for detailed products. The lack of available reported data – and the differences in types of detailed products between 6-digit industries – was a prohibitive barrier. Methodologists on the team recommended using a ratio imputation model for each detailed product, known in-house as “category average” imputation. Although this method did not prove optimal with the referenced broad product research, the model is supported by the literature and almost always produced unbiased broad product estimates in the studied industries (Garcia, Morris, and Diamond 2015).

Table 2 presents a categorization of donor and recipient establishments based on the presence of usable broad and detailed products. Using 2012 EC data, we estimated the percentage of donors that would fall in each of these categories as follows: complete (86.6%), partial (7.4%), and minimal (6.0%). Obviously these percentages may be different in 2017, given the changes to the questionnaire.

Table 2: Establishment Classification for Imputation

Donors	Broad products usable
Complete	All broad and detailed products usable (contribute to category average)
Partial	All broad products usable and some detailed products usable (contribute to category average)
Minimal	All broad products usable; detailed products missing and required (receive detailed products from category average)
Recipients	Missing products
Full	Need broad and detailed products (receive all products from hot deck)
Partial	Need some (designated) detailed products (receive detailed products from category average)
Ineligible	All products usable, but not “typical”; excluded from donor pool

To simplify the operational procedure, the implementation team decided to create “complete” donor records (i.e. fill in the missing detailed products for partial and minimal donors) prior to hot deck imputation. To accomplish this, category averages were computed for each detailed product within a broad product for each potential imputation cell (generally industry-by-state-by-unit type, industry-by-state, industry) with a required minimum of one establishment in the cell reporting the detailed products. Designated missing detailed products were imputed from their associated broad product total by using the appropriate category averages. Once this process was finished and all donors were made “Complete,” the hot deck process was performed to impute products for all “Full” recipients. This approach of completing the partial and minimal donors maximized the use of reported data in the hot deck imputation procedures, but later complicated the variance estimation of detailed products due to the partial donor/recipient establishments.

The implementation team met regularly over a two-year period. During this collaborative period, methodologists met separately each week (along with the team leader) to develop the missing data procedures and treatments that were not addressed by the research team. Specifications were reviewed first by this subgroup, then by the entire team. Testing was a larger problem. Using small, single industry test decks, we were able to verify that the category average and hot-deck processes were working correctly. However, one of the concerns about hot-deck imputation was the time it would take to run the process to impute missing products for the entire EC. To determine estimated run-times, as well as test more scenarios, we created a full size test deck with roughly 2.4 million donors (with over 20 million products) and 1.1 million full recipients covering all NAICS sectors in-scope to the EC. Using a concordance that mapped 2012 product codes to 2017 NAPCS codes, we converted the 2012 EC product data to a 2017 NAPCS basis, again making simplifying assumptions, while ensuring that certain specific scenarios were included in the test data. The performance testing using this test deck took approximately 80 minutes. This was a reassuring result, although it might not directly translate to run times using actual 2017 EC production data and systems.

4. Discussion and Conclusion

When developing a research plan that applies to an ongoing survey, finding balance is hard. On one hand, making the scenario as simple as possible reduces the probability of treatment effects (solutions) being confounded by factors such as sample size or random noise. On the other hand, oversimplification can lead to very impractical solutions. Of course, it is crucial to limit the scope so that the research can be timely enough to be relevant when completed. However, it should be acknowledged that

compressing the scope can lead to hasty decisions later in the implementation process, when there is no time left for careful further investigation.

There are real advantages in establishing (almost) separate research and implementation teams as discussed in this paper. Having two teams approach the same problem from different perspectives leads to innovative applications. Often, these teams provide practical opportunities for methodologists to learn about data and data collection and for subject matter experts to learn about alternative methodologies. From an administrative perspective, these teams can help with succession management planning, especially when junior staff are included. Lastly, they provide justification for the production procedures under the umbrella of data-driven decision making.

Of course, there are equally real disadvantages. The limited scope in research can lead to missed requirements, which can be revealed as unexpected results in implementation testing or in production. Delaying decisions until implementation can preclude having sufficient time for careful investigation, and quick decisions are made for convenience based on anecdotal justification, with no alternatives tested. Having two separate teams increases management challenges as well, as appropriate leaders need to be recruited and team members struggle with competing duties (and on occasion, motivation and morale challenges).

When the end-product is a theoretically solid and operationally viable system, this approach is a success. It certainly was in the case study presented in this paper. The two-phase team approach has been used for other 2017 Economic Census applications such as determining and implementing a variance estimation method for product estimates (Thompson and Thompson 2018) and for developing standard response rates (Lineback, Oliver, and Willimack 2012). Certainly in these examples, the advantages outweighed the disadvantages, with workable solutions and buy-in as well as shared understanding of implemented methods. And of course, the imperfect solutions provide plenty of exciting research ideas and opportunities for the next Economic Census.

References

1. Ellis, Y. and Thompson, K.J. (2015). Exploratory Data Analysis of Economic Census Products: Methods and Results. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
2. Fink, E.B., Beck, J.L. and Willimack, D.K. (2015). Data-Driven Decision Making and the Design of Economic Census Data Collection Instruments. *Proceedings of the FCSM Research Conference*.
3. Garcia, M., Morris, D.S., and Diamond, L.K. (2015). Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on

- Economic Census Products. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
4. Lineback, F., Oliver, B., and Willimack, D.K. (2012). Developing Response Metrics for the Economic Census. *Proceedings of the FCSM Research Conference*
 5. Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*. 27(1), pp. 85–95.
 6. Thompson, K.J. and Liu, X. (2015). On Recommending a Single Imputation Method for Economic Census Products. *Proceedings of the Government Statistics Section*, American Statistical Association.
 7. Thompson, M. and Thompson, K.J. (2018). Variance Estimation for Product Sales in the 2017 Economic Census: Utilizing Multiple Imputation to Account for Sampling and Imputation Variance. *Proceedings of the Government Statistics Section*, American Statistical Association.



The need for granular data in evidence-based policies: The case of housing affordability in Malaysia



Suraya Ismail

Khazanah Research Institute, Kuala Lumpur, Malaysia

Abstract

This paper highlights the need to have more data collected at the right spatial scale to provide evidence-based policies for managing the rising unaffordability of the housing market. The importance of aggregation is not to be underestimated but the need to have more granular data is equally important. This is because cities demonstrate different trends of affordability due to the inherent structural differences of the housing markets and yet data is aggregated at the state level. Examples are given where data is available only at state and federal level for ownership rates and the calculation of affordability. Whilst this is useful, it is still not accurate.

Keywords

homeownership rate, household incomes, house prices, affordability ratio, evidence-based policies

1. Homeownership and informality in Malaysia

According to the latest available official figures in 2010, Malaysia has a home ownership rate of 72.5%. This is a relatively high number considering that home ownership rates in developed countries – apart from Singapore – were below 70% in the same year (UK-67.4 % and US 66.5 %). However, Malaysia's home ownership rates, which are published by the Malaysian Department of Statistics (DoS), also include ownership of informal houses. For instance, houses built by families at buffer zones of rivers are illegal but are still considered as owned homes in the Population and Housing Census. Formal housing stock is defined as housing which has been built with development orders from local authorities being issued. Conversely, informal housing stock are houses built without development orders and/or houses built by the community and may include 'kampung' houses.

There is a significant amount of housing stock that falls within the housing unit count in the 2010 Population and Housing Census that is not included in the estimates for housing stock published by the National Property and Information Centre (NAPIC), which only takes into account formal housing. In 2010, the former exceeded the latter by 2.9 million. This means there are effectively near 3.0 million informal housing and this might well be the reason why ownership rates are higher in rural areas. Informality in this case has

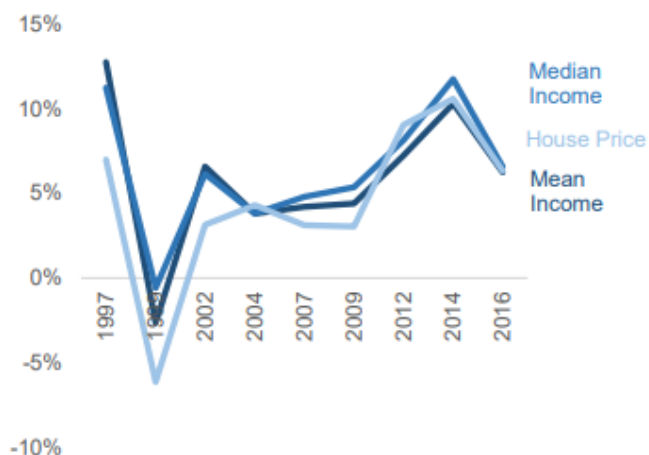
increased the ownership rates for Malaysia. If we take the example of urban Kuala Lumpur, where informal houses do not exist, the ownership rate is 53.5%.

Table 1: Percentage of home ownership in Malaysia, 2010

Percentage of home ownership			
	Total	Urban	Rural
Malaysia	72.5%	69.1%	81.2%
Johor	72.2%	69.6%	78.5%
Kedah	81.8%	77.2%	89.9%
Kelantan	80.5%	72.3%	86.4%
Melaka	72.9%	72.0%	78.5%
N. Sembilan	71.7%	68.2%	77.8%
Pahang	72.0%	65.8%	78.6%
Perak	75.7%	73.4%	80.8%
Perlis	78.1%	72.8%	83.5%
P. Pinang	77.5%	76.8%	83.5%
Sabah	68.1%	65.3%	71.7%
Sarawak	79.4%	75.9%	83.2%
Selangor	67.9%	66.9%	78.2%
Terengganu	78.5%	75.5%	82.8%
K. Lumpur	53.5%	53.5%	-
Labuan	57.0%	54.4%	73.5%
Putrajaya	9.9%	9.9%	-

2. Have household incomes have grown in tandem with house prices?

Figure 1 shows that household incomes have generally moved in tandem with house prices since 1997. It therefore suggests that there should not be a problem with housing affordability. However, Table 2 demonstrate the limitation of this claim when house prices are divided into the different house products specific to the states. For example, in Kuala Lumpur an average terrace house price have increased 3 times from 2000 to 2016. Household incomes however have not increased as much. Granularity of data would depict the problems of affordability more clearly.

Figure 1: Growth in household incomes and house prices, 1997-2016

Compound annual growth rate, 1997 – 2016

Year	1997	1999	2002	2004	2007	2009	2012	2014	2016
Mean Income	12.7%	-2.6%	6.6%	3.8%	4.2%	4.4%	7.2%	10.3%	6.2%
Median Income	11.2%	-0.6%	6.1%	3.8%	4.8%	5.4%	8.1%	11.7%	6.6%
House Price	7.0%	-6.1%	3.1%	4.3%	3.1%	3.1%	9.0%	10.6%	6.3%

Table 2: Breakdown of house prices by state, 2016

Area	2016 house price as a multiple of 2000 house price				
	All House	Terrace	Semi-Detach	Detach	High-Rise
MALAYSIA	2.42x	2.36x	2.42x	2.58x	2.54x
Sabah	3.19x	3.30x	3.32x	2.85x	2.76x
K. Lumpur	2.93x	3.00x	2.92x	3.26x	2.54x
Terengganu	2.81x	2.74x	2.93x	2.84x	
Pahang	2.76x	2.66x	2.88x	3.33x	
P. Pinang	2.72x	2.90x	1.94x	1.84x	3.00x
Perlis	2.54x	2.50x	2.64x		
Kelantan	2.02x	2.17x	2.30x	1.51x	
Perak	2.48x	2.43x	1.86x	2.82x	
Sarawak	2.40x	2.23x	2.50x	2.57x	
Selangor	2.40x	2.24x	2.36x	2.31x	2.01x
Kedah	2.28x	2.17x	2.47x	2.44x	
N. Sembilan	2.23x	2.27x	1.94x	2.09x	1.23x
Melaka	2.07x	2.21x	2.53x	1.19x	1.38x
Johor	1.81x	1.76x	2.03x	1.84x	2.05x

CAGR 2000 - 2016

Area	All House	Terrace	Semi-Detach	Detach	High-Rise
MALAYSIA	6.1%	5.9%	6.1%	6.5%	6.4%
Sabah	8.0%	8.3%	8.3%	7.2%	7.0%
K. Lumpur	7.4%	7.6%	7.4%	8.2%	6.4%
Terengganu	7.1%	6.9%	7.4%	7.2%	
Pahang	7.0%	6.7%	7.3%	8.4%	
P. Pinang	6.9%	7.4%	4.5%	4.2%	7.6%
Perlis	6.4%	6.3%	6.7%		
Kelantan	4.8%	5.3%	5.7%	2.8%	
Perak	6.2%	6.1%	4.2%	7.1%	
Sarawak	6.0%	5.5%	6.3%	6.5%	
Selangor	6.0%	5.5%	5.9%	5.7%	4.8%
Kedah	5.6%	5.3%	6.2%	6.1%	
N. Sembilan	5.5%	5.6%	4.5%	5.0%	1.4%
Melaka	5.0%	5.4%	6.4%	1.2%	2.2%
Johor	4.0%	3.9%	4.8%	4.2%	4.9%

Min  Max

The similar example is for the affordability threshold for states. At the state level, limitations in the data collection process inhibit the analysis of housing affordability for all states in Malaysia. Although DOS' household income statistics measure the household income of those living in both informal and formal housing, NAPIC's data on housing stock excludes a significant amount of housing units that are informal. Given that NAPIC only collects data on formal housing, the housing affordability analysis is limited only to states where these types of houses represent a significant proportion of the state housing stock (60% or more) i.e. Kuala Lumpur, Pulau Pinang, Johor, Selangor, Negeri Sembilan and Melaka.

Table 3: Median multiple affordability by state, 2002 – 2016

State	2002	2004	2007	2009	2012	2014	2016	Affordability category	% formal housing
Kelantan	5.1	5.4	4.4	4.5	6.2	7.1	5.5		16
Sabah	6.3	6.7	10.0	6.2	5.8	5.6	5.5	5.1 & over	24
Pulau Pinang	4.1	4.3	4.1	4.0	4.1	5.8	5.5	Severely unaffordable	74
Negeri Sembilan	3.4	3.1	3.3	3.4	2.8	5.0	5.1		74
Pahang	5.0	4.2	3.7	3.9	3.8	5.3	5.0		58
Johor	4.9	4.9	3.5	3.7	3.7	4.3	5.0		73
MALAYSIA	4.1	4.3	4.4	4.4	4.0	5.1	5.0	4.1 to 5.0	60
Terengganu	4.7	4.8	5.0	5.2	5.3	6.2	5.0	Seriously unaffordable	22
Kuala Lumpur	4.7	5.4	5.0	4.6	4.9	5.6	4.9		88
Selangor	3.7	3.5	3.6	3.6	3.6	5.2	4.7		81
Perak	3.9	4.1	3.5	3.5	3.3	5.1	4.6		57
Kedah	4.6	4.1	4.1	4.0	3.6	3.4	4.3		50
Sarawak	n.a.	n.a.	3.7	4.1	4.0	4.2	4.0	3.1 to 4.0	32
Perlis	4.4	3.7	3.6	4.5	4.3	4.5	4.0	Moderately unaffordable	34
Melaka	3.4	3.5	2.9	2.9	2.6	3.1	3.1		64

Source: NAPIC (2017a), DOS (various years.a), DOS (various years.b) and KRI calculations

Note: The states highlighted in orange are those where 60% or more of housing stock is accounted for by NAPIC.

Even then, these calculations are based on state-wide numbers and do not reflect the relative scarcities of housing units in different housing markets of cities and towns. For example, a housing market in Kuala Muda would be vastly different than Alor Star, and yet official data is collected to reflect a state level median house price that includes both housing markets. Indeed, some of the supply mismatches of house prices to income (and therefore reducing housing affordability and creating a supply glut) have occurred since housing development projects treat these markets as similar, ostensibly with the intention of increasing the carrying capacity of impending growth of cities and towns. Therefore, data on household income levels and house prices must be collected at the right scale of existing housing markets. The projections of the future growth of towns and cities could be estimated from the strategic directions laid down in State's Structure plan and District local plan. In lieu of this, the spread of housing markets that goes beyond state boundaries must be acknowledged. Take for instance the case of Negeri Sembilan; the median multiple has remained moderately affordable during the period of 2002-2016 but with the advent of more houses needed for the Kuala Lumpur/ Selangor

housing market, the state median multiple has crossed over the severely unaffordable threshold. Melaka, on the other hand, has behaved as a state-contained housing market.

While the calculated median multiples of the remaining states are shown in Table 3, it should be noted that the figures are not reflective of the current affordability levels of the housing markets in those states. For states that are close to the 60% threshold, for example Pahang and Perak, upcoming housing developments, given increased urbanisation in the states, may increase the proportion of more formal housing in the state, making the calculated median multiple more reflective of the housing market in the future.

Between 2002 and 2016, housing affordability has worsened for all states. For most states, the deterioration in housing affordability occurred most significantly between 2012 and 2014. In critical states like Negeri Sembilan and Johor, the median house price increased at a CAGR of 36.5% and 26.2% respectively over the stated period. By comparison, median incomes in those states grew considerably more slowly at 7.2% and 17.7%, respectively, over the same period.

In 2016, housing affordability improved slightly for Pulau Pinang, Kuala Lumpur and Selangor, which saw a slight reduction in their calculated median multiple affordability. Nevertheless, their median multiple affordability of 5.5, 4.9 and 4.7 respectively still render housing markets in these states “severely unaffordable” and “seriously unaffordable”.

The lack of housing affordability in these states have stemmed partly from the unresponsiveness of housing supply to effective demand¹. In 2014, the calculated market median-3 house price—the price of an “affordable” house—for Malaysia was RM165,060. In 2016, this figure was RM188,208. As shown in Figure 4, newly launched housing units that are priced below RM200,000 made up less than 20% of the total units launched in 2014 – 2016.

Table 4: Comparison of median house price against the corresponding market median-3 price for Malaysia, 2002 – 2016

Year	Median house price (a)	Market median-3 house price (b)	Difference (a)-(b)
2002	100,000	73,764	26,236
2004	115,001	79,596	35,405
2007	135,000	91,872	37,128
2009	149,000	102,276	46,724
2012	175,000	130,536	44,464
2014	280,000	165,060	114,940
2016	313,000	188,208	124,792

Source: NAPIC (2017a), DOS (various years.a) and KRI calculations

¹ KRI (2015)

At a more granular level, similar observations can be seen at the state level. New launches of property developments in states like Kuala Lumpur and Pulau Pinang in 2016 have generally skewed towards higher priced properties, even as income dynamics in the state suggest that these upcoming units would be unaffordable for the population. Most strikingly, in Pulau Pinang, all properties launched in 2016 were priced above RM250,000, with the bulk of the newly launched properties situated in the RM500,000 to RM1,000,000 price bracket. To put this in perspective, the calculated market median-3 house price—the price of an “affordable” house—for Pulau Pinang in 2016 was RM194,724.

3. Conclusions

This paper has used existing data available to explain housing affordability trends in Malaysia. However, more could be done to collect data at the right spatial scale in order to devise more specific evidence-based policies for the country.

References

References for this paper can be found in “Making Housing Report”, at Khazanah Research Institute’s website: www.krinstitute.org

Index

A

Abd Aziz Latip, 118
Alessandro Fassò, 217
Alphonse L. MacDonald, 300
Arturo Márquez-Cerda, 368
Atilla Karaman, 153

B

Barry Schouten, 272
Belkacem Abdous, 162

C

C. Crea, 208
Charlie Russell, 377
Chibuzor Christopher Nnanatu, 231
Choo Kit Hoong, 346

D

Danielle Jade Roberts, 239
Dennis Matanda, 231
Dickson Lukose, 118

E

Etienne Saint-Pierre, 263

F

Fabio Madonnaz, 217
Fabrizio Maturo, 200
Fauzana Ismail, 351
Firdaus Abhar Ali, 194
Francesca Fortuna, 200
François Laflamme, 263
Fred Barzyk, 263
Frederick Kin Hing Phoa, 1
Fuziah Md Amin, 351

G

Gan Chew Peng, 48
Geoff Bowlby, 263
Ger Snijkers, 272
Gigih Fitrianto, 137
Glory Atilola, 222, 231
Guangquan Li, 222

H

Hasnah Mat, 316, 332
Hideatsu Tsukahara, 124
Holly Mullin, 263
Hsin-Cheng Huangy, 217

I

Igor Valli, 217
Irene Salemin, 253
Ismat Mohd Sulaiman, 118

J

Jeremy Tan, 346
Joe Winton, 377
Joerg Beutel, 11

Joseph Mariasingham, 39
Jukka Hoffrén, 73
Junji Nakano, 1

K

Karin Hansson, 81
Katherine Jenny Thompson, 386
Keisuke Honda, 1

L

L. Leticia Ramírez-Ramírez, 368
Li Jiajing, 360
Lubanzadio Mavatikua, 231

M

Maciej Truszczynski, 64
Maria Denise M. Peña, 324
Mario Palma Rojo, 284
Mazreha Ya'akub, 316, 332
Md. Khadzir Sheikh Ahmad, 118
Mohd Azman Mohd Ismail, 194
Mohd Syazrin Mohd Sakri, 118
Muhammad Aiman Mazlan, 118

N

Nabihah Jasri, 56
Nancy McBeth, 185
Nebil Dabur, 153
Neo Soo Khee, 346
Ng Kok Haur, 48
Ngianga-Bakwin Kandala, 222, 231
Nobuhiro Okamoto, 30
Noraliza Mohamad Ali, 88, 99
Norshahida Shaadan, 56
Norul Anisa Abu Safran, 316, 332
Nur Hurriyatul Huda Abdullah San, 88
Nur Layali Mohd Ali Khan, 88, 99
Nurfarahin Harun, 88

O

Ola Awad, 177
Omar Ismail, 118

P

Paul Komba, 231
Pilar Martin-Guzman, 309
Pooi Ah Hin, 48

R

R. Ayesha Ali, 208
Riaan de Jongh, 248
Rocío Maribel Ávila-Ayala, 368
Roeland Beerten, 293
Ryuei Nishii, 137

S

Sabir Al Harbi, 185
Saidi Hedi, 168

Index

Samuel Manda, 222
Sara Frankl, 81
Sasongko Yudho, 110
Sevgui Erman, 263
Shafiq Naim Shahrudin, 194
Shojiro Tanaka, 137
Siti Asiah Ahmad, 88
Sofie de Broe, 272
Stéphane Dufour, 263
Suraya Ismail, 396
Syirahaniza Mohd Salleh, 118
Sylvie Bonhomme, 263

T

Takaki S., 129
Temesgen Zewotir, 239
Tonio Di Battista, 200

V

Vaijyenthi Gurdayal Singh, 194
Vince Galvin, 377

W

Wan Mohd Haffiz Mohd Nasir, 351
Wei-Yin Loh, 340
William C. Davie Jr., 386

Y

Yang Cuihong, 21
Yasumasa Matsuda, 129
Yoshihiro Yajima, 146
Yuji Mizukami, 1

Z

Zhang Junrong, 21
Zhuzhi Moore, 231



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



ISBN 978-967-2000-64-8



9 789672 000648

#ISIWSC2019