62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur
18 - 23 August 2019

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 6

62nd ISI WORLD STATISTICS CONGRESS 2019

18 - 23 August 2019, Kuala Lumpur

Come | Connect | Create

www.isi2019.org

PROCEEDING

ISI WORLD STATISTICS CONGRESS 2019

CONTRIBUTED PAPER SESSION (VOLUME 6)

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

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

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

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

Published in February 2020

Suggested citation:
Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Contributed Paper Session: Volume 6, 2019. 428 pages

Preface

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

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

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

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

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

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

Thank you.

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

Scientific Programme Committee of the 62nd ISI WSC 2019

Chair
Yves-Laurent Grize, Switzerland

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

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

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

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

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

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

Local Programme Committee of the 62nd ISI WSC 2019

Chairperson
Rozita Talha

Co-Chairperson
Prof. Dr. Ibrahim Mohamed

Vice Chairperson
Siti Haslinda Mohd Din
Daniel Chin Shen Li

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

TABLE OF CONTENTS

Contributed Paper Session (CPS): Volume 6

Using machine learning technologies for coding economic activities of Businesses– The NOGAuto System

Claude Macchi, Michel Chételat, Cindia Duc-Sfez, Christophe Joyon
Swiss Federal Statistical Office, Neuchâtel, Switzerland

Abstract

Classifications are basic elements for the production of statistics. The quality of the coding of the observed units has a direct impact on the entire data production process, on the credibility and on the quality of the statistical outcome. This is even more important in the context of register and administrative data, which are the starting point of countless statistics, the base of sample frames and of data analysis.

With a view to a continuously improving the quality of the coding of units in the Swiss statistical business register (SBER) as well as to decreasing the burden of businesses in their obligations to deliver information to the statistical offices, the Swiss Federal Statistical Office (FSO) launched early 2018 a project to automatise the attribution of the economy activity code to businesses. This project is one of the five projects currently being developed in line with the FSO's data innovation strategy (https://www.bfs.admin.ch/bfs/en/home/news/whats-new.assetdetail.3862240.html) with the goal to argument and/or complement the existing basic official statistical production at the FSO.

The coding procedures are currently quite standardised. The encoders analyse and interpret information on the businesses activities such as inputs from the businesses themselves, from surveys as well as descriptions in company registers and different administrative data. Based on this they define keywords that are compared with a list of keywords- and concepts, linked to each of the positions of the classification and their related explanatory notes, and select on this way the code to be attributed to the observed business.

Using innovative new ways, the FSO is building a machine learning system to automatise the manual coding procedure. This artificial intelligence undertakes the reading and interpretation steps from the coder and automatically associates the business to a classification code. In addition, it proposes new keywords and concepts. It is actually learning from an existing dataset that has already been tested manually. In a sequent step, the horizon of the system could be enlarged by looking directly on the web for additional sources of information on the business to be coded.

Keywords

Machine Learning; Coding; Classifications; Key Words; Artificial Intelligence

1. Introduction

Each company and establishment stored in the SBER has a Swiss Economic Activity Classification (NOGA) code, which is based on the Classification of Economic Activities in the European Community (NACE). The codification is carried out in two main steps: the first, with the assignment of a provisional code when the business is integrated into the registers, and the second, which takes place a few months later, with the validation of the first code assigned.

During the first codification, the FSO uses the economic activity code assigned by the source itself, which provides SBER with information on the new company. Depending on the data source, this coding may have been done either by the source itself (mainly in the case of administrative data or registers external to the FSO), or by third party companies (in the case of announcements from commercial registers). The first code is then validated by the FSO as part of a specific survey of all new companies registered in the SBER. Based on the descriptions of economic activities provided by the companies themselves, coders identify terms or concepts deemed relevant that are compared to a list of keywords, currently containing more than 11'000 items and concepts in four different languages (German, French, Italian and English) and related to each of the NOGA positions. This allows coders to select a code and assign it to the observed company. After this phase, the codes assigned may be updated or corrected at any time, on the basis of inputs from surveys carried out in the context of statistical production, administrative sources, external registers, the companies themselves and information obtained on the Internet. The codes defining the economic activity of companies are all assigned based on oral or written information provided by the companies themselves. Codification therefore consists mainly of reading, understanding and interpreting a text, followed by the definition of terms or concepts that are compared with a list of keywords linked to the classification codes.

With the NOGAuto system, the FSO aims to build a machine learning system with the aim of automating the assignment of economic activity codes to SBER companies. This will make it possible to

- reduce to a minimum the interpretation made by coders of texts describing the economic activities of companies,
- harmonise and standardise the assignment of codes and
- minimise the time spent on the coding activity.

NOGAuto is not built in one go, but, like an onion, in different layers. The central nucleus of the onion, the first phase of construction, makes it possible to validate the codes currently associated with the units already registered in the SBER. The following layer will assist coders with code proposals for the activities of companies to be codified. The interaction of coders, who will accept or reject the codes proposed by the system, as well as the continuous

input of new keywords proposed thanks to artificial intelligence, will allow the system, with the construction of additional layers, to continuously improve the quality of the proposed codes and to reach a stage where the codes can, in most cases, be assigned automatically, without any human intervention.

2. Methodology

The process workflow

Figure 1 below shows the processes of the system, the purpose of which is to support and facilitate the coding of business activity at the FSO.

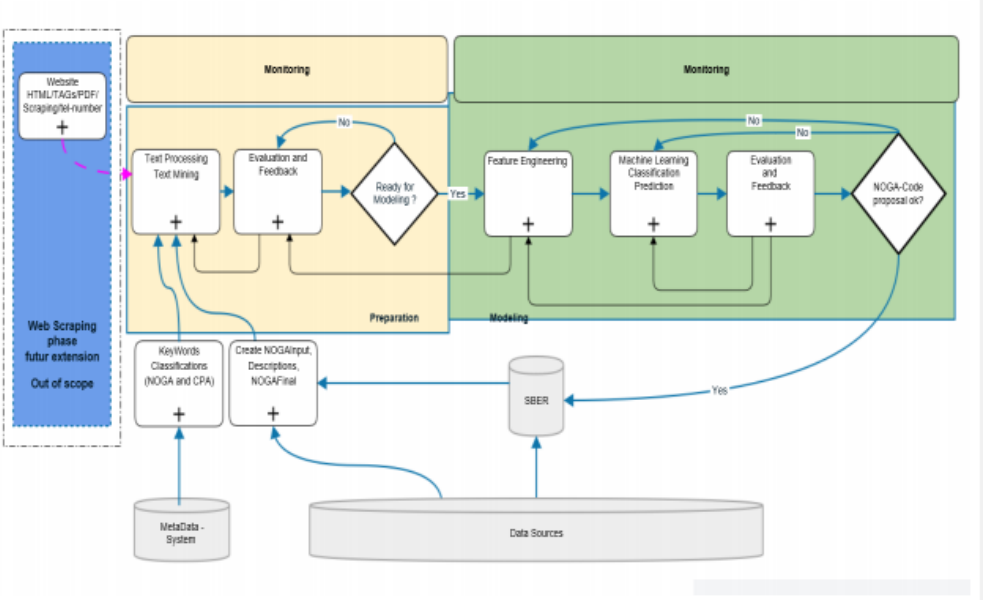


Figure 1: The NOGAuto process workflow

The « Preparation » phase

Texts describing the economic activities of companies registered in the SBER, those from electronic surveys and metadata from the FSO's central metadata system are imported into the NOGAuto system and subject to Natural Language Processing (NLP) operations. In this context, the texts are first of all subjected to a language detection process, followed by a stemming operation, which allows a normalisation of the elements of the texts (verbs, disinences, etc.) as well as a cleaning, with the elimination of stop words which cannot be directly related to an economic activity (articles, acronyms, prepositions, etc.).

The output of the NLP operations is then analysed and the various information cross-referenced. Particular attention was paid to the complex classification positions that are more difficult to assign and where more coding errors are committed.

Figure 2 shows, for example, an extract from the analysis of NOGA positions 45 "Wholesale and retail trade and repair of motor vehicles and

motorcycles", 46 "Wholesale trade, except of motor vehicles and motorcycles" and 47 "Retail trade, except of motor vehicles and motorcycles". We can thus see how the key words "workshop", "car body" and "automobile", which should only identify position 45, have also been linked with companies codified in positions 46 and 47 of the classification.

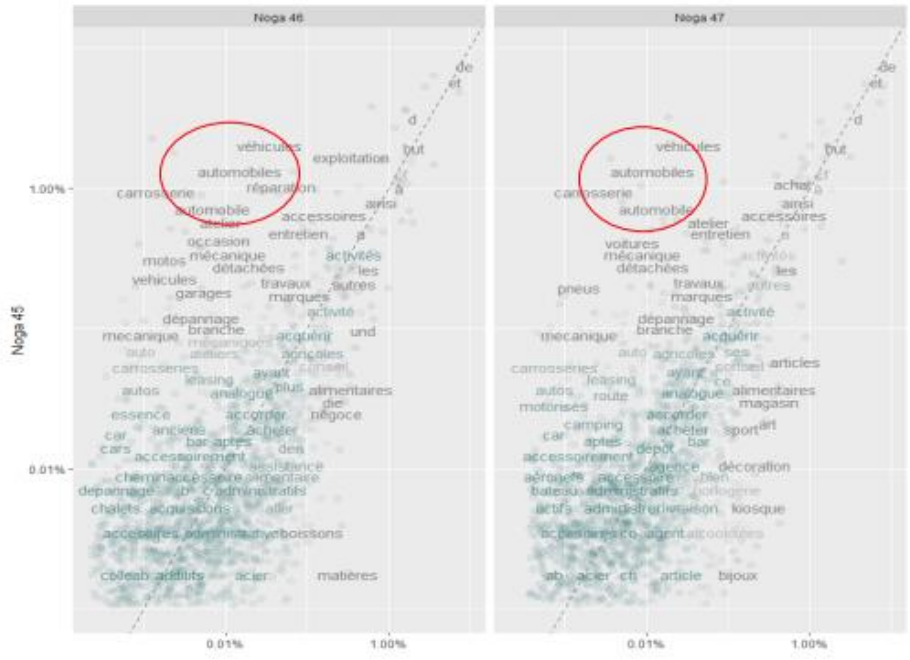


Figure 2: Cross-reference keywords NOGA 45, 46 and 47

These crosses also make it possible to discover elements that have not been completely cleaned during the NLP operations (words in other languages, isolated letters, etc.). This is the trigger for a new language detection and stemming operation phase to correct errors in keywords. And this loop will be repeated systematically until the keywords are of a quality considered good enough to launch the next phase of the process.

This continuous feedback to the previous phase of the process is a key element of NOGAuto's philosophy, which is based on a correction and continuous improvement of the outcomes of the previous steps.

The « Modelling » phase

The companies to be codified are linked in a single matrix to the (1) keywords resulting from the description of the economic activity produced in the "Preparation" phase as well as to (2) variables (address, jobs, turnover, legal form, etc.) from the SBER and which are intended to provide additional input for the definition of the NOGA code to be assigned. From this matrix a prediction model will be generated that defines the keywords and concepts to

be linked with the different activity codes to be assigned to companies. This model will be continuously enriched with elements from the descriptions of the activities of the new companies to be codified or additional information from the SBER. Once again here as well, feedback to the previous phase of the process is essential, so that continuous correction and improvement of the model can be achieved.

The last step in this phase is the evaluation of the codes that the system proposes to the coders, who validate them, integrate them into the SBER or reject them, which will generate feedback at earlier stages of the process, until the defined success criteria are met.

3. Result

The NOGAuto project was launched in early 2018 and is still under construction. After having built and tested the "Preparation" phase, we are currently building the "Modelling" part. Full process testing, based on an existing dataset that already has been tested manually, including the evaluation of the codes at aggregate NOGA 2-digit level that the system proposes, is scheduled for spring 2019. Its productive implementation is planned in stages. In a first period, and until the quality level defined for the most detailed NOGA code level is reached – expected by mid-2020 – the system will only be used as a support tool for coders.

NOGAuto is not only a tool that can be used to codify the economic activities of businesses, but can also be adapted to the needs of other classifications. The more structured, standardised and targeted the information to be codified, the easier it is to propose an automatic codification. Initial discussions for an adaptation and an implementation of the system in the context of the classifications of occupations and of diseases and health problems have already been launched.

A central point that has accompanied this machine learning project from the beginning is the question of acceptance. The word "automation" has quickly been linked to "work reduction" and "loss of job", which caused quite a lot of opposition to the project, principally among the staff responsible for coding. Especially at the beginning of the work the cooperation with the people who were supposed to give the initial inputs on the codification processes was quite complex. An exercise of communication, explanation and clarification was necessary to gain the trust and collaboration of the staff. For the ISI 2019 conference, it is planned to present the complete system as well as the results of the tests performed with data at NOGA 2-digit level.

4. Discussion and Conclusion

Thanks to machine learning, the NOGAuto system will take over the reading and interpretation tasks of coders, propose new keywords and

concepts and associate companies with an economic activity code. This allows a reduction of the interpretation of texts describing the activities of businesses and facilitate the attribution of codes. The system is built in stages, like an onion, layer after layer. In the first step, companies are coded only at the aggregate level of 2-digits of the NOGA classification. A codification at the most detailed level of 6-digit will be undertaken at a later stage, once the system has made its first experiments and has reached a sufficient level of stability and quality.

NOGAuto is based on the principle of feedback, correction and continuous improvement. Each action can be challenged and thus trigger the relaunch of one or more steps in the process. This approach is fundamental not only for the construction phase, but must also be continued once the system will be in production.

In machine learning projects, communication is fundamental from the beginning. Any change often causes resistance, and especially the word "automation" is easily linked by the staff to "loss of jobs". It is therefore essential to integrate future users of the system into the project from the beginning and involve them in the development. "Automation" should in no way be seen as a reduction of tasks, but as a chance to be able to use the time saved for new tasks. The NOGAuto system is not limited to the codification of economic activities of enterprises, but can be adapted and used in the context of codification based on classifications in any other field of statistics.

Estimators of goodness-of-fit measures for a Poisson regression model

Takeshi Kurosawa[1], Kousuke Shinmura[2], Francis K. C. Hui[3], A. H. Welsh[3], Nobuoki Eshima[4]

[1]Department of Applied Mathematics Science, Faculty of Science, Tokyo University of Science, Tokyo, Japan

[2]Department of Applied Mathematics Science, Graduate School of Science, Tokyo University of Science, Tokyo, Japan

[3]Mathematical Sciences Institute, The Australian National University, Acton ACT, 2601, Australia

[4]Center for Educational Outreach and Admissions, Kyoto University, Kyoto, Japan

Abstract

In this study, we discuss a measure of predictive power m_{pp} which is one of the goodness-of-fit measures for generalized linear models (GLMs) proposed by Eshima and Tabata (2007). This measure expresses average amount of decreasing uncertainness of a response variable Y by a vector $\boldsymbol{X}$ of regressors. We apply it to a Poisson regression model with a random vector $\boldsymbol{X}$ of the regressors. Moreover, we propose an estimator of m_{pp} and compare it with other estimators.

Keyword

coefficient of determination; entropy; generalized linear models; measure of predictive power; correlation coefficient

1. Introduction

In regression analysis, it is often desirable to numerically summarize the overall fitted model through a goodness-of-fit measure. Perhaps the most well known of these is the Akaike Information Criterion (AIC). Being a relative measure, the actual values of AIC do not have a clear interpretation, and instead it is differences in AIC values between candidate models which drive its usage. This is in contrast to the multiple correlation coefficient R in the linear model, whose value can be interpreted explicitly as a ratio between the conditional variance of the fitted values $\hat{Y}$ based on the candidate model, and the overall variance of Y. This article focuses on one particular goodness-of-fit measure, based on the covariance between the response and the canonical parameter in a exponential family of distributions.

Approaching the problem of goodness-of-fit measures for GLMs, Zheng and Agresti (2000) proposed the regression correlation coefficient (RCC) which is the correlation between the response variable Y and the its conditional

expectation $E(Y|\boldsymbol{X})$. The RCC lies between 0 and 1, and similar to the multiple correlation coefficient R judges a candidate model as performing if it is close to one. Recently, Takahashi and Kurosawa (2016) studied the performance of the RCC in Poisson GLMs and derived explicit forms for it under the condition that the vector of explanatory variables $\boldsymbol{X}$ has a certain distribution.

Inspired by the work on RCC, Eshima (2004) and Eshima and Tabata (2007) proposed an alternate entropy correlation coefficient (ECC) measure for GLMs, based on the Kullback-Leibler divergence (which is equivalent to the symmetric Kullback-Leibler distance) between the marginal distribution of the response variable Y and the conditional distribution $f(Y|\boldsymbol{X})$. They showed that the form of ECC reduces to simply calculating the correlation between the response variable Y and the canonical parameter in the exponential family. Like the RCC and R, the ECC varies between 0 and 1. In this article, we study the *unstandardized version* of the ECC for Poisson GLMs. We refer to this as the measure of predictive power or m_{pp}, and it is defined as the covariance (as opposed to the correlation) between Y and the canonical parameter in the exponential family.

The article is structured as follows. In Section 2, we provide formal definitions of m_{pp}. We propose a new estimator of m_{pp} in the next section. Finally, we conduct a real data analysis using the proposed estimator of m_{pp} in Section 4.

2. Section Measure of Predictive Power m_{pp}

We introduce the form of m_{pp} for GLMs more generally, before focusing on the case of Poisson regression. In a GLM, conditional on a vector of explanatory variables $\boldsymbol{X}$, the responses are assumed to be independent observations from the exponential family of distributions. That is, $f(y|\boldsymbol{X}) = \exp\{a(\phi)^{-1}\{y\,\theta - b(\theta)\} + c(y,\phi)\}$ for known functions $a(\cdot), b(\cdot)$, and $c(\cdot)$, where θ is the canonical parameter and ϕ is a scale parameter which may be known or require estimation. The conditional mean, $E(Y|\boldsymbol{X}) = b'(\theta)$, is then modeled as $g\{E(Y|\boldsymbol{X})\} = \eta = \alpha + \beta^{\mathrm{T}}X$ for some specified link function $g(\cdot)$, where η is the linear predictor, α is the intercept, and β is the vector of regression coefficients. We focus on the following goodness-of-fit measure.

Definition 2.1. *The measure of predictive power for a GLM, denoted as* m_{pp}*, is defined to be the covariance between the response* Y *and the canonical parameter* θ

$$m_{\mathrm{pp}}(\alpha,\beta) = \frac{cov(\theta, Y)}{a(\phi)},$$

where the value θ is determined from the mean model i.e., $b'(\theta) = g^{-1}(\eta)$.

One particularly appealing characteristic of m_{pp} is that it can be derived in terms of in terms of Kullback-Leibler distances.

Theorem 2.1. *(Theorem 2.1, Eshima and Tabata, 2007) The measure of predictive power* m_{pp} *can be expressed as*

$$m_{\mathrm{pp}}(\alpha,\beta)=\int KLD\{f(y),f(y|\boldsymbol{x})\}f(x)dx=E[KLD\{f(y),f(y|\boldsymbol{X})\}],$$

where $f(y)=\int f(y|\boldsymbol{x})\}f(x)dx$ *is the marginal distribution of the response and*

$$KLD\{f(y),f(y|x)\}=\int f(y|x)\log\left(\frac{f(y|x)}{f(y)}\right)dy+\int f(y)\log\left(\frac{f(y)}{f(y|x)}\right)dy, \quad (1)$$

is referred to as Kullback-Leibler divergence in this article.

Note that $KLD\{f(y),f(y|\boldsymbol{x})\}$ is also often referred as the symmetric Kullback-Leibler distance between $f(y)$ and $f(y|\boldsymbol{x})$, and differs from the more common (asymmetric) Kullback-Leibler distance which comprises only the first term in (1).

The m_{pp} is a relative measure is the sense that $m_{\mathrm{pp}}(\alpha,\beta)\geq 0$ (as proved by Eshima and Tabata, 2007), but it is not bounded. On the other hand, by standardizing the formula in Definition (2.1) we can obtain the entropy correlation coefficient (ECC, Eshima and Tabata, 2007). Unlike $m_{\mathrm{pp}}(\alpha,\beta)$, the ECC is an absolute measure in the sense that $0\leq ECC\leq 1$. As an illustration of this, if we consider a linear model with $E(Y|\boldsymbol{X})=\alpha+\beta^{\top}\boldsymbol{X}=\theta$ and $a(\phi)=\sigma^2=var(Y|\boldsymbol{X})$, then it can shown that (Eshima and Tabata, 2007, see Example 1)

$$m_{\mathrm{pp}}(\alpha,\beta)=\frac{\mathrm{var}\,(E[Y|\boldsymbol{X}])}{E[\mathrm{var}(Y|\boldsymbol{X})]}=\frac{R^2}{1-R^2},$$

where R is the multiple correlation coefficient.

3. Estimation of m_{pp} in Poisson GLMs

We now focus on the specific case of Poisson GLMs with log link function:

$$\log\{E(Y|\mathbf{X})\}=\alpha+\beta^{\top}\boldsymbol{X},\ \ Y|\boldsymbol{X}\sim P\{\exp(\alpha+\beta^{\top}\boldsymbol{X})\}, \quad (2)$$

where $P(\lambda)$ denotes the Poisson distribution with parameter $\lambda>0$. Note that a consequence of using the log link is that $\theta=\alpha+\beta^{\top}\boldsymbol{X}$.

We can simply consider an estimator of m_{pp}. For a dataset comprising N observations $\{(\boldsymbol{X}_i,y_i);i=1,\dots,N\}$ let $\boldsymbol{Y}=(Y_1,\dots,Y_N)^{\top}$ denote the vector of responses and $\hat{Y}=(1/N)\sum_{i=1}^{N}Y_i$ denote the sample mean. Also, suppose a Poisson GLM as in (2) is fitted to the data, yielding maximum likelihood estimates $(\hat{\alpha},\hat{\beta}^T)^T$. Then let $\hat{\theta}_i=\hat{\alpha}+\hat{\beta}^{\top}\boldsymbol{X}_i$ denote the estimated canonical parameter, $\hat{\theta}=(1/N)\sum_{i=1}^{N}\hat{\theta}_i$ its sample mean, and $\hat{\theta}=\left(\hat{\theta}_1,\dots,\hat{\theta}_N\right)^{\top}$. The unbiased covariance estimator is

$$U(\widehat{\boldsymbol{\theta}}, \boldsymbol{Y}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_i - \hat{\theta})(Y_i - \bar{Y}) = \frac{N}{N-1}\mathrm{Cov}(\widehat{\boldsymbol{\theta}}, \boldsymbol{Y}).$$

We consider an another estimator, assuming that $\boldsymbol{X}$ is a vector of random variables with distribution characterized by the vector of parameters ψ. By Definition 2.1, integrating with respective to the distribution of the covariates, we can derive an explicit form of $m_{\mathrm{pp}}(\alpha, \beta)$ for Poisson GLMs. Substituting estimates of α, β, ψ into the explicit form of m_{pp} , we get the estimator

$$m_{\mathrm{pp}}(\hat{\alpha}, \widehat{\beta} \,|\hat{\psi}).$$

The notation makes explicit the dependence of the measure of predictive power on parameters $\hat{\psi}$.

4. Application to Horseshoe Crab Data

This section applies m_{pp} to the horseshoe crab data provided in Agresti (2002). Briefly, the dataset consists of 173 female crabs, with the response variable being the number of male crabs satelliting with each female crab Sa. There are also four explanatory variables: 1) weight of a female crab (Wt); 2) the carapace width a female crab (W); 3) the body color of a female crab (C); 4) the spine condition of a female crab (S). Both weight Wt and carapace width W are continuous variables, and a test of normality applied to both predictors suggested no strong evidence that either deviated substantially from a normal distribution (Takahashi and Kurosawa, 2016). Body color is a factor variable with levels C = 1: light medium, 2: medium, 3: dark medium, and 4: dark. We converted body colour into a binary predictor C_2, such that $C_2 = 1$ if C = 4 and $C_2 = 0$ otherwise. Analogously, the spine condition is a factor with levels S = 1: both good, 2: one worn, 3: both worn. We also converted this to a binary categorical variable S_2 such that $S_2 = 1$ if S = 1 and $S_2 = 0$ if otherwise.

Assuming a Poisson distribution for the count response Sa, we fitted 15 candidate models involving different subsets of the four covariates included as main effects, and Table 1: Values of $m_{\mathrm{pp}}(\hat{\alpha}, \widehat{\beta} \,|\hat{\psi})$ for 15 candidate Poisson regression models fitted to the horseshoe crab dataset. There were four explanatory variables: 1) weight of a female crab (continuous variable; Wt); 2) the carapace width a female crab (continuous variable; W); 3) the body color of a female crab (binary variable; C_2); 4) the spine condition of a female crab (binary variable; S_2).

Model	$m_{\mathrm{pp}}(\hat{\alpha}, \hat{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\psi}})$
W	0.347
Wt	0.334
S_2	0.046
C_2	0.044
W+Wt	0.348
W+C_2	0.337
W+S_2	0.328
Wt+C_2	0.327
Wt+S_2	0.320
C_2+S_2	0.068
W+Wt+C_2	0.339
W+Wt+S_2	0.334
W+C_2+S_2	0.321
Wt+C_2+S_2	0.317
W+Wt+C_2+S_2	0.328

calculated $m_{\mathrm{pp}}(\hat{\alpha},\hat{\beta} \mid \hat{\psi})$.

Table 1 shows the values of $(\hat{\alpha},\hat{\beta} \mid \hat{\psi})$ fom_{pp}r each model. m_{pp} suggested that the models involving only one or both of the two binary predictors C_2 and S_2 have low predictive power. This result is not surprising given one would expect a model fitted based on continuous as opposed to categorical predictors would offer better predictive performance when the response itself is not categorical. Also, the inclusion of both W and Wt simultaneously seems to not be favorable compared to including either one only, which is perhaps a reflection of the fact that W and Wt are highly correlated (Takahashi and Kurosawa, 2016). Our proposed measure $m_{\mathrm{pp}}(\hat{\alpha},\widehat{\beta} \mid \hat{\psi})$ exhibited similar trends to AIC, although noting that AIC also takes account the penalty for the number of parameters into it.

References

1. Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. John Wiley and Sons.
2. Eshima, N. (2004). Canonical exponential models for analysis of association between two sets of variables. *Statistics & Probability Letters*, 66:135–144.
3. Eshima, N. and Tabata, M. (2007). Entropy correlation coefficient for measuring predictive power of generalized linear models. *Statistics & Probability Letters*, 77:588–593.
4. Takahashi, A. and Kurosawa, T. (2016). Regression correlation coefficient for a Poisson regression model. *Computational Statistics and Data Analysis*, 98:71–78.
5. Zheng, B. and Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, 19:1771–1781.

A new composite wellbeing index: (Egypt & Malaysia comparison study)

Mahmoud Mohamed ElSarawy
Central Agency for Public Mobilization and Statistics (CAPMAS), Cairo, Egypt

Abstract

Gross domestic product (GDP) per capita or income is not the only factor to measure well-being of the community, there are a lot of variables that affect it such as achievement of equality and availability of health care and education services, availability of adequate housing, how to get the major needs for life, political participation and social activities, availability of a clean healthy environment. The objective of this study is building an index to measure the well-being for Egypt and Malaysia, using some economic and social variables and use this index to make a comparison between these two countries. The importance of this study relies on making a statistical analysis for the population characteristics in these countries and also stands on the development of variables and indicators used through a time series data when compiling the wellbeing index for each country, and also use these indicators in international comparisons. This study used the recent available data for these variables almost 2017, using 5 individual indices contain 19 variables to calculate Wellbeing index for Egypt and Malaysia. The study revealed that Egypt and Malaysia wellbeing index was 0.7 and 0.84 respectively.

Keywords

Community satisfaction; International comparisons; Quality of Life

1. Introduction

When talking about economics in some of African and Asian countries, it is really related to economic and social behaviour type. Measuring well-being involves identifying the key components of a good life and then selecting a set of indicators and variables that provide information about the progress of society with respect to these outcomes. Therefore, it is important to study the level of well-being and its indicators in all countries. In this study will focus on Egypt and Malaysia to calculate wellbeing index. The index of well-being appeared in the last decade and it has been measured using a variety of methodologies such as the Australian unity wellbeing index, Canadian index of wellbeing, and better life index which was adopted by the "Organization for Economic Cooperation and Development" (OECD).

The concepts of living standards are controversial concepts in the economic history and the challenging of raising the standard of living is the most important challenges facing planners and decision makers. Accordingly, the planners should determine the position of the components of the standard of living and also a clear plan that taken into account all of the income and standard of living variables before deciding to increase rates or improvement in living standards.

2. Importance of Study

We can get from the introduction that it is important to make a statistical analysis for the population characteristics for very high and medium human development countries and also stands on the development of variables and indicators used through a time series data when compiling the wellbeing index, and also use its indicators in international comparisons, also to monitor the value of the change in the levels of well-being for this country, identifying the variables responsible for the largest share of change, whether positively or negatively.

3. Objectives

Building an indicator to measure the well-being in Egypt and Malaysia in 2017, using 19 indicator and 5 indexes (GDP Index, Education index, Employment Index, Women Participation Index, Health Index). Also, we will use this indicator to do comparisons between Egypt and Malaysia community.

4. Methodology

This paper will use a new technique to measure the level of wellbeing for Egypt and Malaysia. The proposed technique includes 19 variables and indicators which related to measure the quality of social and economic life like equality in education, health care, economic security from job loss and unemployment indicators.

Before calculating the index of well-being, it must build the five guides of the key dimensions of well-being separate. And to calculate the evidence of these dimensions, it should first determine by maximum and minimum values (Hypotheses for the guide):

Table no. (1): Well-being guide Minimum and Maximum values

Guide	Indicator	Min	Max
1. Education Guide	Expected years of schooling (years)	0	15
	Mean years of schooling (years)	0	12
	Literacy rate (Adult (% ages 15 and older))	0	100
	Gross enrolment ratio (Tertiary (% of tertiary school–age population))	0	50
	Primary school dropout rate (% of primary school cohort)	0	10
2. GDP Guide	Gross domestic product (GDP) Per capita (Annual growth (%))	0	5
	Gross national income (GNI) per capita ((2011 PPP $)) USD	100	75000
	General government final consumption expenditure (Total (% of GDP))	0	15
	Consumer price index (2010=100)	0	250
3. Employment Guide	Employment to population ratio (% ages 15 and older)	0	70
	Youth Unemployment (% ages 15-24)	5	100
	Unemployment (Youth not in school or employment (% ages 15-24))	0	100
4. Women Participation Guide	Share of seats in parliament (% held by women)	0	15
	Woman Population with at least some secondary education (% ages 15 and older)	0	100
	Woman Labor force participation rate (% ages 15 and older)	0	70
5. Health Guide	Mortality rates (Under-five (per 1,000 live births))	5	50
	HIV prevalence, adult (% ages 15–49)	0	5
	Healthy life expectancy at birth (years)	10	70
	Current health expenditure (% of GDP)	0	5

5. Calculate the well-being guides

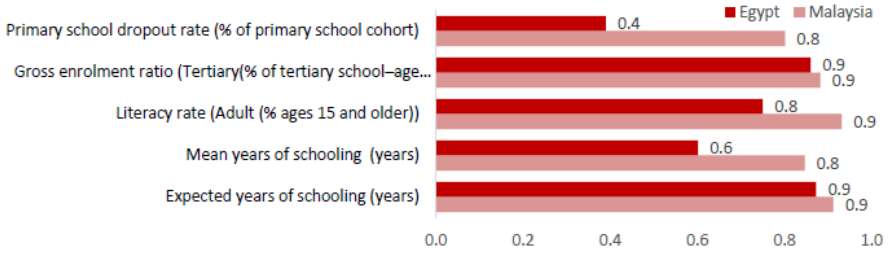
5.1 Well-being guide no. (1): Education Guide

Table no. (2): Education Guide

Indicator	year	Indicator Value		Partial Index Value		Total Index Value	
		Malaysia	Egypt	Malaysia	Egypt	Malaysia	Egypt
Expected years of schooling (years)	2017	13.7	13.1	0.9	0.9	0.87	0.69
Mean years of schooling (years)	2017	10.2	7.2	0.8	0.6		
Literacy rate (Adult (% ages 15 and older))	2006-2016	93.1	75.1	0.9	0.8		
Gross enrolment ratio (Tertiary (% of tertiary school–age population))	2012-2017	44	43	0.9	0.9		
Primary school dropout rate (% of primary school cohort)	2007-2016	8.0	3.9	0.8	0.4		

Source: Calculated by author.

Figure no (1): Education Partial Index Value



We can find that Malaysia achieve high value for all indicators in Education Guide compared with Egypt indicators and also Total Index Value for Malaysia is 0.87 and Egypt is 0.69.

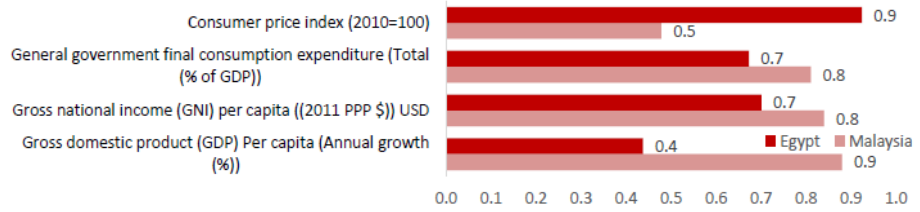
5.2 Well-being guide no. (2): GDP Guide

Table no. (3): GDP Guide

Indicator	year	Indicator Value		Partial Index Value		Total Index Value	
		Malaysia	Egypt	Malaysia	Egypt	Malaysia	Egypt
Gross domestic product (GDP) Per capita (Annual growth (%))	2017	4.4	2.2	0.9	0.4	0.75	0.68
Gross national income (GNI) per capita ((2011 PPP $)) USD	2017	26,107	10,355	0.8	0.7		
General government final consumption expenditure (Total (% of GDP))	2012-2017	12.2	10.1	0.8	0.7		
Consumer price index (2010=100)	2017	120	231	0.5	0.9		

Source: Calculated by author.

Figure no (2): GDP Partial Index Value



We can find that Malaysia achieve high value for most indicators in GDP Guide compared with Egypt indicators and also Total Index Value for GDP Guide for Malaysia is 0.75 and Egypt is 0.68.

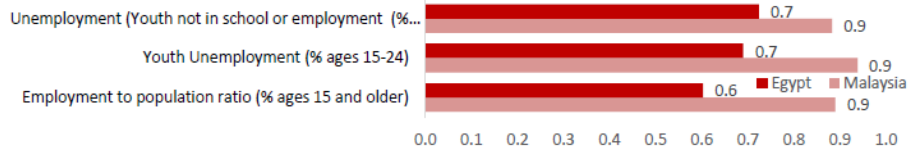
5.3 Well-being guide no. (3): Employment Guide

Table no. (4): Employment Guide

Indicator	year	Indicator Value		Partial Index Value		Total Index Value	
		Malaysia	Egypt	Malaysia	Egypt	Malaysia	Egypt
Employment to population ratio (% ages 15 and older)	2017	62.3	42.2	0.9	0.6	0.90	0.67
Youth Unemployment (% ages 15-24)	2017	10.9	34.4	0.9	0.7		
Unemployment (Youth not in school or employment (% ages 15-24))	2012-2017	11.7	27.6	0.9	0.7		

Source: Calculated by author.

Figure no (3): Employment Partial Index Value



We can find that Malaysia achieve high value for all indicators in Employment Guide compared with Egypt indicators and also Total Index Value for Employment Guide for Malaysia is 0.90 and Egypt is 0.67.

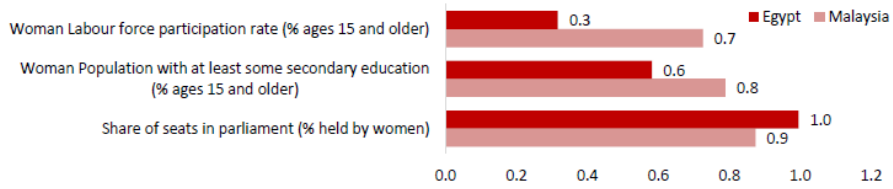
5.4 Well-being guide no. (4): Women Participation Guide

Table no. (5): Women Participation Guide

Indicator	year	Indicator Value		Partial Index Value		Total Index Value	
		Malaysia	Egypt	Malaysia	Egypt	Malaysia	Egypt
Share of seats in parliament (% held by women)	2017	13.1	14.9	0.9	1.0	0.80	0.63
Woman Population with at least some secondary education (% ages 15 and older)	2017	78.9	58.2	0.8	0.6		
Woman Labour force participation rate (% ages 15 and older)	2017	50.8	22.2	0.7	0.3		

Source: Calculated by author.

Figure no (4): Women Participation Partial Index Value



We can find that Malaysia achieve high value for most indicators in Woman Participantion Guide compared with Egypt indicators and also Total Index Value for Woman Participantion Guide for Malaysia is 0.80 and Egypt is 0.63.

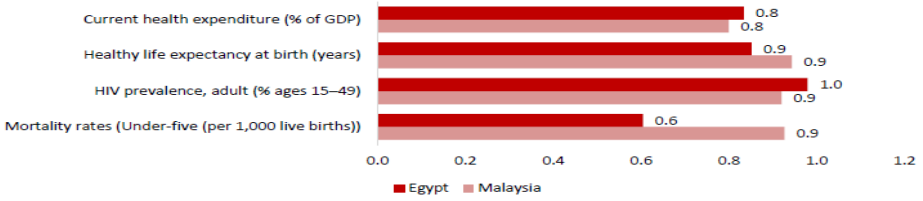
5.5 Well-being guide no. (5): Women Participation Guide

Table no. (6): Health Guide

Indicator	year	Indicator Value		Partial Index Value		Total Index Value	
		Malaysia	Egypt	Malaysia	Egypt	Malaysia	Egypt
Mortality rates (Under-five (per 1,000 live births))	2016	8.3	22.8	0.9	0.6	0.90	0.82
HIV prevalence, adult (% ages 15–49)	2016	0.4	0.1	0.9	1.0		
Healthy life expectancy at birth (years)	2016	66.6	61.1	0.9	0.9		
Current health expenditure (% of GDP)	2015	4.0	4.2	0.8	0.8		

Source: Calculated by author.

Figure no (5): Health Partial Index Value

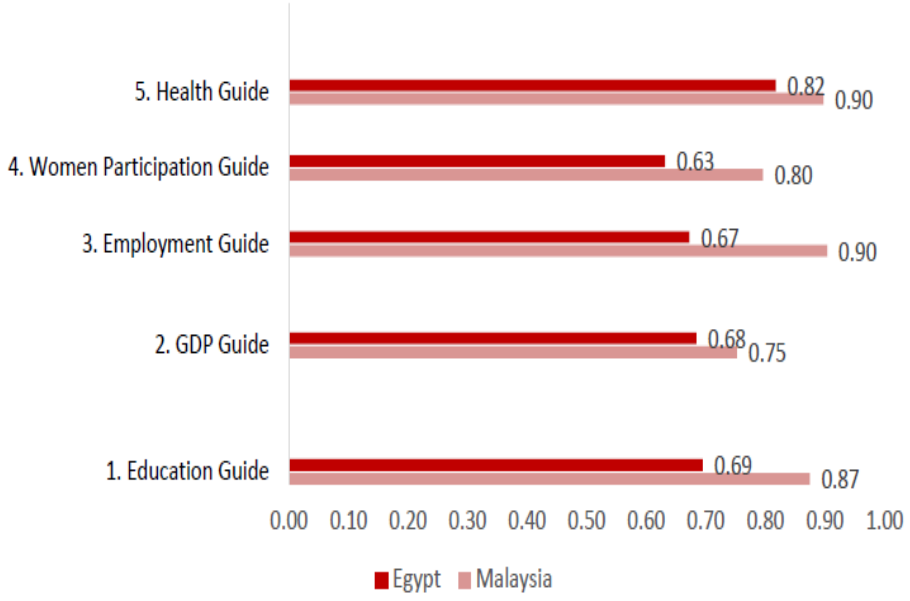


We can find that Malaysia achieve high value for most indicators in Health Guide compared with Egypt indicators and also Total Index Value for Health Guide for Malaysia is 0.90 and Egypt is 0.82.

6. Results

1. Well-being index = (GDP guide, Education guide, Employment guide, Women Participation guide, Health guide) /5
2. We found that Wellbeing Index values for Egypt and Malaysia are different, Malaysia has the largest index value of well-being with 0.84 and Egypt is 0.70 as shown in Figure (6)

Figure (6) : Wellbeing index guide



7. Conclusion

Although the indicators of well-being that can make new light for Economic and social Research, and measuring the welfare of the population, but it also received a number of criticisms. We can conclude that all objective standards for measuring wellbeing calculated partially and should be used exclusively to carry out a comprehensive assessment of the entire human well-being.

All countries should increase studies and researches aimed to measure the level of well-being in its societies to determine the extent to achieve the appropriate level of education, health and all economic and social indicators.

References

1. Axel Dreher, (2005), "The Influence of Globalization on Taxes and Social Policy – an Empirical Analysis for OECD Countries", Thurgau Institute of Economic, Switzerland and University of Konstanz, Department of Economics.
2. http://cdn.media70.com/national-accounts-of-well-being-report.pdf
3. McGillivray, Mark. (2007). Human Well-being: Issues, Concepts and Measures. In Mark McGillivray, ed. Human Well-Being: Concept and Measurement.
4. Van de Ven, Brugt Kazemier and Steven Keuning (1999), "Measuring Well-being with an Integrated System of Economic and Social Accounts," Di vision of Presentation and Integration, Department of National Accounts, Statistics Netherlands, May 17, unpublished report.
5. World Bank, (2006), "Equity and Development", World development report. Co-publication of the World Bank and Oxford University Press.

Mobile phone use and incidence of brain tumour

Ken Karipidis[1], Mark Elwood[2], Geza Benke[3], Masoumeh Sanagou[1], Lydiawati Tjong[1], Rodney J. Croft[4,5,6]

[1]Australian Radiation Protection and Nuclear Safety Agency
[2]School of Population Health, University of Auckland, Auckland, New Zealand
[3]School of Public Health and Preventive Medicine, Monash University, VIC, Australia
[4]Australian Centre for Electromagnetic Bioeffects Research
[5]Illawarra Health and Medical Research Institute
[6]University of Wollongong, NSW, Australia

Abstract

The incidence time trends of brain tumour in Australia were examined and the influence of improved diagnostic technologies and increase in mobile phone use on the incidence of brain tumours were identified. In a population based ecological study we examined trends of brain tumour over the periods 1982-1992, 1993-2002 and 2003-2013, using National Australian incidence registration data on primary cancers of the brain diagnosed between 1982 and 2013. We compared the observed incidence during the period of substantial mobile phone use (2003-2013) with predicted (modelled) incidence for the same period by applying various relative risks, latency periods and mobile phone use scenarios. The study included 16,825 eligible brain cancer cases aged 20 to 59 from all of Australia (10,083 males and 6,742 females). The main outcome measure was the Annual Percentage Change (APC) in brain tumour incidence based on Poisson regression analysis. The overall brain tumour rates remained stable during all three periods. There was an increase in glioblastoma during 1993-2002 (APC = 2.3, 95% Confidence Interval = 0.8-3.7) which was likely due to advances in diagnosis due to increases in the use of MRI during that period. There were no increases in any brain tumour types or sub-types during the period of substantial mobile phone use from 2003-2013. During that period there was also no increase in glioma of the temporal lobe (0.5, -1.3-2.3), which is the location most exposed when using a mobile phone. Predicted incidence rates were higher than the observed rates for latency periods up to 15 years. In Australia, there has been no increase in any brain tumour that can be attributed to mobile phone use.

Keywords

Poisson regression; brain tumour; mobile phone use

1. Introduction

Since its introduction in the mid-80s mobile phone use has grown rapidly worldwide. When using a mobile phone against the head, the brain is exposed to much higher levels of radiofrequency (RF) radiation than the rest of the body (1) and there has been continuing concern of a possible association with brain cancer. Several case–control and registry-based cohort studies have found little evidence to support such an association.(1) However a few other case-control studies have reported modest to large associations with glioma, the most common type of primary brain tumour.(2,3) These studies have found no association with other brain tumour types. Based on these results the International Agency for Research on Cancer (IARC) has classified RF as "possibly carcinogenic to humans".(4) a limited number of ecological studies have shown that although the prevalence of mobile phone use) has seen a massive increase, the time trends of brain tumour incidence have remained fairly stable.(5,6) There is also limited data on brain tumour histological types and anatomical location.

In this study, we analysed the incidence trends of brain tumour for three distinct time-periods to ascertain the influence of improved diagnostic methods and increase in mobile phone use. The analysis considered different histological types and sub-types, and glioma anatomical sites. We further compared the observed incidence during the period of substantial mobile phone use (2003-2013) with predicted incidence for the same period based on relative risks (RRs) reported by the two epidemiological studies forming the basis of the IARC classification. (2, 3)

2. Methodology

Incidence data on primary cancers of the brain and central nervous system diagnosed between 1982 and 2013 inclusive were obtained from the Australian Institute of Health and Welfare (AIHW).

Statistical Analysis of Observed Incidence: We analysed intracranial brain cancer incidence in adults aged 20-59Annual age-standardized incidence rates per 100,000 person-years were calculated separately for males, females and both genders by using the World Health Organization's (WHO) standard population. Histology was analysed by categorising glioma, meningioma, other histological types and brain cancers with unspecified histology.(7) We further analysed glioma by categorising glioblastoma (which is the most common brain tumour sub-type) and glioma location (frontal lobe, temporal lobe, parietal lobe, other locations, overlapping lobes and unspecified).

The incidence rates were low compared to the population at risk so the variability in the observed cases was assumed to follow a Poisson distribution.(8) Analyses of incidence time trends were carried out using Poisson regression to estimate the annual percent change (APC) in the

incidence, with corresponding 95% confidence intervals (CI) over three time-periods: 1982-1992 (representing increased CT and MRI use), 1993-2002 (representing advances in MRI) and 2003-2013 (representing substantial and increasing mobile phone use; more than 65% of the population).(9) Lowess smoothing was used in the graphical representation of the time trends.

Mobile Phone Use Data Sources: Mobile phone use was estimated using information on mobile phone accounts and survey data on actual use. Data on the annual number of mobile phone accounts from 1987, when mobile telephony first commenced in Australia, to 2013 was obtained from the national telecommunications regulator, the Australian Communications and Media Authority (ACMA). The number of mobile phone accounts per capita for each year was calculated by dividing the number of accounts by the total Australian population in that year (obtained from the Australian Bureau of Statistics), noting that since 2008 the annual number of accounts has been exceeding the number of people in the population. This data is not a true indication of mobile phone use as some users may have had more than one account and other users no account. A consumer survey conducted by ACMA reported that approximately 90% of the population used mobile phones in the years 2009 to 2013.(10) We estimated the annual prevalence of mobile phone use by multiplying the annual number of accounts per capita by a factor of 0.9.(10) It was not possible to stratify prevalence of use by age or gender; thus an overall estimate of prevalence is provided equally for all ages across the 20-59 age range and for both males and females.

Statistical Analysis of Predicted Incidence: With the assumption that mobile phone use is associated with glioma in adults as reported by the Interphone and Swedish studies, we calculated predicted incidence rates and time trends by applying various relative risks (RRs, 1.5, 2, 2.5, 3) and latency periods (1, 5, 10, 15, 20 years) for three different mobile phone use scenarios:

a) All users – RRs were applied to all mobile phone users
b) Heavy users – RRs were applied to heavy mobile phone users (defined as 19% of mobile phone users by the Interphone study)
c) Regular users and heavy users - RR of 1.5 applied to regular users (81% of all users) and RRs of 2, 2.5 and 3 applied to heavy users (19% of all users)

The annual predicted incidence rates were calculated for the period 1987-2013 using the formula:

$$\text{Predicted Incidence} = (P \times RR \times I_B) + ((1 - P) \times I_B)$$

where P denotes the annual prevalence of mobile phone use, RR the relative risk and I_B the pre-mobile phone baseline incidence from 1982-1987. Confidence intervals and statistical significance of observed and expected incidence rates were calculated using Poisson confidence intervals.(11) Analyses of predicted incidence time trends were carried out by estimating the

APC for the period 2003-2013, representing the time that mobile phone use increased rapidly.

We used Stata/SE 15.0 for all analyses. The reporting of our study conforms to the STROBE statement.(12)

3. Results

Observed Incidence: There was a total of 16,825 eligible brain cancer cases aged 20 to 59 (10,083 males and 6,742 females) that were diagnosed between 1982 and 2013. The observed incidence trends (given as APC) over the time-periods 1982-1992, 1993-2002 and 2003-2013 are shown in Table 1 for both genders The overall brain tumour rates remained stable in all three time-periods and the trends were similar for males and females. Glioblastoma increased during the period that saw advances in MRI (1993-2002) whilst it remained stable during the period of substantial mobile phone use (2003-2013); this later period also saw a decrease in other glioma sub-types. There was a strong decreasing trend in brain tumours with unspecified histology during the period of increased CT and MRI use (1982-1992). With the redistribution of unspecified tumours there were no significant changes to these histological trends (Table 2)

Table 1. Observed age-standardised brain tumour incidence trends in adults

	1982-1992		1993-2002		2003-2013	
	N	APC* (95% CI)	N	APC (95% CI)	N	APC (95% CI)
All	4793	0.1 (-0.8,1)	5270	0.5 (-0.5,1.5)	6762	-0.8 (-1.6,0)
Histology						
Glioma	4347	1.1 (0.2,2.1)	4990	0.4 (-0.6,1.4)	6421	-0.6 (-1.4,0.2)
Glioblastoma	1638	1.4 (-0.1,2.9)	2397	2.3 (0.8,3.7)	3291	0.8 (-0.4,2)
Other glioma	2709	1 (-0.2,2.2)	2593	-1.2 (-2.6,0.1)	3130	-1.8 (-2.9,-0.7)
Meningioma	82	-0.4 (-6.9,6.6)	110	2.4 (-4.2,9.4)	120	-4.4 (-10.1,1.7)
Other	79	-7.3 (-13.6,-0.6)	66	-1.5 (-9.5,7.2)	94	-5.3 (-11.3,1)
Unspecified	285	-13.4 (-16.6,-10)	104	4.6 (-2.3,12)	127	-4.8 (-10.3,0.9)
Glioma Location						
Frontal	933	7.8 (5.6,10.1)	1345	3.7 (1.8,5.7)	2144	3 (1.6,4.5)
Temporal	599	7.3 (4.6,10.1)	982	2.8 (0.6,5.2)	1371	0.5 (-1.3,2.3)
Parietal	655	6.4 (3.9,9.1)	801	-1.3 (-3.7,1.1)	816	-0.4 (-2.7,2)
Other locations	605	5.1 (2.5,7.8)	778	0.5 (-1.9,3)	989	-1.7 (-3.8,0.3)
Overlapping	298	3.5 (-0.1,7.3)	296	-8.8 (-12.5,-5)	374	-2.3 (-5.6,1.1)
Unspecified	1257	-10.8 (-12.4,-9.2)	788	-2.9 (-5.2,-0.4)	727	-10.5 (-12.7,-8.2)

*APC = Annual percent change

Table 2. Observed age-standardised brain tumour incidence trends in adults after redistribution of unclassified tumours

	1982-1992		1993-2002		2003-2013	
	N	APC* (95% CI)	N	APC (95% CI)	N	APC (95% CI)
All	4793	0.1 (-0.8,1)	5270	0.5 (-0.5,1.5)	6762	-0.8 (-1.6,0)
Histology						
Glioma	4623	0.2 (-0.7,1.2)	5094	0.5 (-0.5,1.5)	6547	-0.7 (-1.5,0.1)
Glioblastoma	1746	0.4 (-1.1,1.9)	2445	2.4 (0.9,3.8)	3353	0.7 (-0.5,1.9)
Other glioma	2886	0.1 (-1,1.2)	2649	-1.1 (-2.5,0.2)	3195	-1.9 (-3,-0.8)
Meningioma	84	-1.6 (-7.9,5.2)	110	2.4 (-4.2,9.4)	120	-4.4 (-10.1,1.7)
Other	82	-8.6 (-14.7,-2)	66	-1.5 (-9.5,7.2)	94	-5.3 (-11.3,1)
Glioma Topography						
Frontal	1447	1.8 (0.2,3.5)	1719	2.3 (0.6,4)	2580	1.6 (0.3,2.9)
Temporal	929	1.8 (-0.2,3.9)	1252	1.5 (-0.5,3.5)	1656	-1.2 (-2.8,0.4)
Parietal	803	3.4 (1.2,5.7)	894	-2 (-4.2,0.3)	880	-1.1 (-3.3,1.1)
Other locations	948	-0.5 (-2.5,1.5)	996	-0.8 (-3,1.4)	1198	-3.3 (-5.1,-1.4)

*APC = Annual percent change

There were increasing trends for all locations and a strong decreasing trend for unspecified location during the period of increased CT and MRI use (1982-1992) Table 1. There were also increases in the frontal and temporal lobes and a smaller decrease in unspecified location during the period of advances in MRI (1993-2002); this period also had a very large decrease in gliomas with overlapping location. During the period of substantial mobile use there were no increases in any of the locations apart from the frontal lobe and there was a strong decrease in unspecified location. With the redistribution of a high number of gliomas with unspecified and overlapping location there was a much lower increasing trend only for gliomas in the frontal lobe during all three periods and a large increase in the parietal lobe during the first period (Table 2).

Predicted Incidence: Assuming a causal association between mobile phone use and glioma, the predicted incidence trends for both genders during 2003-2013 by applying various relative risks, latency periods and mobile phone use scenarios are shown in Table 3. The predicted incidence trends showed an increase for most mobile phone use scenarios and latency periods that were modelled apart from a 20-year latency period. The highest expected trends were generally seen for a 10-year latency period, which was the latency period associated with mobile phones and brain tumour as reported in the Interphone and Swedish studies.

With a RR of 2 for all mobile phone users and a latency of 10 years, the predicted incidence rate for both genders in 2013 was 7.3 per 100,000 people (6.7 to 7.9) compared to the observed 4.5 per 100,000. The predicted rates increase to 8.7 (8.1 to 9.3) and 10.2 (9.5 to 10.8) per 100,000 for RRs of 2.5 and 3 respectively. With a RR of 1.5 for regular users and a RR of 2 for heavy users and a latency of 10 years the predicted rate was 6.1 per 100,000 (5.6 to 6.6); increasing to 6.4 (5.9 to 6.9) and 6.7 (6.1 to 7.2) when applying RRs of 2.5 and 3 to heavy users, respectively. Assuming a latency of 15 years, the predicted incidence rates in 2013 were also higher compared to the observed rate. The model did not show an increasing trend for a latency of 20 years.

Table 3. Predicted glioma incidence trends in adults 2003-2013

Scenario	Latency	RR=1.5 APC* (95% CI)	RR=2 APC (95% CI)	RR=2.5 APC (95% CI)	RR=3 APC (95% CI)
All Users	1	1.1 (0.3,1.8)	1.6 (1,2.3)	2.0 (1.4,2.6)	2.3 (1.7,2.8)
	5	2.8 (2,3.5)	4.5 (3.8,5.2)	5.7 (5.1,6.4)	6.6 (6,7.3)
	10	2.7 (1.9,3.6)	4.9 (4.1,5.7)	6.7 (5.9,7.5)	8.2 (7.4,8.9)
	15	1.3 (0.5,2.2)	2.5 (1.7,3.4)	3.7 (2.8,4.5)	4.8 (3.9,5.6)
	20	0.2 (-0.7,1)	0.3 (-0.5,1.2)	0.5 (-0.4,1.3)	0.6 (-0.2,1.5)
High Users	1	0.3 (-0.6,1.1)	0.5 (-0.3,1.3)	0.7 (-0.1,1.5)	0.9 (0.1,1.6)
	5	0.6 (-0.2,1.5)	1.2 (0.4,2)	1.8 (1,2.6)	2.2 (1.5,3)
	10	0.3 (-0.6,1.1)	0.5 (-0.3,1.4)	0.8 (-0.1,1.6)	1.0 (0.2,1.9)
	15	0.3 (-0.6,1.1)	0.5 (-0.3,1.4)	0.8 (-0.1,1.6)	1.0 (0.2,1.9)
	20	0.0 (-0.8,0.9)	0.1 (-0.8,0.9)	0.1 (-0.8,0.9)	0.1 (-0.7,1)
Regular users and high users			RR=1.5 (R), 2 (H)	RR=1.5 (R), 2.5 (H)	RR=1.5 (R), 3 (H)
	1		1.2 (0.5,1.9)	1.3 (0.6,2)	1.4 (0.8,2.1)
	5		3.2 (2.4,3.9)	3.5 (2.8,4.3)	3.9 (3.1,4.6)
	10		3.2 (2.4,4)	3.6 (2.8,4.4)	4.0 (3.2,4.8)
	15		1.5 (0.7,2.4)	1.8 (0.9,2.6)	2.0 (1.2,2.9)
	20		0.2 (-0.7,1)	0.2 (-0.6,1.1)	0.2 (-0.6,1.1)

*APC = Annual percent change

4. Discussion and Conclusion

The results of our study showed that the overall brain tumour rates in adults aged 20 to 59 years showed no increasing or decreasing trend. This is in line with studies showing stable brain tumour trends in other countries.(6) Furthermore, the trends in our study were stable for different histological types, like glioma, which has been reported in some case-control studies as being associated with mobile phone use.(2, 3) The all glioma incidence rates were stable in both the periods before (1982-1992, 1993-2002) and the period after (2003-2013) substantial mobile phone use. For a causal relationship between mobile phone use and brain cancer, one would expect an increasing trend in the later period and no trend in the earlier periods.

In our study there was an increasing trend for glioblastoma when looking at the entire observation period (1982-2013). However, when looking at different time periods there was no increase in the glioblastoma rates during the period of substantial mobile phone use but there was an increase in the glioblastoma rates in the earlier periods: 1982-1992 (non-statistically significant increase), which saw increased use of CT and MRI, and, 1993-2002 (statistically significant increase) which saw further advances in MRI. Technological developments in MRI during 1993-2002, including diffusion and perfusion imaging, improved significantly the discrimination of brain tumour types and sub-types.(9) Other factors, such as improved access to care and an increase in the number of specialists, may also have played a role in the increase.(5)

The results on anatomical location showed that there was an increase in gliomas located in the temporal and parietal lobes prior to the period of substantial mobile phone use, but not during it. Cardis et al (2008) reported that depending on the type of mobile phone and the manner in which it is

used, the RF energy absorption is at least several times higher in the temporal lobe than in the frontal lobe.(13) In our data there was a large number of gliomas with unspecified or overlapping location. Reclassification of these did reduce the trends for the temporal lobe during the periods before substantial mobile phone use, and for the frontal lobe during all the periods.

In our study we also compared the observed incidence with a modelled predicted incidence assuming a causal association between mobile phone use and glioma as reported in the Interphone and Hardell studies. The results suggest that, if the effects of mobile phones on glioma risk are real, then the incidence rates would be far higher than those observed. The present study has some limitations. The accuracy of the Australian cancer registration system in the early periods when it began in the 80s is unknown for all the states and territories.

We estimated mobile phone use using information on mobile phone accounts, and this may not be a true indicator of actual use as some people may have multiple accounts and others may use a phone without having an account. We mitigated this by also using data from a consumer survey conducted by the national telecommunications regulator on the proportion of the population using mobile phones. Information from the survey was only available from the years 2009 to 2013 and this was applied to data on the annual number of mobile phone accounts from 1987. However, mobile phone use patterns have likely changed from 1987 to 2009. Further, the exposure metric is unclear when investigating whether mobile phone use is implicated in brain cancer risk. Prevalence of phone use is a de facto measure for the amount of RF energy a person is receiving when using a mobile phone, and changes in technology and patterns of individual use were not taken into account in this investigation.

We estimated the prevalence of mobile phone use equally across the 20-59 age range and both males and females. The use of subscription data in early years is likely to underestimate prevalence of use in males and overestimate it in females given that users in early years were middle-aged working men on company mobile phone subscriptions.(14) In later years mobile phone use became equal between the two genders.(15)

For information on the proportion of regular and heavy mobile phone users we used data from the Interphone study, which also included data from Australia. Mobile phone use in the Interphone study was self-reported, relying on participants' recall of past phone use.(2) Sensitivity analyses on the Interphone methodology reported that for short term recall (up to a year) there was underestimation of phone use by regular users and overestimation by heavy users.(16) For longer recall (3 to 5 years) there was an underestimation of number of calls and an overestimation on the duration of

calls for all users.(17) Based on these findings it is likely that the proportion of heavy users in our study is overestimated.

Finally this is an ecological observational study, not based on individual data thus it is not possible to account for confounding factors. This study design is appropriate to define global trends. Further, the stable trend in brain tumour incidence could have concealed a true increasing risk related to mobile phone use which appeared flat due to declines in other risk factors.

In conclusion, we found no evidence that mobile phone use increased any brain tumour histological types or subtypes. There was an increase in the incidence of glioblastoma prior to the rapid increase in mobile phone use which was most likely due to improved diagnosis from MRI. Furthermore, there was no increase in gliomas of the temporal lobe, which is the most exposed location, during the period of substantial mobile phone use. The increase in gliomas of the temporal lobe and decrease in gliomas of unspecified location during the periods prior to substantial mobile phone use are in line with the theory of improved diagnosis from CT and MRI. Further, the predicted rates were higher than the observed rates for latency periods up to 15 years. These results do not support an association between mobile phone use and brain tumour, although the possibility of a small risk or a latency period of more than 15 years cannot be excluded. Future research should continue to investigate trends in brain tumour histological types, and anatomical location for a possible increase with a longer latency period.

References

1. Cardis E, et al. Distribution of RF energy emitted by mobile phones in anatomical structures of the brain. Physics in Medicine & Biology. 2008;53(11):2771.
2. SCENIHR. Final opinion on potential health effects ofexposure to electromagnetic fields (EMF). 2015.
3. INTERPHONE Study Group. Brain tumour risk in relation to mobile telephone use: results of the INTERPHONE international case–control study. International journal of epidemiology. 2010;39(3):675-94.
4. Ostrom QT, et al. The epidemiology of glioma in adults: a "state of the science" review. Neuro-oncology. 2014;16(7):896-913.
5. World Health Organization. WHO research agenda for radiofrequency fields. 2010.
6. Inskip PD, et al. Brain cancer incidence trends in relation to cellular telephone use in the United States. Neuro-oncology. 2010;12(11):1147-51.
7. Chapman S, et al. Has the incidence of brain cancer risen in Australia since the introduction of mobile phones 29 years ago? Cancer epidemiology. 2016;42:199-205.

8. Louis DN, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta neuropathologica. 2016;131(6):803-20.
9. Jensen P. Cancer registration: principles and methods: IARC; 1991.
10. Castillo M. History and evolution of brain tumor imaging: insights through radiology. Radiology. 2014;273(2S):S111-S25.
11. ACMA. Communications report 2015–16. 2016.
12. Ulm K. Simple method to calculate the confidence interval of a standardized mortality ratio (SMR). American journal of epidemiology. 1990;131(2):373-5.
13. Ho VK, et al. Changing incidence and improved survival of gliomas. European journal of cancer. 2014;50(13):2309-18.
14. Little M, et al. Mobile phone use and glioma risk: comparison of epidemiological study results with incidence trends in the United States. Bmj. 2012;344:e1147.
15. Barr ML, et al. Inclusion of mobile phone numbers into an ongoing population health survey in New South Wales, Australia: design, methods, call outcomes, costs and sample representativeness. BMC medical research methodology. 2012;12(1):177.
16. Vrijheid M, et al. Validation of short term recall of mobile phone use for the Interphone study. Occupational and environmental medicine. 2006;63(4):237-43.
17. Vrijheid M, et al. Recall bias in the assessment of exposure to mobile phones. Journal of exposure science and environmental epidemiology. 2009;19(4):369.

Parametric bootstrap based test for detecting changes in marriage patterns

Neela A Gulanikar, Akanksha S Kashikar
Department of Statistics Savitribai Phule Pune University, Pune, India

Abstract

Indian marriage markets have traditionally seen hypergamy, where females tend to marry up. Recent increase in the female literacy rates and increase in the female enrolment in higher education has resulted in some changes in the educational hypergamy. We try to develop a parametric bootstrap based test to examine the significance of changes in the marriage rates across different educational levels. The test can further be extended to other socio-economical groupings such as income groups, age groups etc. This will lead to better understanding of changes in marriage markets which further affect the future demographic structure of the country. Simulation studies are carried out to examine whether the test maintains the level and the changes in the power of the test for different parameter combinations.

Keywords

Educational hypergamy; Marriage markets; Parametric bootstrap

1. Introduction

Skewed sex ratio is an important problem in Indian context. Many parents still consider female children to be burden and hence, female foeticide and infanticide have been serious issues since long. At the time of independence, the literacy rates in India, especially those in female population were seriously low. However, the efforts by the government and different non-governmental organization (NGOs) such as free school education / scholarships to female children have resulted in the significant increase in the female literacy rates. The Figure 1 shows that the proportion of literate female has more than doubled over the last 35-40 years. Traditionally, hypergamy (marrying up) is a common phenomenon in the Indian society. One of the major types of hypergamy is the educational hypergamy, i.e., generally females tend to marry men who have attained a comparatively higher degree. However, the recent increase in the education levels in females are expected to have created some disturbances in this educational hypergamy. Since the number of educated females is on the rise, not all of them will be able to find a partner more educated than them. As a result, some women may resort to hypogamy (marrying down) and marry the men having lower education than them. As per our knowledge, so far this change has not been statistically examined in

the literature. We plan to address this issue in the current paper. Our data request for the same is pending and hence, in the current version we discuss our methodology using simulated datasets.

The long term effects of this include the marriage markets being affected by the phenomenon of `marriage squeeze'. The phrase marriage squeeze refers to the demographic imbalance in which the number of potential brides does not approximately equal the number of potential grooms. When not everyone has an opportunity to marry, some will be squeezed out of the marriage market. An excess of eligible women is called a female marriage squeeze; an excess of eligible men is called a male marriage squeeze. It may be caused due to unavailability of eligible partners in appropriate population strata defined by age, education, religion, or other social categories. Hence, if the society continues to value educational hypergamy, the females in the highest educational level and males in the lowest educational level are expected to be victims of marriage squeeze. Such marriage squeeze can further lead to changes in birth rates and thereby to changes in the entire population structure. Hence, examining these changes is crucial. Rest of the paper is organized as follows. Section 2 discusses the methodology for developing the test, section 3 discusses the simulation study and the observations from that study and section 4 gives some concluding remarks and discusses the future scope.

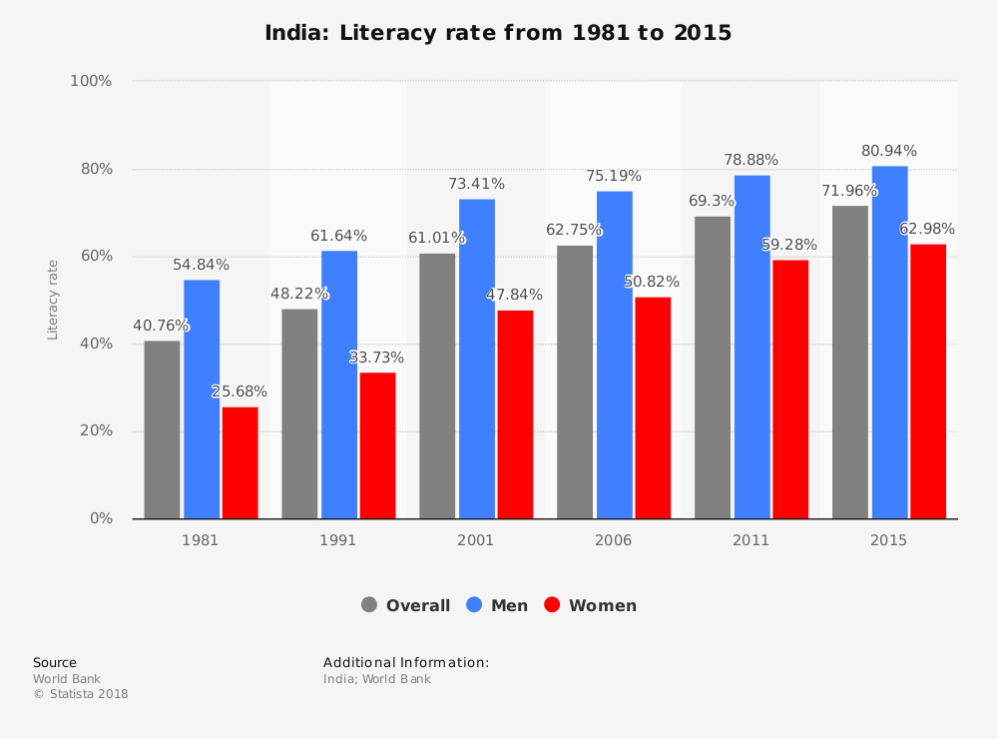


Figure 1: Female literacy rate in India

2. Methodology

2.1 Pair Formation

For modeling the number of marriages we plan to use the pair formation models or marriage functions. The homogeneous functions used to model the rates of pair-formation are commonly referred to as marriage functions. Most pair-formation models have been developed to study the dynamics of heterosexual populations that only include one single group of males and females. Heterogeneity, in a heterosexually-mixing population, is usually introduced by dividing the population of interest into subgroups (within each sex) based on attributes of interest to the modellers or the scientists (e.g., age, education, etc.).

In population theory Birth, death within certain ranges can be described by linear processes (Hadeler et al. 1988). The marriage or pair formation is an essentially nonlinear phenomenon. Pair-formation function is nonlinear, homogeneous of degree one and has certain monotonicity properties. An appropriate modeling the pair formation process has been called the two-sex problem".

Kendall (1949), Keyfitz (1972), Fredrickson (1971), MacFarland (1972), and Pollard (1973,Ch.7) have discussed ordinary differential equations and integrals describing the age structure. Hadeler et al. (1988) have developed an approach to homogeneous evolution equations, and their theory provides the appropriate framework for pair formation models. The process of pair formation is essentially nonlinear. They have suggested various functional forms for the rate of pair-formation or marriage function using harmonic mean, geometric mean and minimum function. Marriage function is a function Ψ of the population sizes of single males M and single females F. The basic properties satisfied by the marriage function are,

- $\Psi(M,F) \geq 0$
- $\Psi(M+u,F+v) \geq \Psi(M,F)$, for $u,v \geq 0$
- $\Psi(\lambda M,\lambda F) = \lambda\, \Psi(M,F)$, for $\lambda \geq 0$
- $\Psi(M,0) = \Psi(0,F) = 0$

Hadeler et al. (1988) given the following functions:

Harmonic mean function (HMF): $\psi(M,F) = 2\rho\left(\frac{MF}{M+F}\right)$

Geometric mean function (GMF): $\psi(M,F) = \rho\sqrt{MF}$

Minimum function (MF): $\psi(M,F) = \rho min\{M,F\}$

M,F describe the population sizes of single males and single females respectively, $\rho > 0$

Schmitz (2000) has also discussed pair formation using typical marriage functions, the Minimum Function (MF) and the Harmonic Mean Function (HMF).

2.2 Test Procedure

Suppose we have n distinct education levels. Hence, the total population T_F of females and total population T_M of males can be partitioned into n disjoint groups each. Suppose the respective population sizes be denoted by $M_1, M_2, \ldots, M_n$ and $F_1, F_2, \ldots, F_n$. From the definition, it is clear that

$$\sum_{i=1}^{n} F_i = T_F \quad \text{and} \quad \sum_{j=1}^{n} M_j = T_M .$$

Further, suppose ψ_{ij} denotes the number of marriages between a female from i^{th} level and a male from j^{th} level and ρ_{ij} denotes the rate of marriages between females from i^{th} level and males from j^{th} level in a given year. Our aim is to examine if the values of ρ_{ij} remain constant over the years. ψ and ρ denote the $m \times m$ matrices of ψ_{ij} and ρ_{ij} respectively.

Given the data on number of marriages across each educational level, i.e., ψ, we first try to estimate ρ. For estimating ρ, any of the following three formulae can be used.

- HMF: $\hat{\rho}_{ij} = \psi_{ij}\left(\frac{M_j + F_i}{2M_jF_i}\right)$
- GMF: $\hat{\rho}_{ij} = \frac{\psi_{ij}}{\sqrt{M_jF_i}}$
- MF: $\hat{\rho}_{ij} = \frac{\psi_{ij}}{min\{M_j, F_i\}}$

To determine which of the three estimates works well, we can compute the fitted values of ψ_{ij} by substituting $\hat{\rho}_{ij}$ the marriage functions given in the above subsection.

Once the best estimation method is chosen, we use those $\widehat{rho}_{ij}$ computed from two different years to determine whether the difference between the marriage rates across different subgroups is significant. To determine the significance of the test statistic, we use cutoffs based on parametric bootstrap. The procedure is described below.

Suppose $\hat{\rho}_1$ be the estimate of marriage rate matrix for the first year. Using the best marriage function chosen above, we simulate the matrices ψ_1^* and ψ_2^* using this estimate of marriage rate matrix for both the years. We then estimate the marriage rate matrices corresponding to these bootstrapped matrices for number of marriages. Suppose the two estimates are denoted by $\hat{\rho}_1^*$ and $\hat{\rho}_2^*$. We then compute the mean of squared relative differences for these two matrices. We repeat this procedure B(= *say,*5000) times and hence obtain B values of the mean of squared relative differences. This constitutes a sample of B mean of squared relative differences under the null hypothesis (as we have used only the estimate of $\hat{\rho}_1$). The 95^{th} percentile of this difference can be used as the cutoff. If the actual value of the mean of squared relative differences between $\hat{\rho}_1$ and $\hat{\rho}_2$ is greater than the above cutoff, we reject the null hypothesis and conclude that there is a significant change in the marriage rates between the two time points under study.

3. Results

As a starting point of the simulation study instead of starting with the completely random values, we use the education attainment data available for US population. The sex-wise population data for US for the years 2008, 2010, 2017 which contains 10 different levels of education namely None (no schooling), NonHighSchool, HighSchoolGrad, SomeCollNoDegree, AssoDegreeOccu, AssoDegreeAca, BachelorDegree, MasterDegree, ProfDegree, DoctDegree are available on the internet. The overall marriage rates are also available. Using these overall marriage rates and using the assumption of hypergamy we construct a 10×10 matrix ρ.

Using this same marriage rate matrix, we then generate three different marriage matrices (ψ) corresponding to three different years by using the three different matrices for educational attainment. This process is repeated 5000 times to get the parametric bootstrap based cutoffs. Further, this process is repeated for each of the three marriage functions. The cutoffs are reported in Table 1.

Table 1: Parametric Bootstrap based Cutoffs

Year	MF	GMF	HMF
2008-2010	0.0266	0.0605	0.0314
2010-2017	0.0249	0.0738	0.0428
2008-2017	0.0283	0.0592	0.0234

The number of rejections out of first 1000 samples generated above are reported in the column 'No change' in Table 2. This helps us in examining whether the test is able to maintain the level of significance as the simulations are done under the null hypothesis (common marriage rate matrix). It can be seen that MF function is the most successful in maintaining the level of significance whereas the performance of the HMF is the worst.

To examine the power of the test, we construct three new ρ matrices by making minor, moderate and major change in the original matrix respectively. In the minor change, the mean of squared relative differences between the two matrices is 0.00007084. For the moderate change this mean is 0.3788 and for the major change matrix, the corresponding value is 15.5437. We then carry out 1000 simulations under each of these setups and compute the number of rejections in each case by using the cutoffs provided in Table 1. The number of rejections are reported in Table 2. It can be seen that the power goes on increasing as the mean of squared relative differences between the two matrices goes on increasing. For the matrix with major change, all the marriage functions have succeeded in achieving power 1.

Table 2: Number of Rejections under Different Settings

Method	Year	Minor Change	Moderate Change	Major Change	No Change
MF	2008-2010	41	240	1000	54
MF	2008-2017	57	1000	1000	55
GMF	2008-2010	39	128	1000	29
GMF	2008-2017	7	112	1000	92
HMF	2008-2010	134	158	1000	150
HMF	2008-2017	137	249	1000	172

4. Discussion and Conclusions

The above results show that the parametric bootstrap based test performs quite well. In the case of MF, the test always maintains its level. The level is pretty much maintained for GMF as well. As far as power is concerned, the test does well for all the marriage functions. From the three different parameter combinations reported over here, it is clear that the test has good power for reasonably distant alternative. This test can further be extended for detecting the changes in the marriage rates across age-groups, income levels etc. Further, as and when the real data become available, we may be able to develop methods for choosing appropriate and more flexible marriage functions, e.g. the GMF and HMF can respectively be modified by using weighted GMF and weighted HMF respectively. The weights for the same may be chosen by comparing the fitted values obtained by different weights. Development of such methodologies will then result in a better understanding of marriage patterns and may lead to better prediction of changes in the demographic patterns which is especially essential for a developing country like India having huge population.

References

1. Hadeler K., Waldstatter R., Worz-Busekros A., (1988), Models for pair formation in bisexual populations, Journal of Mathematical Biology, (26), 635-649.
2. Fredrickson, A., (1971), A mathematical theory of age structure in sexual populations: Random mating and monogamous marriage models. Math.Biosciences, (10), 117-143.
3. Kendall, D., (1949), Stochastic processes and population growth. Roy.Statist.Soc., Ser B, (2) , 230-264.
4. Keytz, N., (1972), The mathematics of sex and marriage. Proc. of the Sixth Berkeley, Symposion on Mathematical Statistics and Probability, Biology and Health, 89-108
5. MacFarland, D.,(1972), Comparison of alternative marriage models. Population Dynamics (T.N.T.Grevilleed.), 89-106, Academic Press, New York London.
6. Pollard, J., (1973), Mathematical models for the growth of human populations, Cambridge University Press, Cambridge.
7. Schmitz, S. F. H. and Casttilo-Chavez C. (2000), A note on pair-formation functions. Mathematical and Computer Modelling, 31(4-5): 83-91

Big data predictive analytics using machine learning for official statistics

Nehall Ahmed Farouk Mohamed
Central Agency for Public Mobilization and Statistics (CAPMAS), Egypt

Abstract

Now one of the most argued topic in the statistical field is big data. Researches, data scientists, and statisticians worked to define it and evaluate the outcome of it over the statistical work on the different NSOs. Big data will not only enhance and improve the NSOs official statistical quality, but also help in predictive analytics in so many sectors. These predictive analytics will have a huge contribution in the national, international, and even the global development. As now it is not the time for knowing the current situation, but to know how the future will be. Big data predictive analytics has two main techniques based on the methodology, which are: regression techniques and machine learning techniques. The paper starts with discussing big data projects, applications, and the challenges that appeared according to the nature of big data. Then it focuses on big data analytics, especially predictive analytics. As it illustrate the challenges that result from the characteristics of big data in predictive analytics and propose methods to overcome it. Using the methodologies of machine learning in predictive analytics is a very important point to produce and to investigate its different techniques. The paper shows that deep learning is the most effective technique in machine learning while working with big data predictive analytics and shows some of its features and challenges. After that, the paper present suggestions to deal and solve machine learning challenges in big data predictive analytics. Also it shows is a modified prediction models over real-life hospital data collected from central China 2013-2015.finally it shows a huge progress in machine learning in this manner through dealing with: (incompleteness-missing values-parallel data and parallel models-the trained data – storage and central processing).

Keywords

Big data and official statistics; Machine learning and big data; Deep learning; Official statistics

1. Introduction

Defining the word big data was a huge demand through the previous years. As big data is considered to be a very huge rapid stream of data with main characteristics, which are: (Volume – Variety – Velocity). However big data as a concept nascent and has uncertain origins. Diebold (2012) argues that the term "big data . . . probably originated in lunch-table conversation at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently". Different levels had been passed through studying big data. There were several studies in big data, either from information technology (IT) aspects or from official statistics aspects. The task of extracting and using big data with official statistics had been discussed by many statistical agencies around the world. According to previous published studies and discussion papers, Australia, Netherland, and Italia have precedence over other countries in using big data with official statistics. Using big data in official statistics still an opened and important theme that needs investigations. On the other hand there are many studies about big data that completely focus on the IT perspectives only. So a shortage is considered in studying big data from both IT aspects and official statistics aspects to gather in one research. Here it should be remarked that, the main aim of this paper is to study big data in official statistics using the latest IT methodologies and techniques. It is important to get benefit from the previous experiences of using big data in official statistics, whether to be used as main source of data or to be integrated with official surveys data. Some of the projects that used big data in official statistical offices should be mentioned, in order to consider their challenges. The experience of Statistics Netherlands in the analysis of Traffic Loop Detection Data and the analysis of social media massages is one of them, also the Big Data Flagship Project of the Australian Bureau of Statistics (ABS). The main effect that is resulted from using big data analytics by national statistical offices (NSOs) should be mentioned as improving the international development.

2. Methodology

The paper integrates big data from official statistics perspectives and big data from IT perspectives. Through the first section, the paper explains using big data in official statistics by NSOs, international organizations, and statistical agencies. This will be clarified through several dimensions. The first dimension is showing examples of huge statistical projects in big data in different NSOs. Then the second one is naming the sources of big data that can be considered officially by UN statistical commission. Thirdly, this section produces the micro level that aggregates macro level effect of big data as development applications. Finally, mentioning the opportunities while using big data. Then the second section focuses on big data analytical techniques

for structured and unstructured big data. This section concentrates on big data predictive analytics. This is the point where it is needed to make a prediction from big data using IT techniques. So the last section illustrates how machine learning techniques that are used in big data predictive analytics. At last will be the result, conclusion, and future.

2.1. Section 1: Big data and official statistics (examples, sources, applications, and opportunities)

NSOs always seek to enhance and develop its statistical framework and improve its official statistics mission. Experiences like the analysis of Traffic Loop Detection Data, the analysis of social media massages, and also the Big Data Flagship Project in Netherlands and Australia. It is preferred to read (Piet J.H. Daas and etc. 2015) and (Siu-Ming Tam and Frederic Clarke 2015), considering the whole levels in these projects in details. According to UN statistical commission the sources of big data for official statistics are categorized into 6 categories which are:

1) The administrative records of governmental or private sector, like: electronic medical records, hospital visits, insurance records and bank records.
2) Tracking device sources, like: tracking data from mobile telephones and the Global Positioning System (GPS).
3) Commercial or transactional sources, like: credit card transactions and online transactions -including transitions from mobile devices - .
4) Sensor networks sources, like: satellite imaging, road sensors and climate sensors.
5) Opinion data sources, like: comments on social media.
6) Behavioral data sources, like: online searches and online page views.

(S.M. Tam & F. Clarke 2015) discussed each of these sources in depth, explaining the differences between identified big data sources and unidentified big data sources. Mentioning identified big data sources might refers for example to satellite images and unidentified refers to online prices. Both can be used in official statistics whether combined with data statistical data or used in statistical calculations.

(Hilbert, M 2016) reviewed empirical evidence and about 180 articles discussing big data for international development. This paper tries to emerge official statistical development using big data from the international development of it. As official statistics main purposes are help decision making to develop economic sector, agriculture, health, banking, and public services. So this is considered to be national and internal development aspects. (Hilbert, M 2016) Focused on the micro level big data application, that aggregates the macro level development applications, which are: (tracking words-tracking locations - tracking nature - tracking behavior - tracking production- tracking transitions -tracing other types).

These applications work on the different types of big data sources which were mentioned before. In the study of (Kalampokis, Tambouris, & Tarabanis 2013), it showed how big data different sources can help improving several different sectors, by providing the suitable big data analytics that can be used in official statistics. As 52 classifications of big data social media studies by source and area shown in figure (1). Big data helps NSOs with several opportunities to" reduce the cost of statistical production, improve the timeliness and frequency of its offerings and create new or richer statistics that meet emerging statistical data needs."(S.M. Tam& F. Clarke 2015) also it might be used as regular official statistics.

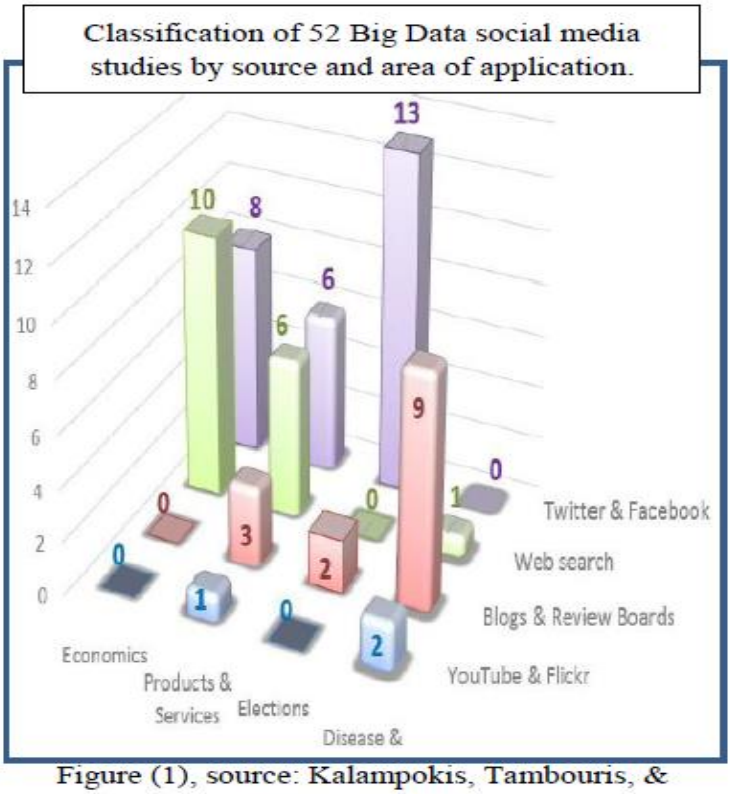


Figure (1), source: Kalampokis, Tambouris, & Tarabanis 2013.

2.2. Section 2: Big data analytical techniques (predictive analytics)

In order to extract the required knowledge from big data, two main processes are required, which are: data management and data analytics. Having a deeper look over the data analytical techniques, (Amir Gandomi, and Murtaza Haider 2015) mentioned 5 main big data analytical techniques which are: (text analytics - audio analytics - video analytics - social media analytics - predictive analytics). But the techniques of predictive analytics are the point where the paper needs to investigate. Predictive analytics of big data can predict the future upon current and historical data. It creates patterns that

consider the features and relations between data. These patterns of predictive analytics have main techniques based on the methodology used or the type of the outcomes variables. The two main techniques based on the methodology are regression techniques and machine learning (ML) techniques. The techniques based on the type of the outcomes variables depend on whether the variables are continuous or discrete, like: linear regression and random forest respectively.

There are main characteristics of big data that need investigation while dealing with predictive analytics. The first one is heterogeneity, which come as a result of different data sources or different populations. This can be overcome through making use of the huge size of data that might almost represent the population through producing sophisticated techniques. From this paper author point of view to overcome heterogeneity is to divide the huge set of data using stratum technique and to create a pattern for each stratum. As the characteristics of target groups in each stratum will be similar, so it will be homogenous. The second characteristic is error accumulation where simultaneous estimations of patterns occur. These consequences some parameters that affect the model might be considered as error accumulation. This might be overcome by dealing with each pattern separately before the simultaneous estimations, in order to be able to define the significant variables for each model but it might be quite difficult. The third characteristic is spurious correlation, where independent variables appear to be correlated as the size of data increase according to the study of (Fan and Lv 2008). However making classifications of each stream of data (granularity) and analysing it might solve this. As in the analysis of Traffic Loop Detection Data, worked on data according to certain time for example a crowded hour in the morning or evening was assumed to have same characteristics and analysis (Piet J.H. Daas and etc. 2015). The fourth one is incidental endogeneity that" refers to a genuine relationship between variables and the error term" (Amir Gandomi, and Murtaza Haider 2015).

2.3. Section 3: Machine learning and big data predictive analytics

These days NSOs and governments do not think only about the current figure of the situation in the different fields, but also there is a huge trend to predict the future. Predicting the future is important, whether it is based on structured data or unstructured data, big data of data from statistics. The previous section showed the main challenges that exist in big data predictive analysis, also tries to mention different ways to overcome it. Briefly

overcoming big data predictive analytics are: using many patterns – granularity- parallelism. Going in depth in using machine learning in big data analytics means the knowledge of the 4 techniques of it. Those techniques are: (data stream learning – deep learning – granular computing – incremental and ensemble learning). Here the reader is advised to read (Ahmed Oussous et al. 2018).

But it needs to be mentioned that deep learning is more effective while working with big data because it has some features, which are: (the ability to work in an environment that consists of a very huge number of data – it is based on hierarchy structure in learning - transfer the raw data into feature vector where the classifier can detect the patterns of the input). Now seeking to the perfect big data predictive analytics needs to investigate about the problem that deep learning in that manner. The challenges in deep learning are emerged of the challenges of big data predictive analytics, as it will affect the deep learning work processes. So deep learning will challenge the following issues: (dealing with continuous data streaming- data incompleteness- running time complexity and model complexity – inability to train data on central processor or storage- difficulty in parallelizing algorithms). Overcoming those challenges emphasize intensive study of each of them, in order to improve deep learning in predictive analysis. (Eric P. Xing et al.2015) mentioned data parallelism and model parallelism for ML big data analytics." In data-parallel ML, the data D is partitioned and assigned to computational workers (indexed by p = 1:P); we denote the data partition by Dp. We assume that the function Δ ( ) can be applied to each these data subsets independently, yielding a data-parallel update equation: = F (,Σ ,))"(Eric P. Xing et al.2015)-see figure (2). So this consideration solves the problem of data and model parallelism.it is suggested to solve the mentioned deep learning challenges, the following: (a) Data labeling and work on missing values. (b) Using decentralized system to train data. (c) Train sample of the data. (d) Data and model parallelism. After investigations it was founded that, some of these suggestions already taken into consideration in a real project China. It is a modified prediction models over real-life hospital data collected from central China in 2013_2015 –read (Min Chen et. al 2017). This project overcome the problem of incomplete data, parallel algorithms, proposed a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm, and with accuracy of the algorithm that reaches 94.8%.the following table (1), shows the different categories of data that were used.

The difference between data and model parallelism

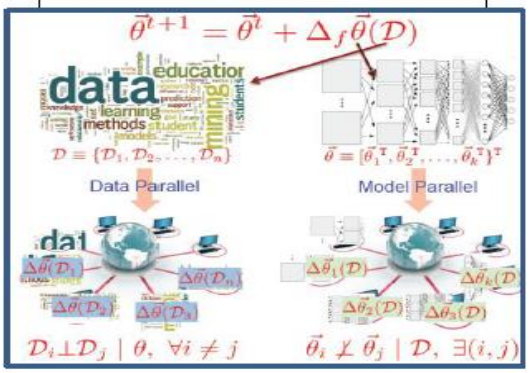


Figure (2), source: Eric P. Xing et al.2015.

were used.

Categories of data that were used in the models

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

Table (1), source: Min Chen et. al 2017.

Considering the work of the team in the project and the methods that was used to overcome ML problems in big data prediction models, the model's basic framework illustrated in figure (3).

The prediction model framework

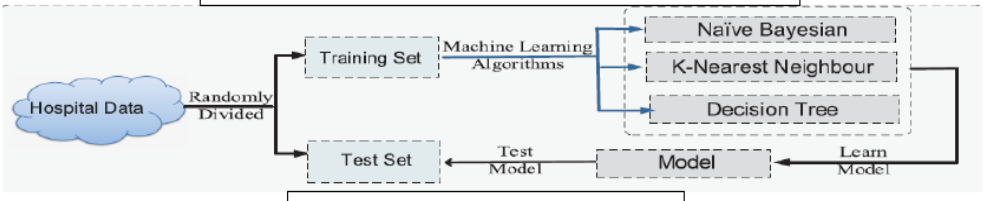


Figure (3), source: Min Chen et. al 2017.

3. Results

Solving the challenges of ML can be progressed through dealing with each problem from its perspective. Missing data and incompleteness can be solved before start working on the predictive mode, by labelling the data and estimating the missing values. Also the main problem of the huge and massive size of big data can be solved through using fine grain technique, where it is divided or classified. Parallel algorithms for data and models helped a lot in improving the running time complexity and model complexity. Training a

sample of the available big data set makes it quite easy, as it shrinks the size of data to be trained, also using decentralized storage.

4. Discussion and Conclusion

A massive work had been done in studying big data analytics, especially predictive analytics. Some problem blocks the completeness of using ML in big data predictive analysis. Recently some of these problems have been solved. The overall work of deep learning in this manner is enhanced than before. But there still an open issues to be studied in the future, like: (a) training a sample of the data, needs to define the adequate sample design. (b) The granularity issue needs to be studied over different categories and classifications to see how it might effect. (c) The issue of the continuous streaming of data.

References

1. Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology),70(5), 849–911.
2. Diebold, F. X. (2012). A personal perspective on the origin(s) and development of "big data": The phenomenon, the term, and the discipline. (ScholarlyPaper No. ID 2202843). Social Science Research Network.
3. Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. Internet Research, 23(5), 544–559. doi:10.1108/IntR-06-2012-0114.
4. O. Y. Al-Jarrah, P. D. Yoo, S Muhaidat, G. K. Karagiannidis, and K. Taha. Efficient Machine Learning for Big Data: A Review. Khalifa University, Abu-Dhabi, UAE, Data Science Institute, Bournemouth University, UK, University of Surrey, Guildford, UK, Aristotle University of Thessaloniki, Thessaloniki, Greece.
5. Bart Buelens, Piet Daas, Joep Burger, Marco Puts, and Jan van den Brakel (2014). Selectivity of Big data. Discussion paper. Statistics Netherlands.
6. Eric P. Xing, Qirong Ho, Wei Dai1, Jin Kyu Kim, Jinliang Wei, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanu Kumar, and Yaoliang Yu (2015). Petuum: A New Platform for Distributed Machine Learning on Big Data. School of Computer Science, Carnegie Mellon University.
7. Amir Gandomi, and Murtaza Haider (2015). Beyond the hype: Big data concepts, methods, and analytics. School of Management, Ryerson University. Canadaa.
8. Siu-Ming Tam and Frederic Clarke (2015). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. Methodology and Data Management Division. Australia.

9. Piet J.H. Daas, Marco J. Puts1, Bart Buelens, and Paul A.M. van den Hurk (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics.
10. Rob Kitchin and Gavin McArdle (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society.
11. Hilbert, M (2016). Big Data for Development: A Review of Promises and Challenges.Development Policy Review. California.
12. Min Chen, (Senior Member, IEEE), Yixue Hao, Kai Hwang, (Life Fellow, IEEE), Lu Wang, and Lin Wang (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE ACCESS.
13. Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih (2018). Big Data technologies: A survey. Journal of King Saud University, Computer and Information Sciences. Morocco.

Determinants of student satisfaction in higher education: A case of the UAE University

Ali Gargoum
College of Business & Economics, Department of Statistics, UAE University

Abstract

Promoting and measuring the students' satisfaction and happiness is a primary focus of higher education institutions during the last few years. The aim of this research is to introduce and test a conceptual model of students' satisfaction with a case study of the United Arab Emirates University (UAEU). A survey instrument designed and used for this purpose. Data (n = 498) were gathered on students from nine colleges of the UAEU. Exploratory and confirmatory factor analyses were implemented using structural equation modelling to test the proposed research model and hypothesis, which were based on a modified Parasuraman's SERVQUAL measurement tool. The results indicated that quality (service/ program) has significant impact on students' satisfaction and consequently on their happiness. Correlations between the quality constructs and the students' satisfaction were statistically significant. The study has shown that program quality, in terms of the university reputation, has the most influence, among quality constructs, towards confirming students' satisfaction. Moreover, results indicated, in general, that UAEU was successful in gaining student's satisfaction.

Keywords: Higher education, Student's Satisfaction, Servqual, Structural Equation model (SEM)

1. Introduction

In recent years, higher education worldwide witnesses great efforts for achieving the qualitative and quantitative development that takes place through the expansion of founding the private and public higher educational institutions. Students' satisfaction was a key factor of competitive lead. (Elwick and Cannizzaro, 2017). In the UAE, education is a priority at both the national level and the individual Emirates, this created an increasing demand on education sector, consequently, the education environment has become increasingly competitive, and many universities have begun to adopt new quality-oriented strategies as a result. In this research, we focus on measuring students' satisfaction based on the quality of service and programs. In particular, we aim to evaluate the student perceptions about the service/program quality of the UAEU. Moreover, to investigate interrelationships between service quality and student satisfaction using structural equations modelling (SEM). The conceptual model adopted in this

research was formulated to explore the influence of quality towards student's satisfaction (Weerasinghe et al., 2017). The supporting underpinning theory is the "equity theory". This theory has earned extensive recognition due to its capability in explaining customer behavior and customer satisfaction (Grigoroudis and Siskos, 2010).

The research framework is illustrated in Figure 1.

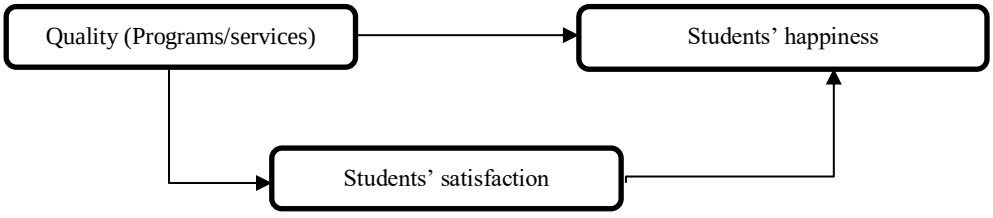


Figure 1: The research framework

The research framework of the study and the literature suggest the following main hypothesis: Quality is positively related to students' satisfaction and consequently to their happiness.

2. Methodology

This article aims to assess service quality and customer satisfaction using SERVQUAL. In spite of theoretical and operational criticisms, SERVQUAL has growing popularity and widespread applications. According to (Parasuraman et. al. 1985), when service quality is high, then this will lead to increase in customer satisfaction. This goes in line with many authors, see, for example, (Lee et. al., 2000, p 226) who acknowledge that customer satisfaction is based upon the level of service quality provided.

This study is performed empirically and results are based on assessment of UAE University students' responses. A questionnaire is designed using a modified SERVQUAL to model the service quality of the UAE University and how the quality of services could help in gaining students' satisfaction. The questionnaire is comprised of two parts. Part I includes 5 demographic questions. Part II includes 43 questions on the students' university experience, which are subdivided into 6 constructs, out of which 5 constructs are used to measure service quality and the 6th construct is used to measure the overall satisfaction about the quality service. The responses were measured on a 5-point Likert scale where [5] is assigned for "extremely satisfied", [4] for "satisfied", [3] for neutral, [2] for unsatisfied and [1] for extremely unsatisfied. The target population of the study comprises the undergraduate and postgraduate UAEU students enrolled in the Fall semester 2018. A total of 498 questionnaires were valid for the analysis. A pretesting of the questionnaire

has been, conducted before the actual study. A face- to -face interview with a representative set of 20 participants was carried out. A feedback from the respondents relevant to question ambiguity, ease of response to questions and the duration of the response has been, obtained.

The data were analyzed using SPSS 25 and AMOS graphics. A structural Equations Modeling (SEM) developed for fitting and testing a conceptual model of students' satisfaction to observed data.
(Awang, 2012).

3. Data Analysis and Result

Demographics of the data are presented in Table 1. A sample of 498 students participated of this study. Male respondents were (130) representing (26.1%) and Female respondents were 368 representing (73.9%). The respondents were from different academic levels and different colleges. Local and expatriates students were represented in the sample. The percentage of hostel students in the sample was (38%). In general, the sample was representative of the UAE University students.

Table 1: Demographics of the study

Gender	Frequency	Percentage	Academic Level	Frequency	Percentage
Male	130	26.1	Undergraduate	435	87.3
Female	368	73.9	Postgraduate	63	12.7
Total	498	100	***Total***	498	100
Nationality			Accommodation		
Emirati	395	79.3	Hostel	198	39.8
Expatriate	103	20.7	Other	300	60.2
Total	498		*Total*	498	100
College	Frequency	Percentage			
CBE	69	13.9			
CHSS	91	18.3			
CIT	48	9.6			
COE	135	27.1			
COL	26	5.2			
CFA	21	4.2			
COS	62	12.4			
CMHS	27	5.4			
COED	19	3.8			

First Exploratory Factor Analysis (EFA) performed to assess sources of variation and covariation in the observed measurements. Table 2 shows all factor loadings which are greater than 0.50. Scales with factor loadings of 0.50 and greater considered significant (Hair et al. 1998). Constructs with their Coronbach's alpha values are shown in Table 2.

Table 2: Standardized Regression Weights, Reliability, and Validity Assessment

Construct	Item	SRW	Cronbach's alphaalpha	CR	AVE
Tangibles	tan1	0.632	0.71	0.72	0.53
	tan2	0.619			
	tan3	0.629			
	tan4	0.630			
Reputation	rep1	0.672	0.77	0.75	0.55
	rep2	0.672			
	rep3	0.626			
	rep4	0.632			
Cooperation	coo1	0.821	0.88	0.87	0.63
	coo2	0.857			
	coo3	0.739			
	coo4	0.761			
Reliability	Rel1	0.653	0.86	0.86	0.55
	Rel2	0.708			
	Rel3	0.831			
	Rel4	0.780			
	Rel5	0.754			
	Res1	0.735	0.88	0.87	0.63
Responses	Res2	0.843			
	Res3	0.777			
	Res4	0.818			
	Sat1	0.850	0.83	0.84	0.63
Satisfaction	Sat2	0.725			
	Sat3	0.802			

SRW = Standardized Regression Weights, **CR** = Composite Reliability, **AVE** = Average Variance Extracted

Confirmatory factor analysis (CFA) conducted to assess the proposed research model (Osman et al., 2017). The study showed a good fit of the model to the data. The $\chi 2$ statistic was 480.073 (degrees of freedom 237, $p < 0.05$), with the ratio ($\chi 2$ /df) = 2.062 which is lower than the value of 5.0 as recommended by Hail et al. (1995, 1998, 2010) with lower values indicating a better fit. The incremental fit indexes were higher than 0.90, with CFI of 0.962, and TLI of 0.956. Of the absolute fit indexes, RMSEA was 0.045.

Moreover, there were some other valid conditions such as composite reliability (CR) and average variance extracted (AVE) checked. AVE exceed 0.50, indicating solid construct validity as shown in Table 2.

Table 3 presents the goodness of fit indexes. According to Hair et al. (1998), at least one index from each category must satisfy minimum acceptable limit for ensuring goodness model fit. Figure 2 shows the final fit structural model, after deleting some items due to factor loading less than 0.60 (Awang, 2012). Table 3 presents the proven fitness indexes.

Table 3: Goodness of fit indexes of structural fit model

Category	Index	Acceptable level
Absolute fit	RMSEA = 0.045	RMSEA < 0.08
Incremental fit	CFI = 0.962 TLI = 0.956	CFI > 0.90 TLI > 0.90
Parsimonious fit	χ^2/df = 2.062	Ratio < 5.0

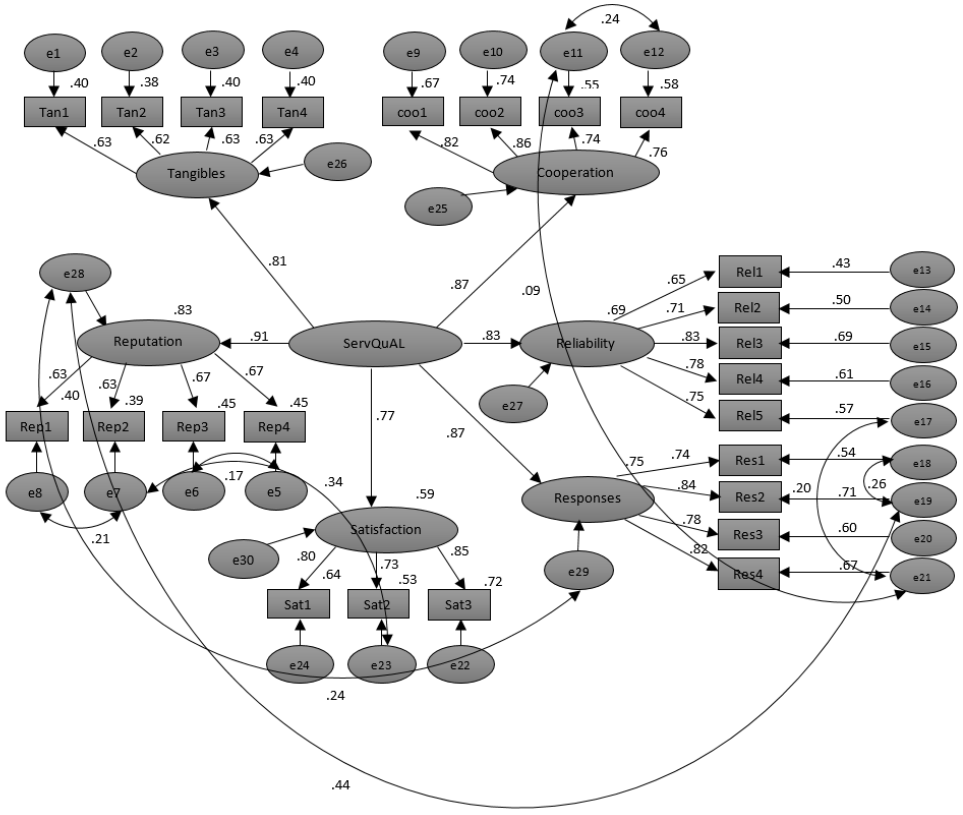


Figure 2: Standardized Structured Model of Student Satisfaction

The first construct tangibles is consisting of 4 items, all of them provide a positive and direct effect on service quality such as library facilities, food courts, safety on campus (0.81 * 0.63 = 0.510), and well equipped classrooms (0.81*0.62 = 0.502). The standardized regression weight for this construct is 0.81. This means that students perceive that tangibles have a positive effect on service quality of the UAEU. Second construct is reputation which is comprised of 4 items, all of them have a positive and direct effect on service quality, research and global degree (0.91*0.67 = 0.61) while high job market, and scholarships and financial assistance (0.91*0.63 = 0.57). All four items of this construct have a positive effect on service quality. The standardized regression weight for this construct is 0.91. This leads to conclude reputation of the UAEU among the students has a positive effect on quality. The third construct is cooperation and support (Empathy) with its four items, faculty concern (0.87*0.82 = 0.71); staff support (0.87*0.86 = 0.75); academic advising (0.87*0.74 =0.610 and management concern (0.87*0.76 = 0.660. All items have

positive and direct effect on quality with standardized regression weight for the construct of 0.87. The fourth construct is reliability that has five items, classes and exam scheduling (0.83*0.65 = 0.54); exams timing (0.83*0.71 = 0.59); faculty evaluation system (0.83*0. = 069.); lecturing times (0.83*0.78 = 0.65) and student fair evaluation (0.83*0.75 = 0.62). All items have positive and direct effect on quality with standardized regression weight for the construct of 0.83. The fifth construct is responsiveness that has four items, complains resolving efficiency (0.87*0.74 =0.64); responding to student matters or issues (0.87*0.84 = 0.73); faculty accessibility and contacts (0.87*0.78 = 0.68) and feedback mechanism (0.87*0.82 = 0.71). All items have positive and direct effect on quality with standardized regression weight for the construct of 0.87. Finley, service quality of the UAEU has a positive impact on student satisfaction with a standardized regression weight of 0.77. This positive quality will be the corner stone for gaining students' loyalty to the UAEU and attract internal and international students.

Table 4 shows correlations between the five quality constructs and student satisfaction. There is a significant correlation between all the constructs and the student satisfaction at 0.01 significance level. The highest correlation is between satisfaction and the reputation of the UAEU, which is 69.7%. All other correlations are similar which implies that service quality influenced student satisfaction.

Table 4: Correlation among quality constructs and student satisfaction

Constructs	Tangibles	Reputation	Cooperation	Reliability	Responses	Satisfaction
Tangibles	1					
Reputation	0.738*	1				
Cooperation	0.705*	0.789*	1			
Reliability	0.675*	0.755*	0.722*	1		
Responses	0.705*	0.698*	0.754*	0.722*	1	
Satisfaction	0.623*	0.697*	0.666*	0.638*	0.666*	1

***Significant at 0.01 level.**

4. Conclusion

This paper introduced a conceptual model for investigating student satisfaction with the service and program quality at the UAE University. The measures proposed were empirically tested and shown to be valid. Exploratory and confirmatory factor analyses empirically performed and validated the underlying dimensions of perceptions of student satisfaction. Results show that program quality and service provided by the UAEU have statistically significant impact on student satisfaction and consequently on student happiness. Findings have shown that UAEU reputation significantly influenced the student satisfaction. This indicates that student are more concern about program quality the most, as it establishes strong reputation and consequently creates more job opportunities. In the last few years, the UAEU assessed its

programs quality through international accreditation agencies such as AACSB, ABET etc. These accreditations added value to the university reputation and internationalized its programs. As a result, university leadership should focus progressively on improving the university programs quality, without ignoring other types of service quality, to gain the students' satisfaction.

References

1. Awang, Z. (2012). Structural Equation Modeling Using Amos Graphic: UiTM Press.
2. Elwick, A. and Cannizzaro, S. (2017). Happiness in Higher education, Higher Education Quarterly, 0951-5224 DOI: 10.1111/hequ.12121, Volume 71, No. 2, April 2017, pp 204–219.
3. Grigoroudis E. and Y. Siskos (2010), Customer Satisfaction Evaluation, Springer, New York.
4. Hair, J.F.J., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate Data Analysis, 5th edition. Prentice Hall, Upper Saddle River, New Jersey.
5. Lee, H., Lee, Y. & Yoo, D. (2000). The determinants of perceived service quality and its relationship with satisfaction, Journal of Service Marketing, 14(3), 217-231.
6. Osman, A., Saputra, R., Saha, J. (2017). Determinants of student satisfaction in the context of higher education: A complete structural equation modeling approach. British Journal of Marketing Studies. Vol.5, No.6, pp. 1-14, July 2017.
7. Parasuraman, A., Zeithaml, V. and Berry, L.L. (1985). A conceptual model of service quality and its implications for future research, Journal of Marketing, Vol. 49, Autumn, pp. 41-50.
8. Weerasinghe, I., Lalitha, R., Fernando, S., (2017). Students' Satisfaction in Higher Education Literature Review. American Journal of Educational Research, Vol. 5, No. 5, pp. 533-539.

Australian Labour Account: A new holistic view of people and jobs

Hayley Collett
Australian Bureau of Statistics, Canberra, Australia

Abstract

The Australian Labour Account has been developed to provide a framework for integrating labour data from a number of sources (including household survey, business survey, and administrative data). The result is consistent estimates of key labour market variables, which more effectively enable the description and analysis of the state and dynamics of the Australian labour market. These core variables can help users make sense of seemingly inconsistent labour related data, which are often based on different reference periods, populations, concepts, definitions and methodologies. The purpose of the Australian Labour Account is to support macro-economic analysis of peoples' participation in employment and related production over time. Its development provides an opportunity to significantly improve the quality of aggregates such as the number of jobs occupied within each industry, measures of hours worked, and improved labour productivity estimation. The concepts and definitions underlying the Australian Labour Account are built on International Labour Organisation (ILO) fundamentals, while expanding them to ensure consistency with the 2008 System of National Accounts (SNA08). The result provides a set of core macro-economic labour market variables derived through data integration, with both an industry focus and time series dimension.

Keywords

Employment; industry; productivity; macro-economy

1. Introduction

The Australian Labour Account provides a conceptual framework through which existing labour market data from different sources can be confronted and integrated, with the aim of producing a coherent and consistent set of aggregate labour market statistics.

The Australian Labour Account is macro-economic in scope, building on the International Labour Organisation (ILO) fundamentals and expanding them to ensure consistency with the Australian System of National Accounts (ASNA). It aims to extend the analytical capacity of national accounts data by providing a labour-specific lens. The Australian Labour Account produces a set

of statistical tables of employment related data that are consistent with the ASNA.

The ILO describes two approaches to compiling a labour account: a cross-sectional approach involving confrontation and reconciliation of key labour market measures, and a longitudinal approach which incorporates changes to population and labour force via births, deaths, and net migration, and includes measures such as duration of employment. The Australian Labour Account focuses on the cross-sectional approach (since this is the approach that supports data confrontation and reconciliation), and also provides a time series dimension.

The Australian Labour Account helps address data coherence by:

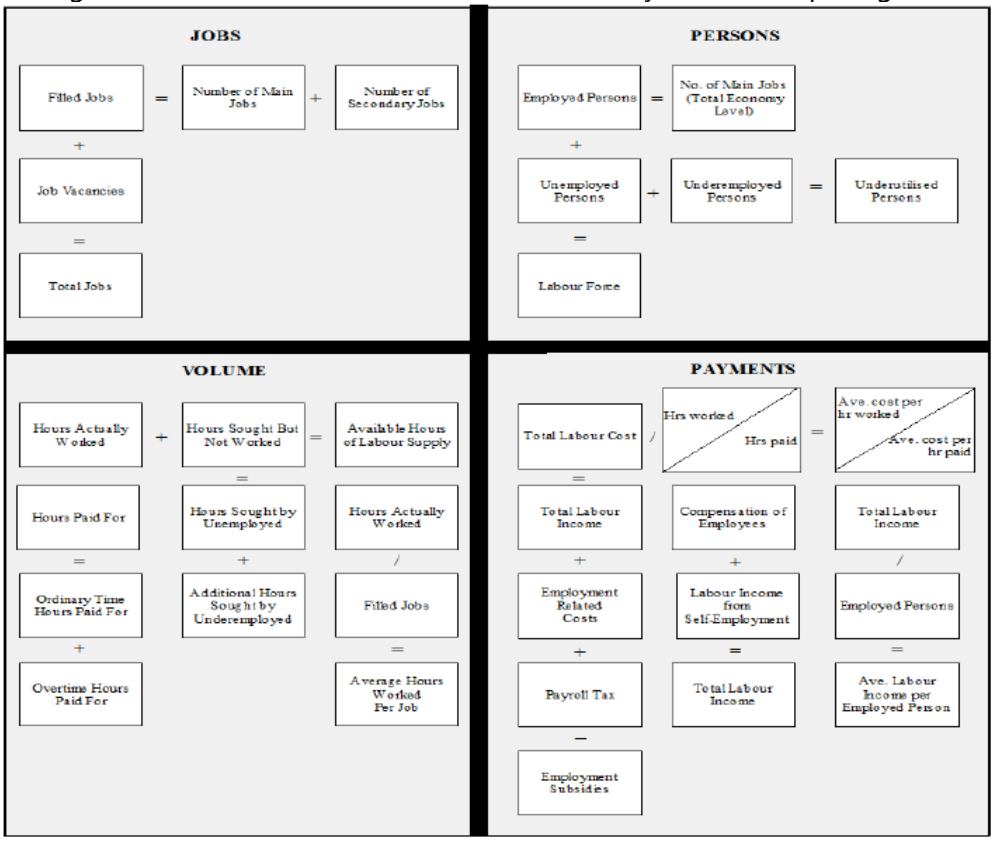
- bringing together related labour statistics from multiple sources in a single set of tables;
- applying a consistent set of concepts across the data to explore statistical anomalies;
- making transparent adjustments to data to offset conceptual and scope differences; and
- making further informed and documented data adjustments to provide a balanced set of labour statistics.

2. Methodology

Conceptual Framework

The Australian Labour Account consists of four distinct quadrants: jobs, persons, volume and payments (see Figure One). Data are available quarterly for 19 high level industry groupings, and annually for 86 detailed industry groups.

Figure One: Australian Labour Account: Identity Relationship Diagram



Scope of the Australian Labour Account

Accounting conventions are necessary to define the scope and treatment of activities that occur within the economy. The production and residency conventions adopted in the Australian System of National Accounts are used in the Australian Labour Account to determine the scope of activities covered, and the size of the economy measured.

The scope of the Australian economy defined by these conventions embraces the activities of all enterprises resident within Australia's economic territory engaged in the production of goods and services, which fall within the scope of the national accounts production boundary. The Australian Labour Account relates to the employment of all persons in jobs created by those enterprises. In this context, an enterprise is a productive undertaking maintained and controlled by one or more households, corporations or "quasi-corporations" that are resident in Australia's economic territory.

Enterprises include (for example):

- businesses operated by unincorporated self-employed trades persons,
- family operated farms,
- large corporations such as the major commercial banks and supermarket chains,

- Government departments and agencies, and
- schools and hospitals operated by the state, or by religious organisations and charities.

The national accounts production boundary embraces the production of all goods and services, with the exception of services produced by household controlled enterprises solely for consumption by the household itself. This exclusion relates to (for example) the cooking of meals for household members, household washing and cleaning and care of dependents. However, the "shelter services" provided by owner occupied dwellings are included within the production boundary.

Australia's economic territory includes all geographies under the control of the Australian Government, i.e. the Australian mainland, off-shore islands, Antarctic territories, Australian embassies and military establishments in other countries, and Australia's exclusive maritime economic zone. It excludes foreign embassies and military establishments in Australia.

An enterprise is considered "resident" if the "economic interest" of its controlling institutional unit (household, corporation or quasi-corporation) is centred in Australia's economic territory.

Scope Adjustments

Adjustments for scope and conceptual differences between data sources are required in compiling the Australian Labour Account. Scope adjustments are made in each of the four quadrants in the Australian Labour Account to ensure coherence. Scope adjustments made in one quadrant may be applied to another quadrant, and flow through to a third quadrant, based on the identity relationships.

Filled Jobs (business sources) is mainly based on summing estimates from two different business surveys. Data from a third source is added to account for employment in an industry division that is outside the scope of the primary sources. The following scope adjustments are made:

- add the number of persons from known industries excluded from primary business survey sources,
- add the number of persons employed in the permanent defence forces,
- add the number of unpaid contributing family workers,
- add the number of child workers who do not work for an employer (as they are excluded from business surveys), and
- subtract the number of persons from specific industry subdivisions duplicated in primary sources to avoid double counting.

Filled Jobs (household sources) is based on the number of jobs held by people employed in main jobs and secondary jobs soured from the Labour Force Survey (LFS), which is a household survey. Scope adjustments made to Filled Jobs (household sources) are similar to those made to Filled Jobs (business sources), to align the employed person estimates from the LFS with

production boundary and residency concepts of the Australian System of National Accounts. The following scope adjustments are made to Filled Jobs (household sources) to address LFS scope exclusions:

- add the number of persons employed in the permanent defence forces,
- add the number of child workers,
- add the number of main jobs held by non-resident visitors to Australia,
- add the number of secondary jobs held by non-resident visitors, and
- subtract the number of Australian residents working in Australia for non-resident enterprises.

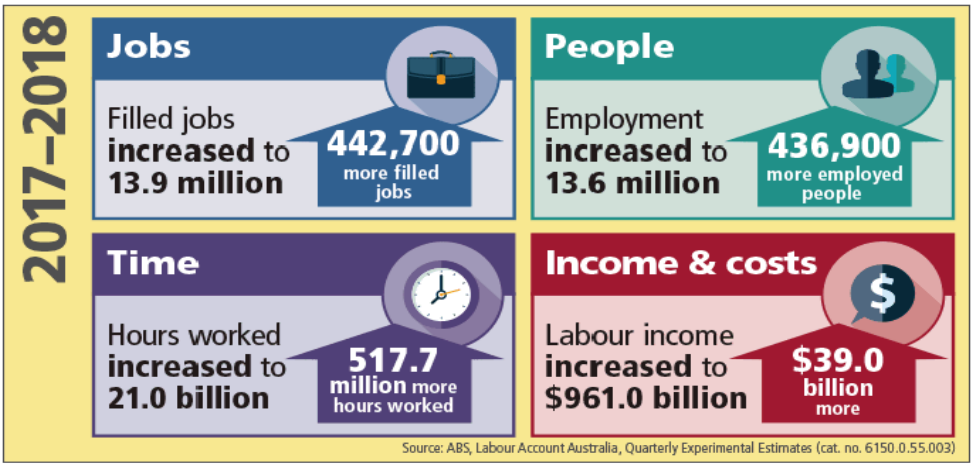
In compiling the Labour Account, residual differences remain between the estimated number of filled jobs based on business sources and those derived from household sources. These differences remain after making adjustments for known conceptual and scope differences. They represent measurement error in the respective sources, and are reflected in the "statistical discrepancy" series presented in the "unbalanced" data tables. In the balanced tables, separate business and household estimates have been replaced by a single "filled jobs" estimate. Consequent adjustments are also made to estimates of employed persons, hours worked and hours paid for. The harmonised, or "balanced", filled jobs series are based on a more detailed industry by industry investigation of the underlying sources of measurement error.

3. Results

Over the past five years health care and social assistance was the fastest growing industry, and remained the largest contributor to the number of jobs in the Australian economy. Of the 13.6 million employed people in Australia, 12.6 per cent work in the health care and social assistance industry. Filled jobs in Australia grew by 3.3 per cent, or 442,700, in 2017-18. The largest contributor to this increase was health care and social assistance filled jobs, which rose by 4.7 per cent. 2017-18 was the eighth consecutive year of jobs growth in this industry.

At the same time, manufacturing filled jobs saw a 1.2 per cent increase. This was mostly due to an increase of 19,300 fabricated metal product manufacturing filled jobs, and 8,700 non-metallic mineral product manufacturing filled jobs. These increases were offset by food product manufacturing filled jobs, which decreased by 5,900, and printing (including the reproduction of recorded media) filled jobs, which decreased by 7,900. The key results from the 2017-18 Australian Labour Account are summarised in Figure Two below.

Figure Two: Key Labour Account results, 2017-18



4. Discussion and Conclusion

The Australian Labour Account can help to inform on a number of macro-economic and policy related questions. Many of these questions are able to be answered for the first time through the holistic and integrated approach to analysing the labour market provided by the Australian Labour Account, while other questions are able to be answered in a more considered and comprehensive way.

For example, consider the answers to the following questions:

How many people are employed in Australia?

It depends on when you ask this, who you ask, and how you ask the question. Based on the answers provided by "responsible adults" from the households where workers live, the basic approach used in the Labour Force Survey, there were 12.5 million people employed in Australia in financial year 2017-18. Based on the answers provided by "responsible representatives" of businesses and other enterprises where they work, the approach adopted in business surveys, there were 13.2 million filled jobs in Australia in financial year 2017-18.

Why are the two figures different?

First, they are counting different things - for example, the Labour Force Survey asks about a person's main job to identify employed and unemployed people, and people not in the labour force. However, a person holding two jobs will be counted twice in a business survey, once by each employer. Business surveys measure the number of "filled jobs", not the number of employed people. When people in households were asked how many jobs they have, they told the Australian Bureau of Statistics (ABS) in 2017-18 they had 13.3 million. Businesses reported they had 13.2 million filled jobs, which was

100,000 (or 0.8%) more than reported by households in the monthly Labour Force Survey.

The second reason for the difference is that, in line with international standards, not everyone who has a job is in the scope of the Labour Force Survey. Similarly, some forms of work are not captured by reporting businesses. People whose main job is in the permanent military forces are not reported by either businesses or households, and household representatives are not asked to report on jobs held by people intending to stay in Australia for less than 12 months. No employment by children under 15 years, either paid or unpaid, is reported by households. In addition, unpaid contributions of work to a family business or farm by family members of any age are not reported by businesses. If the ABS adjusts for these known differences, then the number of filled jobs reported by businesses would be raised to 13.4 million, and the number of filled jobs reported by households would increase to 13.9 million. The remaining difference of 510,000 jobs, or 3.7% of the household based estimate, reflects the unavoidable measurement limitations related to measuring filled jobs and employment.

Likely sources of measurement error in household based data include lack of knowledge about the jobs held by household members on the part of the person responding to the Labour Force Survey. On the business survey side, there is no single ABS business survey that collects employment data from businesses across the whole economy, and business based estimates of filled jobs are compiled from multiple sources, potentially resulting in a larger overall total measurement error than in any of the individual sources. Both business and household surveys are also subject to sampling variability. Divergences can further arise when estimating missing data, or modelling is required to offset data gaps and lags in the supply of information.

How many hours were worked in Australia during 2017-18?

Based on hours worked reported by households, and after adjusting for defence force personnel, short-term visitors and children, 21,198 million hours were worked in 2017-18. Businesses reported the number of "hours paid for" at 21,946 million hours. These numbers imply that hours paid for but not worked, mainly various forms of paid leave, exceeded hours of unpaid overtime (hours worked but not paid for). This pattern was consistent over time at a whole of economy scale.

4. Conclusion

The Australian Labour Account provides a comprehensive picture of the labour market by building on and complementing other measures produced from survey based collections and administrative data sources. It allows for the analysis of both main and secondary jobs, and produces industry based estimates of the total number of persons employed for the first time. In

addition, the Australian Labour Account provides hours worked data which are on a consistent basis with measures of economic production, greatly improving the measurement of labour productivity.

Quantile residual life regression based on semi-competing risks data

Jin-Jian Hsieh, Jian-Lin Wang
Department of Mathematics, National Chung Cheng University Chia-Yi, Taiwan, R.O.C.

Abracts

This paper investigates the quantile residual life regression based on semicompeting risk data. Because the terminal event time dependently censors the non-terminal event time, the inference on the non-terminal event time is not available without extra assumption. Therefore, we assume that the non-terminal event time and the terminal event time follow an Archimedean copula. Then, we apply the inverse probability weight technique to construct an estimating equation of quantile residual life regression coefficients. But, the estimating equation may not be continuous in coefficients. Thus, we apply the generalized solution approach to overcome this problem. Since the variance estimation of the proposed estimator is difficult to obtain, we use the bootstrap resampling method to estimate it. From simulations, it shows the performance of the proposed method is good.

Keywords

Archimedean copula model; Bone marrow transplant data; Dependent censoring; Quantile residual life regression; Semi-competing risks data.

1. Introduction

Quantile regression can provide covariate effects for different quantile, which is more robust than ordinary least squares regression. Quantile regression was originally introduced by Koenker and Bassett (1978), and it has been widely investigated by many literatures, such as Powell (1984, 1986), Ying, Jung and Wei (1995), Portnoy (2003), Peng and Huang (2008) for censored data. Peng and Fine (2009) studied quantile regression for competing risks data, which constructs the model based on conditional quantiles with the cumulative incidence function. Hsieh et al. (2013), Hsieh and Hsiao (2015), and Hsieh and Wang (2017) studied quantile regression for semi-competing risks data based on the inverse probability weight technique, a weighted approach, and the counting process approach, respectively. In many medical research, the residual life is of interest. The residual life of a patient can be prolonged by a medical treatment. The quantile residual life regression also has been widely investigated by many literatures, such as Gelfand and Kottas (2003), Jeong, Jung and Costantino (2008), Jung, Jeong and Bandos (2009) and Ma and Yin (2010) for censored data. Gelfand and Kottas (2003)

formulated a semiparametric median residual life regression model based on a accelerated failure time regression model. Jeong, Jung and Costantino (2008) used a simple approach to estimate the median residual lifetime, which applied the technique by inverting a function of the Kaplan-Meier estimators. Jung, Jeong and Bandos (2009) applied the inverse probability weight method for log-linear quantile residual life regression model. Ma and Yin (2010) suggested a general class of semiparametric median residual life models. However, quantile residual life regression has not been studied for semi-competing risks data yet. Based on this motivation, we study the quantile residual life regression for semi-competing risks data in this article. This article investigates the quantile residual life regression based on semi-competing risk data. Because the non-terminal event time is dependently censored by the terminal event time, the inference on the non-terminal event time is not available without extra assumption. Therefore, we assume the non-terminal event time and the terminal event time follow an Archimedean copula. Then, we apply the inverse probability weight technique to constructing an estimating equation of quantile residual life regression coefficients.

2. Methodology

In this paper, we study quantile residual life regression based on semi-competing risk data. Assume that T is the non-terminal event time and D is the terminal event time. In addition, T may be dependently censored by D. Let C be the right censoring time which is assumed to be independent of (T, D) given covariates. Therefore, the observed variables are $\{(X_i, Y_i, \delta_{xi}, \delta_{yi}), i = 1, \ldots, n\}$, where $X = T \wedge D \wedge C, \delta_x = I(T \leq D \wedge C), Y = D \wedge C, \delta_y = I(D \leq C)$, $\wedge$ is the minimum operator and $I(\cdot)$ is the indicator function, which is called as semi-competing risks data. With covariates, we usually investigate the relationship between the response and covariates via regression model. However, the quantile residual life regression model can provide a more intuitive interpretation in medical or other related research. Suppose $\theta_{\xi|t}$ defines the $\xi-$quantile residual life function at time t. Then, $\theta_{\xi|t} = \xi-\text{quantile}(T_i - t|T_i \geq t)$ satisfies the relation $P(T_i - t \geq \theta_{\xi|t}|T_i \geq t) = 1 - \xi$, which is equivalent to $P(T_i \geq t + \theta_{\xi|t}) = (1 - \xi)P(T_i \geq t)$. It can be more clearly aware of the impact in test drug or clinical trial. Then, we consider the log-linear quantile residual life model as:

$$\xi - quantile\,\{\log(T_i - t_0)|T_i \geq t_0, Z_i\} = \beta_{\xi|t_0}^T Z_i, \qquad (1)$$

where $\beta_{\xi|t0}$ is a vector of the quantile regression coefficients, and Z_i is a vector of discrete covariates for a subject i. Because T is dependently censored by D, it is necessary to specify the relationship between T and D. In this paper, we assumed that T and D follow an copula model.

Nextly, we introduce the inference procedures for quantile residual life regression based on semi-competing risks data. Under model (1), it has $1-\xi = P\{T_i \geq t_0 + \exp(\beta_{\xi|t0}^T Z_i) | T_i \geq t_0, Z_i\}$. Under the semi-competing risks data, it has

$$E\left[I\{X_i \geq t_0 + \exp(\beta_{\xi|t0}^T Z_i)\}\delta_{x_i}/H_{z_i}(X_i)|Z_i\right] = P\left(T_i \geq t_0 + \exp(\beta_{\xi|t0}^T Z_i) \,|\, Z_i\right) \quad (2)$$

Similarly,

$$E\left[I\{X_i \geq t_0\}\delta_{x_i}/H_{z_i}(X_i)|Z_i\right] = P(T_i \geq t_0|Z_i), \quad (3)$$

where $H_{z_i}(X_i) = G_{z_i}(X_i) \times S_{D|T,Z}(x_i|x_i, z_i) = P(C > x_i|Z = z_i) \times P(D > x_i|T = x_i, Z = z_i)$

Therefore,

$$1-\xi = \frac{E\left[I\{X_i \geq t_0 + \exp(\beta_{\xi|t0}^T Z_i)\}\delta_{x_i}/H_{z_i}(X_i)|Z_i\right]}{E\left[I\{X_i \geq t_0\}\delta_{x_i}/H_{z_i}(X_i)|Z_i\right]},$$

which is equivalent to

$$E\left[\frac{I\{X_i \geq t_0 + \exp(\beta_{\xi|t0}^T Z_i)\}\delta_{x_i}}{H_{z_i}(X_i)}\middle| Z_i\right] - (1-\xi)E\left[\frac{I\{X_i \geq t_0\}\delta_{x_i}}{H_{z_i}(X_i)}\middle| Z_i\right] = 0.$$

Hence, we can construct the following estimating equation of $\beta_{\xi|t0}$ as:

$$S_n(\beta_{\xi|t0}) = \sum_{i=1}^{n} Z_i \left\{\frac{I\{\log(x_i - t_0) \geq \beta_{\xi|t0}^T Z_i\}\delta_{x_i}}{\widehat{H}_{z_i} x_i} - (1-\xi)\frac{I\{x_i \geq t_0\}\delta_{x_i}}{\widehat{H}_{z_i} x_i} = 0\right\} \quad (4)$$

where $\widehat{H}_{z_i}(x_i) = \hat{G}_{z_i}(x_i) \times \hat{S}_{D|T,Z}(x_i|x_i, z_i)$. $\hat{G}_{z_i}(x_i)$ could be estimated by Kaplan and Meier (1958) based on the data $\{(Y_i, 1-\delta_{yi}), i = 1, \ldots, n \text{ and } Z = z_i\}$ within each discrete covariate stratum, and since the non-terminal event time T may be dependently censored by the terminal event time D, it becomes more difficult to make inference on T. Therefore, we assume that (T, D) follows an Archimedean copula on the upper wedge as $P(T > t, D > d|Z) = \phi_{\alpha_z}^{-1}\{\phi_{\alpha_z}(S_T(t|Z)) + \phi_{\alpha_z}(S_D(d|Z))\}, t < d$. $\hat{S}_{D|T,Z}(x_i|x_i, z_i)$ can be derived as $\hat{S}_{D|T,Z}(x_i|x_i, z_i) = \hat{P}(D > x_i|T = x_i, Z = z_i = \frac{\phi_{\alpha_z}(\hat{S}_{T|Z}(x_i|z_i))}{\phi'_{\alpha_z}(\hat{S}_{W|Z}(x_i|z_i))}$, and the survival function of T and α can be estimated by the copula-graphic estimator by Lakhal, Rivest, and Abdous (2008). By the uniform convergence properties of the Kaplan-Meier estimator, the consistency property of $\hat{\alpha}$, and the continuous mapping theorem, we can construct the uniform convergence property of $\widehat{H}_z(x)$. Then, by the same way in the Appendix A, B, C in Jung et al. (2009), we can construct the consistency property and the
asymptotic normality property of $\hat{\beta}_{\xi|t0}$.

Because the equation (4) contains an indicator function of $\beta_{\xi|t0}$, it may not be continuous. An exact zero-crossing of $S_n(\beta_{\xi|t0})$ may not exist. However, Peng and Fine (2009) provided a generalized solution to estimate $\beta_{\xi|t0}$. The

generalized solution of $S_n(\beta_{\xi|t0})$ can be rewritten as the minimizer of the following function,

$$U_n(\beta_{\xi|t0}) = \sum_{i=1}^{n} \delta_{x_i} \left| \frac{\log(x_i - t_0) - \beta_{\xi|t0}^T Z_i}{\widehat{H}_{z_i}(x_i)} \right| + \left| M - \sum_{i=1}^{n} \frac{\beta_{\xi|t0}^T Z_i \delta_{x_i}}{\widehat{H}_{z_i}(x_i)} \right| + \left| M + 2(1-\xi) \sum_{i=1}^{n} I\{X_i \geq t_0\} \frac{\beta_{\xi|t0}^T Z_i \delta_{x_i}}{\widehat{H}_{z_i}(x_i)} \right|,$$

where M is an extremely large positive value larger then $\sum_{i=1}^{n} \beta_{\xi|t_0}^T Z_i \delta_{x_i} / \widehat{H}_{z_i}(x_i)$ and $\sum_{i=1}^{n} 2\beta_{\xi|t_0}^T Z_i (1-\xi) I\{x_i \geq t_0\} \delta_{x_i} / \widehat{H}_{z_i}(x_i)$. Because the variance of $\hat{\beta}_{\xi|t0}$ is difficult to estimate, we use the bootstrap resampling method to estimate the variance of $\hat{\beta}_{\xi|t0}$. Firstly, we obtain a resampling data from the original data as $\{(x_i^*, y_i^*, \delta_{xi}^*, \delta_{yi}^*, z_i^*), i = 1, \dots, n\}$. Secondly, estimate the parameter $\beta_{\xi|t0}$ based on the bootstrapping sample, denoted as $\hat{\beta}^*$. Then, repeat this process B times. We obtain the estimators $\{\hat{\beta}_b^*, b = 1, \dots, B\}$. Therefore, we can estimate the variance of $\beta_{\xi|t0}$ by $\widehat{var}(\hat{\beta}_{\xi|t0}) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_b^* - \bar{\beta}^*)^2$, where $\bar{\beta}^* = \sum_{b=1}^{B} \hat{\beta}_b^* / B$. Hence, we can construct the $100(1-\alpha)\%$ confidence interval for $\beta_{\xi|t0}$ as $\beta_{\xi|t0} \pm z(\alpha/2)\widehat{SD}$, where $\widehat{SD} \sqrt{\widehat{Var}(\hat{\beta}_{\xi|t_0}^T)}, z(\alpha/2) = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.

3. Results

In this section, we conduct simulation studies to examine the finite sample performance of the proposed approach. We consider the log-linear quantile residual life model as:

$$\xi - \text{quantile}\{\log(T_i - t_0) | T_i \geq t0, Zi\} = \beta_0 + \beta_1 Z_i, \qquad (5)$$

where we take the true values of $\beta_0 = -1.5$ and $\beta_1 = -0.5$. The covariate Z_i is generated from bernoulli distribution with mean 0.5. The terminal event time D is generated from exponential distribution with mean a, and the non-terminal event time T is generated from exponential distribution with mean $1/\{-log(1-\xi)exp(-(\beta_0 + \beta_1 Z))\}$ which is associated with model (5). Further, (T, D) follow the Clayton copula, where $\phi_{az}(v) = (v^{-\alpha z} - 1)/\alpha_z$ and $C_{az}(u, v) = (u^{-\alpha z} + v^{-\alpha z} - 1)^{-1/\alpha z}$. The right censoring time C follows a uniform distribution on $[0, b]$. We set Kendall's $\tau = 0.3, 0.5, 0.7$, quantile $\xi = 0.5$, sample size $n = 200$ and $t_0 = 0, 0.07, 0.17, 0.35$ which is the $0, 25\%, 50\%$ and 75% quantile of T. We consider the settings $(\xi, a, b) = (0.5, 0.57, 3)$. For each case, we replicate 400 simulation runs with 100 bootstrapping times.

We compare our proposed method with the method by Jung et al. (2009), which didn't consider the association between T and D. We present the bias of the proposed estimator (Bias), the empirical standard deviation (EmpSd), the average of estimated standard deviation (AveSd) based on the bootstrap

method, the coverage probability of the 95% confidence intervals (CP%) and the mean squared error (MSE). The results are shown in Tables 1. We denote the method by Jung et al. (2009) as the old method. The mean squared error of our method is smaller than the old method. In particular, the bias of β_1 of our method is smaller than the old method and the standard error of β_1 of our method is smaller than the old method. The average of estimated standard deviation of our method is reasonably close to the empirical standard deviation, and the coverage probabilities of the 95% confidence interval are close to 95%.

4. Discussion and Conclusion

This paper investigates the quantile residual life regression based on semi-competing risk data. Because T is dependently censored by D, we can't make inference on T without extra assumption. Therefore, we assume that (T,D) follow an Archimedean copula. To check the copula assumption, we can apply the checking approach by Hsieh, Wang, and Ding (2008). Then, we apply the inverse probability weight technique to constructing an estimating equation of $\beta_{\xi|t_0}, S_n(\beta_{\xi|t_0}) = 0$. But, $S_n(\beta_{\xi|t_0})$ may not be continuous in $\beta_{\xi|t_0}$. Thus, we apply the generalized solution approach to overcoming this problem. From the simulation studies, it shows the performance of the proposed method is good. When the covariates are continuous, we can group it as categorical variables or handle it with smoothing technique, which is treated as a future work.

References

1. Gelfand, A. E. and Kottas, A. (2003). Bayesian Semiparametric Regression for Median Residual Life. *Scandinavian Journal of Statistics*, **30**, 651-665.
2. Hsieh, J. J., Ding, A. A., Wang, W., and Chi, Y. L. (2013). Quantile regression based on semicompeting risks data. *Open Journal of Statistics*, **3**, 12-26.
3. Hsieh, J. J. and Hsiao, M. F. (2015). Quantile Regression Based on A Weighted Approach under Semi-Competing Risks Data. *Journal of Statistical Computatiion and Simulation*, **85**, 27932807.
4. Hsieh, J. J. and Wang, H. R. (2017). Quantile regression based on counting process approach under semi-competing risks data. *Accepted by Annals of the Institute of Statistical Mathematics*.
5. Hsieh, J. J., Wang, W and Ding, A. A. (2008). Regression analysis based on semi-competing risks data. *Journal of Royal Statistic Society, Series B*, **70**, 3-20.
6. Jeong, J. H. Jung, S. H. and Costantino, J. (2008). Nonparametric Inference on Median Residual Life Function. *Biometrics*, **64**, 157-163.

7. Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer.
8. Jung, S. H. and Jeong, J. H. and Bandos, H. (2009). Regression on Quantile Residual Life. *Biometrics*, **65**, 1203-1212.
9. Kaplan, E. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481.
10. Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.
11. Lakhal, L., Rivest, L. P. and Abdous, B. (2008). Estimating Survival and Association in a Semicompeting Risks Model. *Biometrics*, **64**, 180-188.
12. Ma, Y. and Yin, G. (2010). Semiparametric median residual life model and inference. *The Canadian Journal of Statistics*, **34**, 665-679.
13. Peng, L. and Fine, J. P. (2009). Competing Risks Quantile Regression. *Journal of the American Statistical Association*, **104**, 1440-1453.
14. Peng, L. and Huang, Y. (2008). Survival Analysis Based on Quantile Regression Models. *Journal of the American Statistical Association*, **103**, 637-649.
15. Portnoy, S. (2003). Censored Regression Quantiles. *Journal of the American Statistical Association*, **98**, 1001-1012.
16. Powell, J. (1984). Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics*, **25**, 303-325.
17. Powell, J. (1986). Censored Regression Quantiles. *Journal of Econometrics*, **32**, 143-155.
18. Ying, Z., Jung, S., and Wei, L. (1995). Survival Analysis With Median Regression Models. *Journal of the American Statistical Association*, **90**, 178-184.

Table 1: Estimations of β_0 and β_1 with quantile $\xi = 0.5$ under Clayton copula.

			Old method			Our method				
β	τ	t_0	Bias	EmpSd	MSE	Bias	EmpSd	MSE	AveSd	CP%
β_0	0.3	0	0.1083	0.2037	0.0532	0.0429	0.1851	0.0361	0.1816	0.9200
		0.07	0.1374	0.2248	0.0694	0.0583	0.2103	0.0476	0.2198	0.9475
		0.17	0.1879	0.2636	0.1048	0.0756	0.2540	0.0702	0.2688	0.9375
		0.35	0.2407	0.3645	0.1908	0.0773	0.3183	0.1073	0.3222	0.9250
	0.5	0	0.1527	0.1739	0.0536	0.0299	0.1628	0.0274	0.1787	0.9275
		0.07	0.1800	0.2035	0.0738	0.0266	0.1890	0.0364	0.2038	0.9475
		0.17	0.1812	0.2288	0.0852	0.0728	0.2347	0.0604	0.2488	0.9475
		0.35	0.1697	0.3068	0.1230	0.0834	0.3141	0.1056	0.3297	0.9575
	0.7	0	0.1300	0.1649	0.0441	0.0379	0.1664	0.0291	0.1620	0.9100
		0.07	0.1400	0.1723	0.0493	0.0229	0.1712	0.0298	0.1832	0.9525
		0.17	0.0952	0.2031	0.0503	0.0493	0.2083	0.0458	0.2177	0.9300
		0.35	0.0504	0.2902	0.0867	0.0756	0.2977	0.0943	0.3010	0.9250
β_1	0.3	0	0.0167	0.2766	0.0768	0.0026	0.2420	0.0586	0.2527	0.9525
		0.07	0.0187	0.3425	0.1177	0.0011	0.2950	0.0870	0.3051	0.9575
		0.17	0.0108	0.4207	0.1771	0.0117	0.3841	0.1477	0.3901	0.9500
		0.35	0.0481	0.5963	0.3579	0.0493	0.4311	0.1883	0.4954	0.9625
	0.5	0	0.0022	0.2622	0.0688	0.0038	0.2224	0.0495	0.2460	0.9650
		0.07	0.0101	0.2968	0.0882	0.0179	0.2691	0.0728	0.2913	0.9525
		0.17	0.0355	0.3588	0.1300	0.0162	0.3366	0.1136	0.3566	0.9525
		0.35	0.0670	0.5264	0.2816	0.0778	0.4369	0.1969	0.4926	0.9750
	0.7	0	0.0107	0.2449	0.0601	0.0126	0.2291	0.0527	0.2285	0.9325
		0.07	0.0153	0.2486	0.0620	0.0039	0.2467	0.0609	0.2639	0.9550
		0.17	0.0551	0.3216	0.1064	0.0345	0.3067	0.0953	0.3229	0.9425
		0.35	0.0674	0.4927	0.2473	0.0771	0.4400	0.1995	0.4713	0.9550

Note that: The sample size is 200 and replications are 400.

A study on the relationship between life cycle of enterprises and affecting variables in Korea

Younyoung Park
Statistics Korea, Daejeon, Korea

Abstract

Information on the life cycle of enterprise provides very useful tool to economic entities participating in industrial activities. The purpose of this paper is to derive the measurable indicators for the life cycle and affecting variables of the enterprises and to analyze the relationship. To increase the efficiency of the research, the scope of the study could be supposed to limit the manufacturing industry, wholesale and retail, accommodation and restaurant business. The main tasks are to address the indicators that could be used to diagnose the business life cycle by industry and scale, to examine the affecting variables, to analyze relationship between them and to seek the findings and policy implications. As a data source to identify the life cycle, business demography statistics which was developed in 2012 from statistics Korea was used. As for the variable indicators, it used the economically active population survey, whole country establishment survey, GRDP, national balance sheet, and trade characteristics statistics considering the measurability by industrial sector. Dataset from each statistics has been reprocessed into 10 indicators. The relationship analysis has been intensively conducted in two years of 2012 and 2015 because of data availability and so forth. The first finding is that the birth and death of enterprises are lower as the number of employees increases. Secondly, industries intensively combined with land, labor and capital in some degree, such as manufacturing, are less likely to be born and disappeared than service industries that are easier to enter the market. The third finding is that manufacturing sector could accumulate the net capital stocks due to the high proportion of imports and exports and value added. These could be regarded as the result of concentrating on the favorable tax system and policies of the government. On the other hand, the hospitality which is one of representative service industry has not been able to accumulate a large amount of net capital stock. Therefore fundamental of this sector has not been improved for a long time. Fourthly, although manufacturing sector is dominating in the portion of exports and imports and capital stock, the proportion of employment is relatively low. On the other hand, the service sector is high proportion in employment and fundamental of industry is still questionable. As for the policy implications, it seems that there is a need to strengthen competitiveness for service industry, which has a high proportion of employment. In the case of manufacturing, the sound

fundamental has been created for decades, but the service sector still shows low levels of exports and value added. Considering the importance of hospitality sector in tourism which is high value added industry, it is urgent to secure international competitiveness. Also, it could be play a pivotal role to create employment. For this reason, supporting policies that are similar to the manufacturing sector should be prepared. It is meaningful that this study examined the various indicators around life cycles of enterprises by subdividing them into industries. However, it could be alleged that in-depth analysis of the variables has not been achieved due to lack of detailed statistics by industry as a limitation. It is hoped that more detailed statistics could be produced in the near future.

Keywords
Enterprise; Life cycle; Birth; Death; Variable indicator; Policy Implications

1. Introduction

Information on the life cycle of an enterprise provides very useful tool to economic entities participating in industrial activities. In addition to analysis of birth and death, which best represents the life cycle; research on variables could be an important touchstone for enhancing industrial competitiveness. The purpose of this paper is to derive demographic indicator and the associate variables using available data and to analyze the association. This study is supposed to presuppose two conditions. The scope of research is limited to the manufacturing, wholesale and retail trade, and accommodation and restaurant business. The second one is that indicator could be obtained from existing approval statistics and could be reprocessed for the study purpose. This paper starts with raising three research questions to be addressed.

Q1. Whether could be justifiable to diagnose the indicators for life cycle of an enterprise by industry and scale?

Q2. Whether could be examined the affecting variables to life cycle of enterprises?

Q3. Whether could be possible to investigate the association between them?

To address the raised questions, it will begin with data searching and create indicators of both sides and will attempt to analyze the relationship.

2. Diagnosis of life cycle

1) Data source and calculation

It has used business demography statistics which was developed in 2012. To calculate the indicator, basic dataset was prepared using time serious table of active, birth, death from it. Firstly, three industrial dataset was separated from basic dataset. Secondly, these dataset was separated into two groups by number of employees. One is one person enterprise. The other is 2 and more

personnel enterprises. At last, it created 4 formula indicators such as birth and death by two types of enterprises. Active enterprise was denominator. Birth and death were used for numerator.

2) Life cycle analysis by employment size

In 2016, the birth rate of one person enterprises was 21.8% in accommodation and restaurant, 17.6% in wholesale and retail trade and 13.6% in manufacturing industry. The birth rate of enterprises with more than two personnel was somewhat low. It has shown 14.1% in the accommodation and restaurant, 7.3% in wholesale and retail trade and 6% in manufacturing.

<Table 1: 1 person enterprise birth rate>

	2012	2013	2014	2015	2016
Manufacturing	0.135	0.136	0.144	0.137	0.136
Wholesale and retail trade	0.183	0.176	0.183	0.172	0.176
Accommodation and restaurant	0.210	0.204	0.226	0.215	0.218

<Table 2: 2 and more personnel enterprise birth rate>

	2012	2013	2014	2015	2016
Manufacturing	0.081	0.076	0.075	0.065	0.060
Wholesale and retail trade	0.085	0.080	0.083	0.074	0.073
Accommodation and restaurant	0.126	0.136	0.156	0.142	0.141

In 2015, the death rate of one person enterprise was 19.4% for accommodation and restaurant, 15.4% for wholesale and retail trade, 12.2% for manufacturing industry. The death rate for two and more personnel enterprises was lower than one person enterprises. By industry, accommodation and restaurant business was 7.7%, 4.5% for wholesale and retail and 4.2% in the manufacturing.

<Table 3: 1 person enterprise death rate>

	2012	2013	2014	2015
Manufacturing	0.133	0.125	0.126	0.122
Wholesale and retail trade	0.178	0.168	0.163	0.154
Accommodation and restaurant	0.223	0.199	0.198	0.194

<Table 4: 2 and more personnel enterprise death rate>

	2012	2013	2014	2015
Manufacturing	0.047	0.043	0.040	0.042
Wholesale and retail trade	0.053	0.056	0.044	0.045
Accommodation and restaurant	0.079	0.078	0.072	0.077

3. Exploration of affecting variables

1) Introduction of data source and processing method

Considering the measurability and availability, it has searched suitable statistics such as Economically Active Population Survey, Whole country establishment Survey, Gross Regional Product (GRDP), National Balance Sheet and Trade Characteristics by Company. From these statistics, appropriate macro data were extracted for the analysis as follows; employment and

workers by industry, value added by industry, net capital stock by industry, export and import data by industry. Lastly, comparable affecting indicators such as the employment, working personnel in establishment, the value added, the net capital stock, the exports and the import proportion by specific industries were generated.

2) Variable analysis by industry

In general, the employment proportion by industry was steady and stable. In 2017, manufacturing was 16.9%, the wholesale and retail trade was 14.2%, and accommodation and restaurant was 8.6%.

<Table 5: Employment proportion of economically active population by industry>

	2012	2013	2014	2015	2016	2017
Manufacturing	0.167	0.168	0.170	0.174	0.171	0.169
Wholesale and retail trade	0.149	0.146	0.148	0.146	0.142	0.142
Accommodation and restaurant	0.077	0.079	0.082	0.084	0.087	0.086

As for proportion of establishment working personnel by industry in 2017, manufacturing was highest in 20.7%, 12.4% of wholesalers and retail trade, and 6.4% of accommodation and restaurants respectively.

<Table 6: Proportion of establishment working personnel by industry>

	2012	2013	2014	2015	2016	2017
Manufacturing	0.222	0.223	0.222	0.217	0.210	0.207
Wholesale and retail trade	0.120	0.122	0.122	0.123	0.124	0.124
Accommodation and restaurant	0.064	0.066	0.067	0.065	0.065	0.064

Likewise, proportion of value added from GRDP in 2016 has shown that manufacturing accounted for 29.5%, 8.4% in the wholesale and retail trade, and 2.8% in the accommodation and restaurant.

<Table 7: Proportion of value Added by Industry>

	2012	2013	2014	2015	2016
Manufacturing	0.310	0.310	0.301	0.297	0.295
Wholesale and retail trade	0.093	0.091	0.087	0.084	0.084
Accommodation and restaurant	0.026	0.026	0.026	0.027	0.028

Regarding to proportion of capital stock by industry in 2016, manufacturing was 19.7%, 2.2% in wholesale and retail trade, and 1.2% in accommodation and restaurant.

<Table 8: Proportion of capital stock by industry>

	2012	2013	2014	2015	2016 p)
Manufacturing	0.189	0.192	0.194	0.196	0.197
Wholesale and retail trade	0.020	0.021	0.021	0.022	0.022
Accommodation and restaurant	0.010	0.011	0.011	0.011	0.012

In terms of exports, the proportion by industry in 2017 was 84.3% for manufacturing, 12.1% for wholesale and retail trade, and 0.0% for accommodation and restaurants.

<Table 9: Proportion of exports by Industry>

	2012	2013	2014	2015	2016	2017
Manufacturing	0.846	0.850	0.846	0.851	0.847	0.843
Wholesale and retail trade	0.122	0.118	0.118	0.111	0.111	0.121
Accommodation and restaurant	0.000	0.000	0.000	0.000	0.000	0.000

The share of imports in the manufacturing sector for 2017 was 65.4%, wholesale and retail trade was 23.9%, and accommodation and restaurant was 0.1%.

<Table 10: Proportion of Imports by industry>

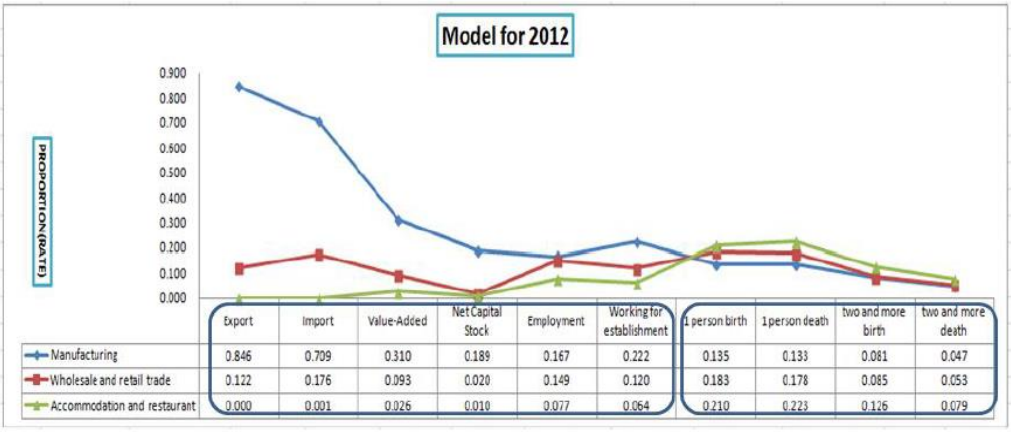
	2012	2013	2014	2015	2016	2017
Manufacturing	0.709	0.693	0.685	0.649	0.643	0.654
Wholesale and retail trade	0.176	0.183	0.192	0.231	0.251	0.239
Accommodation and restaurant	0.001	0.001	0.001	0.001	0.001	0.001

4. Relationship Analysis

1) Relationship modeling

This study is supposed to assume that finding indicators are directly or indirectly related to the business life cycle of enterprises. Analysis model has been set up for 2012 and 2015 because those years are suitable to compare both sides considering data availability, time interval, additional study and so forth.

2) 2012 Model



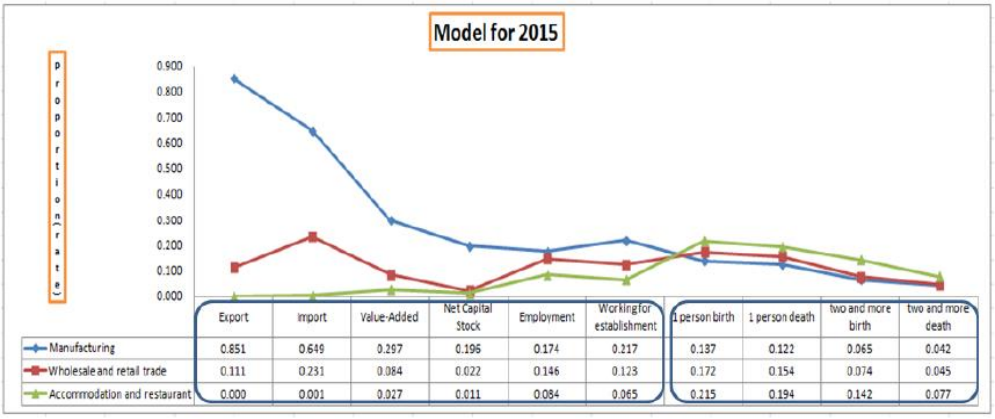
	Export	Import	Value-Added	Net Capital Stock	Employment	Working for establishment	1 person birth	1 person death	two and more birth	two and more death
Manufacturing	0.846	0.709	0.310	0.189	0.167	0.222	0.135	0.133	0.081	0.047
Wholesale and retail trade	0.122	0.176	0.093	0.020	0.149	0.120	0.183	0.178	0.085	0.053
Accommodation and restaurant	0.000	0.001	0.026	0.010	0.077	0.064	0.210	0.223	0.126	0.079

In manufacturing sector, exports accounted for 84.6% and imports accounted for 70.9%. The value added ratio was 31%, and the net capital stock was the highest at 18.9%. Owing to the stable industrial structure, the ratio of employment accounted for 16.7% and proportion of establishment employee account for 22.2% respectively. As a result of the relatively high need for land, labor, and capital combination, birth rate of two or more personnel enterprises was 8.1% and the death rate was 4.7%. The birth rate of one person enterprise was 13.5% and the death rate was 13.3%. In brief, big enterprise was slightly lower than small one in birth and death rate. It could be explained that barriers to entry are somewhat higher because big one needs more production factors such as labor, capital, skill and so forth.

In terms of wholesale and retail trade, exports account for 12.2% and imports account for 17.6%. The proportion of exports and imports of accommodation and restaurant businesses was negligible level. Value added was also low in both industries (9.3%, 2.6%), and net capital stocks (2% and 1%, respectively). the ratio of employment accounted for 14.9% and the

proportion of establishment working personnel account for 12.0% respectively whereas accommodation and restaurant were 7.7% and 6.4%, respectively. Demographically, birth and death rate were much higher than manufacturing. Especially, birth and death rate of 1 person enterprise in accommodation and restaurant showed 21.0% and 22.3% respectively.

3) 2015 Model



	Export	Import	Value-Added	Net Capital Stock	Employment	Working for establishment	1 person birth	1 person death	two and more birth	two and more death
Manufacturing	0.851	0.649	0.297	0.196	0.174	0.217	0.137	0.122	0.065	0.042
Wholesale and retail trade	0.111	0.231	0.084	0.022	0.146	0.123	0.172	0.154	0.074	0.045
Accommodation and restaurant	0.000	0.001	0.027	0.011	0.084	0.065	0.215	0.194	0.142	0.077

In the case of manufacturing, exports accounted for 85.1% and imports accounted for 64.9%. Although exports have increased and imports have decreased compared to 2012, the proportion of exports was extremely high. The value-added portion was 29.7%, while the net capital stock was 19.6%. The ratio of employment accounted for 17.4% and proportion of establishment employee account for 21.7% respectively. The birth rate was 6.5% and the death rate was 4.2% for the 2 and more personnel enterprises whereas birth rate and death rate was 13.7% , 12.3% for 1 person one.

In the wholesale and retail trade, exports accounted for 11.1% of total and imports accounted for 23.1%. The proportion of exports and import of accommodation and restaurant businesses was negligible. The added value of the two industries was low at 8.4% and 2.7%, respectively, and net capital stocks have accumulated 2.2% and 1.1% respectively. The ratio of employment and proportion of establishment working personnel accounted for 14.6%, 12.3% respectively for wholesale and retail trade and 8.4% and 6.5% for accommodation and restaurants. Next, the birth and death rate for more than two employees in the wholesale and retail trade was 7.4% and 4.5%, respectively. For the 1 person enterprises, birth and death were 17.2% and 15%, respectively. In accommodation and restaurant, the birth and death rate were 14.2% and 7.7% for more than 2 personnel enterprises and 21.5% and 19.4% for one person.

5. Conclusion

The first finding is that the births and deaths are lower as the number of employees increases. Secondly, industries combined with land, labor, skill and capital in some degree, such as manufacturing, are less likely to be born and disappeared than service industries that are easier to enter the market. The third finding is that manufacturing sector could accumulate the net capital stocks due to the high proportion of imports and exports and value added with the help of concentrating on the tax system and favorable policies of the government. On the other hand, the hospitality industry which is belong to service sector, has not been able to accumulate a large amount of net capital stock. Therefore basic fundamental of this sector has not been improved for a long time. Fourthly, although manufacturing sector is dominating in the portion of exports and imports and capital stock, the proportion of employment is relatively low. On the other hand, the service sector is high in employment even though industrial fundamental is not enough stable.

As for the policy implications, it seems that there is a need to strengthen competitiveness for the representative service industry, which has a high proportion of employment. In the case of manufacturing, the industrial base has created solid infrastructure for decades, but the service industry still shows low levels of exports and value added. Considering the importance of hospitality sector in terms of supporting tourism which is high value added industry, it is urgent to secure international competitiveness. Also, it could be play a pivotal role to create employment. For this reason, supporting policies that are similar to the manufacturing level should be provided. It is meaningful that this study examined the variables affecting the business demographic events by subdividing them into industries. However, there could be a limitation for in-depth analysis because of data lacking in detailed statistics by industry. It is hoped that more detailed statistics could be produced in the near future.

References

1. Business Demography Statistics, viewed in Jan 2019, http://kosis.kr/
2. Economically Active Population Survey, viewed in Jan 2019, http://kosis.kr/
3. Whole country establishment Survey, viewed in Jan 2019, http://kosis.kr/
4. Gross Regional Product (GRDP), viewed in Jan 2019, http://kosis.kr/
5. National Balance Sheet, viewed in Jan 2019, http://kosis.kr/
6. Trade Characteristics by Company, viewed in Jan 2019, http://kosis.kr/

Married women's experience of domestic violence in Malawi: New evidence from a cluster and multinomial logistic regression analysis

Lana Clara Chikhungu[1], Mark Amos[2], Ngianga II Kandala[2], Saseendran Palikadavath[2]

[1]School of Area Studies, History, Politics and Literature, Faculty Humanities and Social Sciences, University of Portsmouth.

[2]School of Health Science and Social Work, Faculty of Sciences, University of Portsmouth. Southampton, UK

Abstract

Violence against women is a global issue with estimates indicating that 35% of all women world-wide have experienced either physical and/or sexual intimate partner violence or non-partner violence in their life time. In Malawi 42% of ever-married women have experienced some form of violence perpetrated by their current or most recent spouse. A number of studies have investigated intimate partner violence in Malawi within the context of HIV/AIDS and girls sexual abuse and a few report on the role of socio-cultural factors in influencing gender based violence. No study has used cluster analysis to systematically analyse the groups affected domestic violence across different dimensions of abuse. Using the 2015 Malawi Demographic and Health Survey data, we employed cluster analysis and multinomial logistic regression to analyse the distribution of different forms of abuse amongst married women in Malawi and the key attributes associated with each form of abuse. Correlates of domestic violence significantly differ by levels of abuse and are distributed as follows; controlling behaviour (11.8%), general controlling behaviour (27.1%), moderate physical and emotional abuse (27.2%) and the high and complete abuse (8.5%). Alcohol consumption, ethnicity and women working status were significantly associated with all four levels of abuse but age and religion were only associated with controlling behaviour and generalised controlling behaviour. The strength of association between *husband's alcohol consumption, woman's working status and marriage type* and domestic violence increased by level of abuse. On each of these factors, the odds of experiencing violence were lowest in the controlling behaviour group and highest in the high physical and emotional abuse group. Policies and programmes that are designed to tackle violence against married women in Malawi should incorporate strategies that discourage excessive drinking, promote messages that women can be bread winners and discourage polygamous marriage.

1. Introduction

The most recent Malawi Demographic and Health Survey report estimates that 42% of ever-married women have experienced some form of physical, sexual or emotional violence perpetrated by their current or most recent spouse(NSO-Malawi & DHS-Program, 2017). A number of studies have investigated intimate partner violence in Malawi within the context of HIV/AIDS and girls sexual abuse and a few have studied the role of socio-cultural factors in influencing gender based violence(Mellish, Settergren, & Sapuwa, 2015). No study has considered multiple ways in which abuse is perpetrated, nor attempted to analyse the patterns of abuse holistically. Using cluster analysis and multinomial logistic regression, this study identifies the distribution of different forms of abuse amongst married women in Malawi and the key attributes associated with each form of abuse.

2. Methods

The data are drawn from the 2015 Malawi Demographic and Health Survey (2015 MDHS) carried out under the DHS programme. The woman's questionnaire collected information from 24,562 women out of 25,146 women aged 15 to 49 years that were eligible for the interview representing a 98% response rate. One third of the sampled households received domestic violence questions (6,379 household). Further details of study design and data collection are reported on the National Statistical Office of Malawi website: http://www.nsomalawi.mw/.

Explanatory Variables

The choice of explanatory variables is guided by findings from previous literature and data availability. These comprised of demographic variables: age of the woman, age of the woman at first sex, age at first cohabitation, geographic/location variables: urban/rural residence, region, religion, ethnicity and socio-economic variables: household wealth status, whether husband takes alcohol or not, woman's education level, whether the woman is currently working or not and type of marriage (polygamous or monogamous)).

Statistical Analysis

Cluster analysis is used to extract meaningful groupings of the experience of domestic violence from the MDHS. 18 binary variables are used in forming clusters, each taking the form of a 0/1 indicator variable where the variable takes the value 1 if the woman reports experiencing that form of domestic abuse and zero otherwise. We include three domains of domestic abuse: physical (formed using reports of the woman being: pushed, slapped, punched, kicked, strangled, threatened with a weapon, limb twisted), sexual (physically forced into sex, coerced into sex, forced to perform a sex act) and controlling behaviour (husband exhibits jealous behaviour, accused of being unfaithful, needs permission to see friends, needs permission to see family,

needs to justify whereabouts, humiliated, threatened, insulted). Cluster analysis is performed using the cluster function in Stata 13.0 for Windows (Statacorp, Tx, USA). Clustering is performed using hierarchical cluster analysis based on Ward's distance. To decide on the number of clusters, we use a combination of indices which indicate optimal clustering pattern (Calinski–Harabasz pseudo-F index (Calinski & Harabasz, 1974) and the Duda–Hart index (Duda & Hart, 1973) and dendritic analysis.

Regression analysis

Once the final cluster profile was selected, membership of a particular cluster was then used as the dependent variable in a multinomial logistic regression. A multinomial logistic regression model is appropriate because instead of running four separate regression models, separate logistic regression model for each indicator variable are estimated simultaneously. The model takes the form of equation 1

$$\ln\left(\frac{\pi^{s}}{\pi^{s=0}}\right) = \boldsymbol{\beta}'\boldsymbol{x}$$
$$s = 0 \ldots S$$

In equation 1 the probability of belonging to cluster s is denoted as π^{s} where there are S clusters. The probability of belonging to cluster S is modelled as a logit function of a combination of a vector of coefficients $\boldsymbol{\beta}$ and associated predictor variables in the vector x. $S-1$ logits are estimated and the baseline cluster $s=0$ is omitted to identify the model. (Anderson & Rutkowski, 2008).

3. Results

Cluster analysis

Five distinct clusters of abuse were extracted based on a combination of fit statistics and dendritic analysis.

The first cluster is termed no abuse (NA), and contains women who reported no experience of abuse on any of the response variables. This cluster comprises 25.39% of the sample.

The second cluster extracted is characterised by Controlling Behaviour (CB). All women within this cluster report that they have experienced their spouse demanding knowledge of their whereabouts, but no other form of abuse is present. This cluster comprises 541 women, or 11.79% of the sample.

The third cluster comprises general controlling behaviour (GCB). Women in this cluster experience high rates of jealousy from their husbands (84%) and their husbands demand to know whereabouts (77%), as well as lower level of accusations of being unfaithful (30%), isolation from friends (22%) and family (17%). All indicators of physical abuse are low (below 5%). This is the second largest cluster in the sample comprising 1245 women (27.13%).

Cluster 4 comprises moderate physical and emotional abuse (MPE). Emotional abuse is common in this cluster, with high rates of jealous behaviour (60%) and control of whereabouts (67%). A substantial minority of women in this cluster have experienced physical abuse largely in the form of being slapped (39%) and being forced to perform a sex act (37%). This cluster is the largest in the sample with 1250 women (27.24% of the sample).

Cluster 5 represents the highest overall level of abuse, and is termed the high and complete abuse (HCA). Nearly all women in this cluster have experienced some form of physical abuse, with particular high rates of being slapped (90%), punched (65%) and kicked (60%). Women in this cluster also report high rates of forced sex (65%). Controlling behaviour and emotional abuse is also common in this cluster, particularly jealous behaviour (90%), knowledge about whereabouts (90%) being humiliated (55%), threatened (65%) or insulted (75%). This cluster comprises 288 women and is 8.46% of the sample.

Multinomial logistic regression results

Results of the multinomial regression are summarised in Table 1. The odds of a woman encountering a particular profile of domestic violence were consistently associated with husband's alcohol consumption, women's working status and ethnicity. For husband's alcohol consumption, women's working status and polygamous marriage, the odds were smaller for controlling behaviour type of abuse and higher in the physical and complete abuse categories. Age, religion, education level and wealth status were only significant in some types/levels of abuse. Where these variables were significant, younger women were more likely to encounter violence than the older women, Muslim women were more likely to experience that type of violence compared to Christian women, women with primary education were more likely to encounter violence than women with no education and poorer women had higher odds of suffering from violence than the poorest.

Table 1 Results of multinomial logistic regression results on type and severity of domestic violence, 2015 MDHS

Variable	Controlling behaviour	Generalised controlling behaviour	Moderate physical and emotional abuse	High and complete abuse
Husband's alcohol consumption	Yes (1.80)	Yes (2.00)	Yes (3.30)	Yes (6.91)
Women's working status	Yes (1.25)	Yes (1.31)	Yes (1.36)	Yes (1.40)
Polygamous marriage	No (1.07)	Yes (1.58)	Yes (1.76)	Yes (2.77)
Ethnicity	Yes	Yes	Yes	Yes
Wealth status	Yes	Yes	No	Yes
Education level	No	No	Yes	Yes
Religion	Yes	Yes	No	No
Age	Yes	Yes	No	No
Age at first sex	No	No	No	Yes

4. Conclusion and Policy Recommendations

Alcohol consumption, ethnicity and women's working status were consistently associated with all levels of abuse. Age of the married women and religion were only significantly associated with controlling behaviour and generalised controlling behaviour. Alcohol consumption, women's working status and polygamy showed a dose response relationship with domestic violence suggesting that these factors are key to tackling domestic violence in Malawi. The Malawi Government and development partners should consider designing policies and programmes that tackle excessive beer drinking, promotion of the acceptance that women can be bread winners and discouraging the practice of polygamy to curb violence against women in Malawi.

References

1. Anderson, C. J., & Rutkowski, L. (2008). Multimomial Logistic Regression. In J. Osborne (Ed.), Best practices in Quantitative Methods. Online: Sage Publications.
2. Calinski, R. B., & Harabasz, J. A. (1974). A dendrite method for cluster analysis. Communications in Statistics-Theory and Methods, 3, 1-27.
3. Duda, R. O., & Hart, P. E. (1973). Pattern Classification and Scene Analysis. New York: Wiley.
4. Mellish, M., Settergren, S., & Sapuwa, H. (2015). Gender-based Violence in Malawi: A Literature Review to Inform the National Response. Washington DC: Futures Group, Health Policy Project.
5. NSO-Malawi, & DHS-Program. (2017). Malawi Demographic and Health Survey Report 2015-2016. Zomba, Malawi and Rockville, Maryland USA: NSO and ICF.

Fuzzy rule base method for forecasting time series data

Nur Fazliana Rahim[1,2], Mahmod Othman[2], Rajalingtam Sokkalingam[2]

[1]Centre for Pre-University Studies, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

[2]Fundamental and Applied Sciences Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia

Abstract

In generating fuzzy rule of forecasting, Weighted Subsethood-Based Algorithm (WSBA) were used to develops Foreign Exchange Rate (FER) forecasting method. In the use of Fuzzy Time Series to develop fuzzy rules, Fuzzy Rule Based Systems (FRBS) concept was implemented. The intention of the recommended method is to improve the efficiency of time series forecasting which offer higher predicting accuracy. In order to validate this method, 5 years' data of FER and three currency pairs was used as testing data sets, which are Malaysian Ringgit (MYR), Japanese Yen (JPY), South Korea Won (KRW) and Singapore Dollar (SGP). The forecasting precision of this method was compared with the prior methods. In this paper, the results proved that this method can minimize forecasting error and so that increase the accuracy of FER forecasting values. The outcomes of this paper could be used as another choice of method to obtain the fuzzy rules to get a superior forecast values of FER.

Keywords

Forecasting Foreign Exchange Rate; Fuzzy Time Series; Weighted Subsethood Based Algorithm

1. Introduction

In dealing with Fuzzy Time Series (FTS) forecasting, exchange rate is found to be a fascinating subject. Taken into consideration, exchange rate takes a crucial part in worldwide trade, handling of business risk and the country economic condition. (Korol, 2014). Several factors can affect forecasting the exchange rate, while affecting currency ratings such as inflation, interest rate, national debt, employment data, political stability and economic performance and etc. (Patel et al., 2014). Currently a rare situation in currency market when central bank intervenes and when doing so it is often successful. In 1997 when the currency depreciates causing Malaysia facing a rough time due to inflation (Leu et al., 2009). This is why forecasting the Foreign Exchange Rate (FER) is vital and by using appropriate forecasting model that can validly forecast the

FER will surely be a great assist for the government in preparing for the instability of the exchange rate as such**.** Furthermore, as it's a major player in forecasting of exchange rates, Bank Negara Malaysia (BNM) will also obtain the benefit of this research. In the near future, the use of forecasting exchange rates will surely be made better.

Based on the prior study, a new FTS model was proposed by Yu (2005) during the creation of FTS model to modify the intervals length. As shown by the outcome, appropriate fuzzy relationships can be improved the forecasting values. Another study by Arumugam et al. (2013) was conducted in order to predict Taiwan export trade using FTS model and ARIMA model. The final results indicate that, FTS model beat ARIMA model with smallest average error and its ability to forecast the Taiwan export trade. A viable practice to forecast correctly and successfully future exchange were needed by the government in dealing with foreign exchange rate as it can affect the country currency. Even though lots of forecasting method were used to predict the foreign exchange rate, still there are no way to show which are the most reliable forecasting method. As stated by Applanaidu et al. (2011), the selection methods should take into account several aspects, such as statistical data, financing, level of accuracy and its significance. Nonetheless, these models are quite expensive, need a great degree of expertise and numerous types of data which may not always be obtainable.

Fuzzy rule-based systems (FRBS) deliver impulsive technique of reasoning based on linguistic models (Dubois and Prade, 2001). Often reasoning based on fuzzy models contribute an option to handle all kinds of unspecific data, which presented the way people think and make decisions. (Rasmani and Shen, 2006). In fuzzy forecasting, the determination of fuzzy rules is one of the aspect to reflect on. The FTS rises the precision of the results made in predicting situations that involve subjective, ambiguous and inaccurate information by implementing these rules. So, in conjunction with this research, Weighted Subsethood-Based Algorithm (WSBA) is suitable for generating fuzzy rules for two reasons, which are easiness of the method and possible to produce higher degree of accuracy indirectly minimize rules compared to other methods (Chen and Tsai, 2008). The development of WSBA includes a modification of the Subsethood-Based Algorithm (SBA) and the use of fuzzy general rules and SBA values as weight.

2. Methodology

This section explained and discussed in detail four parts that have been carried out in the research. The detailed are as follows.

A. *Part 1:* Compilation and Processing Data

Foreign Exchange Rate (FER) data is collected by referring to the to the secondary data of monthly FER for five years' data from year 2010 to 2015.

The data was obtained from the Bank Negara Malaysia. In this part, the dataset of Foreign Exchange Rate (FER) data were separated into two subsets; FER-1 used for training and FER-2 used for testing. Each dataset comprises 30 cases. The FER data includes three characteristic; One month previous, Two months previous and Three months previous. There are three classification outcomes of the FER rank; Small, Intermediate and Large.

B. *Part 2:* Creation of Fuzzy Rules

To generate the required fuzzy model using WSBA was introduced in this phase. To create a system that is more readily understandable, WSBA uses a rule generation algorithm based on fuzzy general rules or addition of a Mamdani-type Fuzzy Rule Based Systems (FRBS). Five steps involve in this part as follows.

First, three subgroups obtained from the training dataset based on the classification outcomes. The measure of location, Q_k method used to classify the outcomes as follows

$$Q_k = L_k + \left(\frac{\frac{k}{4}N - F_k}{f_k} \right) C_k \tag{1}$$

where $k = 1, 2, 3, L_k =$ cumulative frequency before the Q_k class, $N =$ total number of observations, $F_k =$cumulative frequency before the Q_k class, $f_k =$frequency of the class where Q_k lies, and $C_k =$ size of the class where Q_k lies.

Second, the fuzzy membership function for FER dataset was constructed based on the measure of location; Q_1, Q_2 and Q_3 calculated respectively from equation (1). Then, the fuzzy partition was defined from the established fuzzy membership function. It was used to convert crisp values into fuzzy values.

Third, for each linguistic term, fuzzy subsethood values were calculated in each subgroup. This generated rules able to deal with classification problems. The fuzzy subsethood value of A taking into account $B, S(B, A)$ refers the degree levels to which A is subset of B [22], [23]:

$$S(B, A) = \frac{M(B, A)}{M(B)} = \frac{\sum_{x \in U} \nabla(\mu_B(x), \mu_A(x))}{\sum_{x \in U} \mu_B(x)} \tag{2}$$

where $S(B, A) \epsilon [0,1]$.

Fourth, using the subsethood values in Step 3, each linguistic term weights then were obtained. In this research, weighting is restricted between 0 to 1, which 0 is referring to the smallest weight (or less significance) and 1 the largest weight (or more significance). The subsethood here is mean to extend

and allow fuzzy sets to be related with different linguistic variables associated. The relative weight, W for linguistic term A_i, with regard to classification X is:

$$W\left(X,A_i\right)=\frac{S\left(X,A_i\right)}{\max\limits_{j=1,\ldots,l} S\left(X,A_j\right)} \tag{3}$$

where $w(x,A_i)\epsilon[0,1]$and $i = 1,2,\ldots,l$. Thus, the compound weight $T(A)$ and $T(B)$ of the weighted conjunction of linguistic terms related with it can be obtained as follows;

$$T(A)=\left(\frac{w_1}{w}(A_i)\ \nabla\ldots\nabla\frac{w_m}{w}(A_m)\right) \tag{4}$$

$$T(B)=\left(\frac{w_1}{w}(B_i)\ \Delta\ldots\Delta\frac{w_n}{w}(B_n)\right) \tag{5}$$

where A is the conditional attribute and the compound weight is $T(A)$, ∇ is the t-norm, $A_i, i = 1,2,\ldots, m$, are the linguistic terms of variable A which are conjunctively combined and W is the largest amongst the m associated weight, $W(X,A_i)$. Similarly, the compound weight $T(B)$ of the weighted disjunction of linguistic terms associated with variable B, where Δ is the t-conorm and $A_i, i = 1,2,\ldots, n$, are the linguistic terms of variable B, which are disjunctively combined.

Fifth, the WSBA then used the weighted conjunction $T(A)$ and weighted disjunction $T(B)$ to generate fuzzy rules.

C. Part 3: Finalize the Classification Output

The classification of FER rank can be performed when the rule set and the study over the three conditional attributes are acquired. Then, using rule set generated and the transform fuzzy values, the rules were calculated. the Min-Max Operator is used in this part.

D. Part 4: Rule set Testing for the Classification Tasks

The training dataset, FER-1 using rule set for classification of FER rank were then tested using the FER-2 dataset. The identified trend for each of the FER data were based on the classification of the rules. The FER distribution was started to forecast when each of the FER data had been classified by the rule and trend of forecasting.

3. Result

The analysis performed in this research were discusses in this section. Table I below depicted the forecast value obtained by using prior method and the current approach. The FER-1 dataset are tested using the rule set trained for classification of Foreign Exchange Rate where the FER-2 dataset is used for testing. The difference between forecasting methods was illustrates in Fig. 1.

TABLE I: COMPARISON OF CPO PRICE FORECASTING BETWEEN FUZZY APPROACH AND PREVIOUS METHOD

Cases	Foreign Exchange Rate (FER)	Forecast value (WSBA)	Forecast value (Prior Method)
1	3.7509	3.7538	3.7680
2	3.7481	3.7511	3.7635
3	3.7467	3.7500	3.7582
4	3.7523	3.7493	3.7411
5	3.7552	3.7549	3.7510
6	3.7495	3.7485	3.7463
7	3.7467	3.7466	3.7524
8	3.7453	3.7455	3.7470
9	3.7481	3.7465	3.7388
10	3.7495	3.7490	3.7532
11	3.7495	3.7495	3.7477
12	3.7580	3.7580	3.7526
13	3.7665	3.766	3.7678
14	3.7679	3.7665	3.7690
15	3.7679	3.7669	3.7645
16	3.7665	3.7656	3.7632
17	3.7679	3.7670	3.7645
18	3.7693	3.7688	3.7654
19	3.7693	3.7693	3.7645
20	3.7693	3.7693	3.7670
21	3.7665	3.7659	3.7646
22	3.7693	3.7680	3.7631
23	3.7509	3.7512	3.7540
24	3.7481	3.7467	3.7529
25	3.7467	3.7458	3.7422
26	3.7523	3.7537	3.7575
27	3.7552	3.7564	3.7593
28	3.7495	3.7487	3.7448
29	3.7467	3.7459	3.7508
30	3.7453	3.7440	3.7411

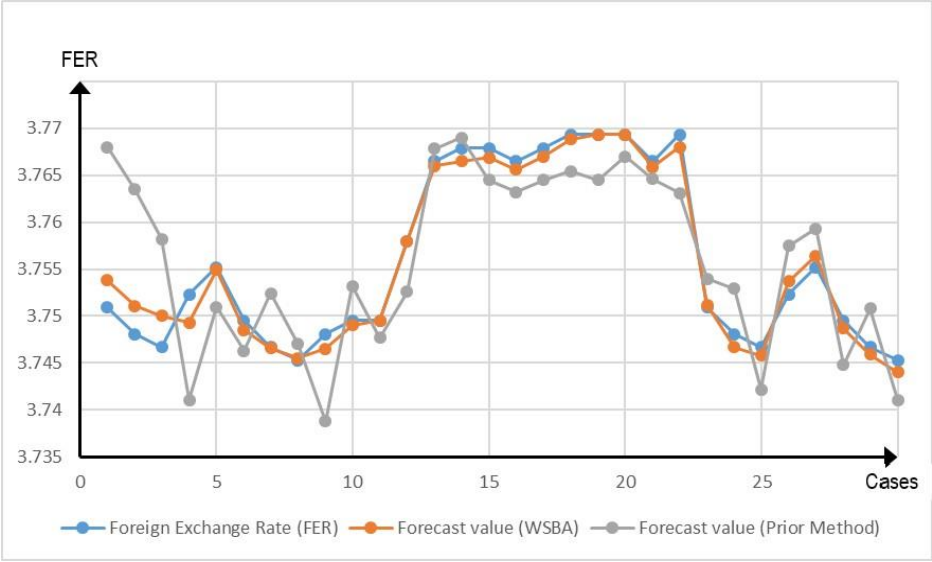


Fig. 1. The Differences between Foreign Exchange Rate (FER) and the Forecasting Method

Based on Fig. 1, the blue line denotes the Foreign Exchange Rate (FER). The orange line denotes the forecast value using Weighted Subsethood-Based Algorithm (WSBA) and the grey line denotes the prior forecasting method. From the plotted graph, it was shown that the forecasting value of FER using WSBA found to be nearly accurate to the Foreign Exchange Rate (FER). This proves that the forecast value using FTS with the proposed of WSBA is better compare to prior method and perform more precise forecasting of the FER.

Next, Table II below summarized the results of evaluation for each forecasting method.

TABLE II: EVALUATION OF METHOD

Forecasting Method	**MSE**	**RMSE**	**Percent Accuracy**
Proposed WSBA	0.17	1.89	98.1 %
Prior Method	0.37	9.52	90.5 %

From the table III, the result shows the value of MSE and RMSE for proposed Weighted SubsethoodBased Algorithm (WSBA) is lesser than the prior method, which is 0.17 and 1.89 respectively. Meanwhile, the result of MSE and RMSE for previous method is 0.37 and 9.52 respectively. Referring to the percent accuracy also shows that forecasting using WSBA more accurate compare to the prior method. Therefore, based on the above result, the proposed WSBA can be used as a method in generating rule prediction of time series forecasting.

4. Discussion and Conclusion

This research presents the field of data driven FRBS in forecasting of Foreign Exchange Rate (FER). It shows that this new approach gave so much advantages to strengthen the prior method. A preliminary data driven FRBS, Weighted Subsethood-Based Algorithm (WSBA) was developed using fuzzy subsethood values. Its provide easiness by generating default fuzzy rules without the need to use any threshold value. This is very valuable in forecasting area, which needs a system that is easy to understand by the people especially for forecaster. The FER data were process first by classifying the outcomes (FER rank). By using the rules generated, the FER forecasting was done and were compared with the prior method. These methods were evaluate using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). As mention in the result, the value of MSE and RMSE results for the proposed WSBA were lesser than the other method. It can be summarizing that WSBA produce more effective and reduce forecasting error compare to the prior method. Thus, the use of this method will lead to the formation of a systematic approach in forecasting application, which help reinforce decision made by alternative methods.

References

1. Korol, T. (2014). A fuzzy logic model for forecasting exchange rates. *Knowledge-Based Systems*, *67*, 49-60.
2. Patel, J. P., Patel, J. N., and Patel, R. A. (2014). Factors affecting currency exchange rate, economical formulas and prediction models. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 3(3)*, 53-56.
3. Leu, Y., Lee, C. P., and Jou, Y. Z. (2009). A distance-based fuzzy time series model for exchange rates forecasting. *Expert Systems with Applications*, *36*, 8107-8114.
4. Yu, H. K. (2005). A refined fuzzy time-series model for forecasting. *Physica A*, *346*, 657-681.
5. P. Arumugam, and V. Anithakumari (2013). Fuzzy time series method for forecasting Taiwan export data. *International Journal of Engineering Trends and Technology, 4(8),* 3342-3347.
6. Applanaidu, S. D., Mohamed Arshad, F., Shamsudin, M. N., and Abdel Hameed, A. A. (2011). An econometric analysis of the link between biodiesel demand and malaysian palm oil market. *Int. J. Bus. Manag., 6(2),* 35–45.
7. Dubois, D. and Prade, H. (2001). Handbook of fuzzy computation. *Fuzzy Sets Syst. Dep. Comput. Sci. Artif. Intell.*, *123(3)*, 397–398.

8. Rasmani, K. A. and Shen, Q. (2006). Data-driven fuzzy rule generation and its application for student academic performance evaluation. *Appl. Intell.*, *25(3)*, 305–319.

Chen, S. M. and Tsai, F. M. (2008). Generating fuzzy rules from training instances for fuzzy classification systems. *Expert Syst. Appl., 35(3),* 611–621.

Research on carbon emissions factors based on the EKC

Lili Chen[1], Qiguang Dong[2]
[1]School of Economics and Management, Tsinghua University, Beijing, 100084, China
[2]PLA Military Science Academy, Beijing, 100142, China

Abstract

The threat of climate change has been a major environmental challenge for the past two decades due to increased global warming. The rise in carbon dioxide emissions is considered to be one of the main causes of global warming and climate instability. Based on the panel data of the G20, this paper verifies the hypothesis of the EKC inverted U-shaped curve of the G20. The results of the study show that the level of urbanization in developing countries is positively related to carbon emissions. With the rapid development of urbanization, energy consumption will increase and carbon emissions will increase. The level of urbanization in developed countries is negatively correlated with carbon emissions. As urbanization improves environmental pollution control and energy use efficiency, it also reduces environmental pressure to a certain extent. The impact of openness of developing countries on carbon emissions is negative but not significant. Investment in environmental governance is conducive to the improvement of the exporting country's own environment, but this effect is not obvious. The openness of developed countries has a positive impact on carbon emissions. The increase in imports has caused some carbon emissions to be transferred to importing countries, resulting in an increase in carbon emissions in importing countries.

Keywords

Carbon emissions; Economic growth; Population; EKC; G20

1. Introduction

The G20 is made up of seven advanced economies, twelve major emerging economies and the European Union (EU). It accounts for two-thirds of the world's population, 60% of the world's GDP, and 90% of the global economy. Beyond that, the overall trade accounts for 80% of the global, including the most developed countries and emerging market economies in the world. With countries and regions which have international influence, the G20 is regarded as a new platform for dialogue between developing and developed countries. It is working to strengthen international communication and cooperation, achieve emission reduction targets and respond to global warming.

As more and more people flood into cities, energy consumption increases, lifestyle changes, and thus more emissions. Urbanization is a major process of

global development and has a profound impact on the relationship between people and the environment. In addition, population aging is also considered to be a major demographic trend of the 21st century. According to the United Nations standards, an ageing population of over 65 years old and above accounts for more than 7% of the total population. Globally, the proportion of people over 65 is increasing rapidly, which will influence the future policies of governments. Therefore, the impact of aging on carbon emissions policies cannot be ignored. The growth of import and export trade also have different effects on the carbon emissions of exporting and importing countries. On the one hand, exporting countries have obtained income through foreign trade, and have also gained access to foreign markets and participation in international transactions, which can optimize production in export-oriented countries according to environmental preferences attached to foreign consumer demand. Moreover, trade income can also be partially converted into environmental governance investment, which is conducive to the improvement of the exporting country's own environment. On the other hand, trade openness can lead trade participants to maintain or increase product attractiveness by relaxing environmental regulations, leading to environmental degradation.

Economic development is a long process that promises a high standard of living, but it can also lead to environmental degradation. Grossman and Krueger (1991) has shown that as the economy grows, the environment will gradually degrade, but when the economic development reaches a certain level, the environmental conditions will improve. This inverted U-shaped relationship is called EKC. Scholars have done a lot of research based on different countries and different time periods. A large number of literature studies based on EKC hypothesis are mainly divided into the following categories: First, there is a linear relationship between carbon emissions and economic growth (Azomahou et al., 2005). Second, carbon emissions and economic growth have an inverted U-shaped relationship (Lean et al., 2010; Al-Mulali et al., 2015). Third, carbon emissions and economic growth are in an N-type relationship (Shafik, 1994; Friedl and Getzner, 2003). Fourth, there is no relationship between carbon emissions and economic growth (Richmond and Kaufmann, 2006). The inconsistency of the research conclusions may be affected by the bias of the missing variables. Therefore, scholars have combined the factors such as trade openness, urbanization, and financial development to study the relationship between carbon emissions and economic growth (Ozturk and Acaravci, 2013).

2. Methodology and Data

2.1 Methodology

This paper attempts to re-verify the EKC curve hypothesis, examines the relationship between carbon emissions and economic development levels, and reveals the role of economic levels in environmental pollution. The expected result is in line with the inverted U-shaped relationship of the EKC curve. With fast-growing economies and per capita income, the carbon emissions will gradually increase, but when the per capita income reaches a certain point, the carbon emissions will gradually decrease. After drawing on and combining the methods of the predecessors and considering the particularity of the data, the expression of the EKC model is:

$$\ln CO_2 = \beta_0 + \alpha_1 * \ln GDP + \alpha_2 (\ln GDP)^2 + \mu \quad (1)$$

Among them, CO2represents carbon emissions. GDP is per capita GDP. β0, α_1, α_2 are estimated parameters. μ is an error term, and obeys a normal distribution. When $\alpha_1<0$, $\alpha_2>0$, U-form is presented; when $\alpha_1>0$, $\alpha_2<0$, it shows an inverted U-shape.

$$\ln CO_{2it} = \beta_0 + \beta_1 * \ln P_{it} + \alpha_1 * \ln GDP_{it} + \alpha_2 * (\ln GDP_{it})^2 + \beta_2 * \ln T_{it} + \beta_3 * \ln Age_{it} + \beta_4 * \ln Urb_{it} + \beta_5 * \ln Open_{it} + \eta_t + \delta_{it} \quad (2)$$

Further, factors such as technical level, aging, urbanization, and trade openness are added to the EKC curve model (Equation 4). Among them, $(\ln \mathrm{GDP}_{it})^2$ indicates the square of the economic development level after taking the logarithm. T represents the technical level. represents the proportion of people over 65 in countries or regions. Urb indicates the level of urbanization in various countries or regions. Open represents the proportion of total imports and exports of goods to GDP. Due to the hysteresis of carbon emissions, the hysteresis term for the dependent variable is introduced for equation (3):

$$\ln CO_{2it} = \theta * \ln I_{i,t-1} + \beta_0 + \beta_1 * \ln P_{it} + \alpha_1 * \ln GDP_{it} + \alpha_2 * (\ln GDP_{it})^2 + \beta_2 * \ln T_{it} + \beta_3 * \ln Age_{it} + \beta_4 * \ln Urb_{it} + \beta_5 * \ln Open_{it} + \mu_{it} \quad (3)$$

Among them, $\ln I_{i,t-1}$ is the first order lag term of $\ln I_{it}$ The magnitude of θ reflects the extent to which the previous carbon emissions affected the current carbon emissions.

2.2 Data

This paper selects (1) the total carbon dioxide emissions of countries or regions over the years to represent carbon emissions (CO2). The carbon emissions data is derived from the European

Commission's Global Atmospheric Emissions Database. (2) The per capita GDP that selected from each country or region is to represent the level of economic development (GDP). (3) The total population of selected countries or regions over the years is to indicate the population (P). (4) The

carbon emission intensity is to indicate the technical level (T), that is, the carbon emissions per 1,000 US dollars of GDP. (5) The improvement of urbanization level has promoted production and consumption levels, resulting in an increase in carbon emissions. This paper selects the proportion of population that aged 65 and over to indicate the level of urbanization. (6) International trade is also one of the reasons for the increasing carbon emissions. The selected proportion of total imports and exports of goods to GDP in this paper indicates trade openness. The selected samples and data descriptive statistics are shown in Table 1.

Table 1 Descriptive statistics of variables

	Variable	Unit	Obs	Mean	Std. Dev.	Min	Max
Developing country	Carbon emission	Million tons	250	115.06	198.46	11.37	1050.00
	Population	Million Peoples	250	4.62	5.62	0.17	17.70
	Economic development	Dollar	250	6578.55	6516.39	298.22	27811.37
	Techinque level	Kilogram/thousand dollars	250	0.42	0.23	0.14	1.18
	Aging	%	250	6.88	3.06	2.86	13.82
	Urbanization	%	250	61.76	22.56	25.52	91.63
	Openness	%	250	124.50	72.31	1.00	249.00
Developed country	Carbon emission	Million tons	250	143.58	169.09	28.16	582.71
	Population	Million Peoples	250	1.17	1.02	0.18	3.41
	Economic development	Dollar	250	31721.77	12985.08	3828.72	67990.29
	Techinque level	Kilogram/thousand dollars	250	0.34	0.14	0.13	0.69
	Aging	%	250	15.14	4.40	4.42	26.56
	Urbanization	%	250	77.89	5.30	66.74	91.46
	Openness	%	250	122.52	72.27	1.00	247.00

3. Results

In this paper, we use the fixed effect and the random effect to empirically test equation (3). The regression estimation results are shown in Table 2. Among them, the model (1), model (3) and model (5) estimate the static equations of the G20, developing countries and developed countries respectively by using the fixed effect (FE). By using the random effects (RE), model (2), model (4) and model (6) estimate the static equations of the G20, developing countries, and developed countries.

Table 2 Fixed effect and random effect regression results

	G20		Developing country		Developed country	
	(1) FE	(2) RE	(3) FE	(4) RE	(5) FE	(6) RE
lnGDP	0.644**	0.675**	1.423**	1.039*	3.601***	4.846***
	(2.68)	(2.70)	(3.08)	(2.30)	(7.44)	(10.80)
lnGDP2	-0.006	-0.012	-0.052*	-0.044	-0.145***	-0.214***
	(-0.50)	(-0.92)	(-1.97)	(-1.66)	(-5.95)	(-9.66)
lnP	1.056***	0.975***	1.027***	0.877***	1.102***	1.093***
	(40.79)	(39.38)	(15.93)	(15.47)	(115.22)	(110.87)

	G20		Developing country		Developed country	
	(1) FE	(2) RE	(3) FE	(4) RE	(5) FE	(6) RE
lnT	0.956***	1.010***	0.921***	0.996***	1.202***	1.263***
	(25.95)	(26.79)	(14.85)	(16.56)	(41.28)	(45.95)
lnAge	0.119*	0.206***	0.184*	0.155	0.136**	0.233***
	(2.45)	(4.18)	(1.99)	(1.67)	(2.70)	(4.62)
lnUrb	0.910***	0.787***	0.724*	0.872**	-0.459**	-0.666***
	(5.78)	(4.90)	(2.24)	(2.85)	(-3.16)	(-4.88)
lnOpen	0.145***	0.080***	0.245***	0.158***	0.083***	0.050***
	(7.57)	(4.57)	(5.76)	(4.25)	(8.50)	(6.63)
_cons	-15.388***	-12.931***	-17.822***	-12.474***	-25.382***	-29.902***
	(-15.53)	(-13.19)	(-8.24)	(-6.64)	(-10.64)	(-12.60)
N	500	500	250	251	252	250

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

From the regression estimation results, the primary and secondary regression coefficients of GDP per capita are positive and negative respectively, and the EKC curve exhibits an inverted U-shaped relationship. The Hausmann test has a P value of 0.000 and a fixed effect model should be chosen. In the fixed effect model, the impact of the G20 per capita GDP is positive and significant, and the secondary term of GDP per capita is negatively correlated with carbon emissions, which is not significant. Population size and technology have the greatest impact on carbon emissions. They are 1.056 and 0.956 respectively. The impact of aging, urbanization and trade openness on carbon emissions are positive and significant. The G20 carbon emissions are in line with the EKC curve inverted U-type assumption. The overall development level of all countries is still on the left side of the inflection point, and has not yet reached the inflection point. The primary and secondary items of GDP per capita in developing countries are significant. The carbon emissions of developing countries are in line with the U-shaped relationship of the EKC curve. The overall development level of all countries is still on the left side of the inflection point and has not yet reached the inflection point. The impacts of population, technology and urbanization on carbon emissions were 1.027, 0.921 and 0.787, respectively, and the impact of aging and trade openness on carbon emissions was positive and significant. The primary and secondary items of per capita GDP in developed countries are significant. The carbon emissions of developed countries are in line with the U-shaped curve of the EKC curve. The overall development level of all countries is still on the left side of the inflection point and has not yet reached the inflection point. The impacts

of population size and technology on carbon emissions were 1.102 and 1.202, respectively. The impact of aging and trade openness on carbon emissions was positive and significant, while the impact of urbanization on carbon emissions was significantly negatively correlated.

The first-order lag term for carbon emissions is introduced in Table 3, and the dynamic panel regression results of developing and developed countries are compared. Among them, model (7) and model (10) are the results of fixed-effect regression in developing and developed countries. Model (8), model (9), model (11) and model (12) are estimates of differential GMM and system GMM for developing and developed countries. The predicted results show that the cumulative effect of carbon emissions in developing countries is stronger than that in developed countries, indicating that developed countries are more effective in environmental governance than developing countries. Both carbon emissions and economic growth have an inverted U-shaped relationship. Population size, carbon intensity and aging levels are positively correlated with carbon emissions. The level of urbanization in developing countries is positively correlated with carbon emissions, while the level of urbanization in developed countries is negatively correlated with carbon emissions. The impact of openness in developing countries on carbon emissions is negative but not significant, while it is positive in developed countries.

Table 3 Dynamic panel regression results for developing and developed countries

	Developing country			Developed country		
	(7)	(8)	(9)	(10)	(11)	(12)
lnCO2i,t-1	0.438***	0.468***	0.299***	0.070***	0.077***	0.080***
	(16.50)	(158.70)	(191.46)	(5.37)	(21.16)	(33.76)
lnGDP	1.069***	1.336***	0.646***	2.117***	2.207***	3.971***
	(3.92)	(22.62)	(18.24)	(5.31)	(8.70)	(17.82)
lnGDP2	-0.034*	-0.049***	-0.034***	-0.069***	-0.069***	-0.165***
	(-2.18)	(-15.53)	(-17.52)	(-3.41)	(-5.51)	(-15.12)
lnP	1.095***	1.174***	0.757***	1.159***	1.162***	1.148***
	(21.24)	(115.77)	(114.22)	(135.99)	(531.33)	(1033.42)
lnT	0.793***	0.771***	0.985***	1.302***	1.309***	1.375***
	(19.76)	(118.57)	(586.60)	(37.63)	(88.99)	(145.97)
lnAge	0.748***	0.714***	0.732***	0.265***	0.187***	0.285***
	(12.48)	(82.25)	(152.01)	(5.79)	(8.76)	(19.03)
lnUrb	1.601***	1.811***	1.569***	-0.890***	-1.071***	-1.223***
	(7.58)	(31.23)	(44.44)	(-7.76)	(-36.48)	(-48.14)
lnOpen	0.101***	0.142***	-0.006	0.096***	0.106***	0.063***
	(3.35)	(12.66)	(-1.67)	(12.74)	(27.07)	(29.85)

	Developing country			Developed country		
	(7)	(8)	(9)	(10)	(11)	(12)
cons	-27.213***		-14.586***	-18.620***		-25.773***
	(-15.02)		(-97.22)	(-9.50)		(-21.90)
N	225	200	225	225	200	225

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

4. Discussion and Conclusion

The static regression results show that the relationship between the G20 carbon emission and the level of economic development conforms to the inverted U-shaped assumption of the EKC curve. Population, technical level and urbanization level of developing countries have a greater impact on carbon emissions, and the impact of aging and trade openness on carbon emissions is positive and significant. The population size and technical level of developed countries have a greater impact on carbon emissions. The impact of aging and trade openness on carbon emissions is positive and significant, while the impact of urbanization on carbon emissions is significantly negatively correlated.

The results of dynamic regression embody that carbon emissions and economic growth in both developed and developing countries show an inverted U-shaped relationship. Population size, carbon intensity and ageing are positively correlated with carbon emissions. The level of urbanization in developing countries is positively related to carbon emissions. With the rapid development of urbanization, energy consumption and carbon emissions will increase. However, the level of urbanization in developed countries is negatively correlated with carbon emissions. As urbanization improves environmental pollution control and energy use efficiency, it also reduces environmental pressure to a certain extent. The impact of openness of developing countries on carbon emissions is negative but not significant. The openness of developed countries has a positive impact on carbon emissions. The increase in imports has caused some carbon emissions to be transferred to importing countries, resulting in an increase in carbon emissions in importing countries.

Firstly, formulating strict and effective environmental regulation policies. Through strict environmental regulation policies, we will tighten emission standards for different types of industry enterprises, promote energy conservation and emission reduction, and thus enhance the development of green economy. It is easier for developed countries to effectively reduce the intensity of carbon emissions, reach the right side of the inflection point as soon as possible, and take the initiative to assume responsibility for energy conservation and emission reduction. Conversely, the developing countries

should actively change their development strategy and promote industrial upgrading and green sustainable development. Thirdly, population structure and the level of urbanization should receive higher priority. With the growth of population size and the rapid development of urbanization, carbon emissions will also increase. Finally, continuing cooperation with countries to transfer technology, and establish a unified carbon trading market. Protecting the environment is the responsibility of every country.

References

1. Al-Mulali, U., Ozturk, I., & Lean, H. H. (2015). The influence of economic growth, urbanization, trade openness, financial development, and renewable energy on pollution in Europe. Natural Hazards, 79(1), 621-644.
2. Azomahou, T., Laisney, F., & Van, P. N. (2006). Economic development and CO_2 emissions: a nonparametric panel approach. Journal of Public Economics, 90(6-7), 1347-1363.
3. Friedl, B., & Getzner, M. (2003). Determinants of CO_2 emissions in a small open economy. Ecological economics, 45(1), 133-148.
4. Grossman, G. M., & Krueger, A. B. (1991). Environmental impacts of a North American free trade agreement (No. w3914). National Bureau of Economic Research.
5. Lean, H. H., & Smyth, R. (2010). Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia. Energy, 35(9), 3640-3648.
6. Ozturk, I., & Acaravci, A. (2013). The long-run and causal analysis of energy, growth, openness and financial development on carbon emissions in Turkey. Energy Economics, 36, 262-267.
7. Richmond, A. K., & Kaufmann, R. K. (2006). Is there a turning point in the relationship between income and energy use and/or carbon emissions?. Ecological economics, 56(2), 176-189.
8. Shafik, N. (1994). Economic development and environmental quality: an econometric analysis. Oxford economic papers, 757-773.

CSI300 volatility forecasting model and its MCS test

Qiguang Dong[1], Hang Li[2], Lili Chen[3]

[1]PLA Military Science Academy, 100142, Beijing, China

[2]Donlinks School of Economics and Management, University of Science and Technology 100083, Beijing, China

[3]School of Economics and Management, Tsinghua University, 100084, Beijing, China

Abstract

In this paper, 5 min frequency observations are taken to forecast the actual volatility of CSI300 stock index intraday returns. Both realized volatility and logarithm-transformed realized volatility are modelled directly in the ARFIMA model specification. Besides, GARCH family models and different distributions are utilized to address the potential heteroscedasticity problem. Applying the out-of sample rolling time window forecasting and Model Confidence Set which is proved superior to SPA test, this paper compares the empirical performance of all specified models. The empirical results show that: (1) Both RV and LnRV series have a long memory due to both Hurst indexes are greater than 0.5 and smaller than 1. (2) The symmetric and skewed generalized error distributions ged and sged are employed more accurate than normal and student-t distributions. (3) The model LnRV-sGARCH-sged is outperformed than the rest in the long memory model as well as in the short memory model.

Keywords

Volatility forecasting; stock index; MCS test; CSI300

1. Introduction

The efficiency of the volatility forecast is crucial for option pricing, but also in many areas of finance. Since ARCH models, GARCH models and Stochastic Volatility models are introduced to estimate and forecast volatility in financial market, these models gradually established their dominance and still in the continuous development. However, these models usually use lower frequency observations to estimate and forecast volatility of financial assets, which lost an amount of intraday trading information and have difficulty to deal with multidimensional problems. High frequency observations can quickly and effectively to capture market information, which can be more accuracy to reflect the actual situation in the financial markets than lower frequency observations. (Andersen and Bollerslev, 1998; BarndorffNielsen and Shephard,2002). RV is the sum of intraday squared returns, which accuracy depends on the returns of intraday high frequency observations (Taylor and

Xu,1997). Hence, both issues inherent in the high frequency observations are considered in this paper.

In the specification of model structure, Barkoulas and Baum (1997) first use the ARFIMA model structure to forecast the Eurocurrency return series and show that the ARFIMA model can improve of the forecasts. Unlike the autocorrelation process decays exponentially in the ARMA model, it decays hyperbolically which can be more slowly in the ARFIMA model. Besides, Andersen et al. (2008) find that the out-of-sample forecasts based on ARFIMA model are performed better than other long memory models, such as FIGARCH, FIEGARCH. In addition, Kanellopoulou and Panas (2008) argue that the ARFIMA models are strictly based on the assumptions of no conditional heteroscedasticity and normal distributions, otherwise the forecasts will be biased. However, previous literature confirms that the volatility of financial assets violates both assumptions. Researchers use the logarithm-transformed realized variance, namely LnRV, instead of RV in the model specification. Though LnRV series are more approximate normal distribution than RV series. Hence, this paper considered these problems during the model specification by combining GARCH family models and non-normal distributions.In the evaluation of models predictive ability, Hansen and Lunde (2005) introduced Superior Predictive Ability (SPA) test based on the bootstrap method, which use a set of loss functions to deliver the optimal models with respect to a given set of loss functions. SPA test needs to set benchmark model at first, then the other models are compared to the benchmark model so that models that produce better forecasts are preferred. However, the procedure may bring two problems: first, the comparing procedure may not deliver a unique result; second, sometimes it is not trivial to asses which model clearly outperforms each other. For overcoming the deficiencies of SPA test, Hansen, Lunde and Nason (2011) further introduced a new test, the Model Confidence Set (MCS) test. MCS test permits to construct a set of "superior" models, and can directly evaluate and compare the models predictive ability in the set.

2. Methodology

In the specification of volatility forecasting model, Andersen et al. (2003) find that the variance pattern of financial asset can be described as Gaussian dynamic process, and the RV series shows long memory features. Based on this, this paper chooses the long term ARFIMA to construct the empirical model.

$$\emptyset(L)(1-L)^d(y_t-\mu)=\theta(L)\varepsilon_t \quad (1)$$

Where d is the degree of long memory fractional integration process with 0<d<1. L is the lag operator. $\emptyset(L)$ is the lag operator of AR, and $\theta(L)$ is the lag operator of MA. $(1-L)^d$ is the difference operator, which represents the long memory.

$$(1-L)^d = \sum_{k=0}^{\infty} \frac{\Gamma(k-d)L^k}{\Gamma(-d)\Gamma(k+1)} \tag{2}$$

ARFIMA models are strictly based on the assumptions of no conditional heteroscedasticity and normal distributions[9]. However, current literatures show that the volatility series violate both assumptions. Hence, this paper considers these problems through two ways. Before the evaluation of models predictive ability, this paper applies the out-of-sample rolling time window forecasting. After then, this paper compared the evaluated volatility with realized volatility (in this paper it means both RV and LnRV) through loss function, so that we can measure the accuracy of each models. Based on this, this paper uses 6 different loss functions as follows:

$$MSE = \frac{\sum_{t=1}^{T}\left(RV_t - \widehat{RV}_t\right)^2}{T} \tag{3}$$

$$HMSE = \frac{\sum_{t=1}^{T}\left(1 - {\widehat{RV}_t}/{RV_t}\right)^2}{T} \tag{4}$$

$$MAE = \frac{\sum_{t=1}^{T}\left|RV_t - \widehat{RV}_t\right|}{T} \tag{5}$$

$$HMAE = \frac{\sum_{t=1}^{T}\left|1 - {\widehat{RV}_t}/{RV_t}\right|}{T} \tag{6}$$

$$QLIKE = \frac{\sum_{t=1}^{T}\left(\ln \widehat{RV}_t - {RV_t}/{\widehat{RV}_t}\right)}{T} \tag{7}$$

$$R^2LOG = \frac{\sum_{t=1}^{T}\left(\ln {RV_t}/{\widehat{RV}_t}\right)^2}{T} \tag{8}$$

3. Results

In this paper, we choose the CSI300 stock index 5 min frequency closing price from December 16th, 2012 to April 13th, 2016. The data are extracted

from Wind Database. Table 1 shows the summary statistics of RV and LnRV. Table 1 shows that both RV and LnRV are significantly skewed and leptokurtic, and the sample autocorrelation coefficients are high and slowly decaying at lags 5, 10 and 20. This paper also use Rescaled Range Analysis (R/S) to test the long memory in both volatility series, and the Hurst values show that both volatility series have significant long memory features. Besides, the ADF test indicates that both volatility series are stationary.

Table 1 Summary Statistics.

	RV	**LnRV**
Mean	2.70329	0.369385
std_dev	4.836594	1.01278
Skewness	5.582336	0.605968
Kurtosis	39.6458	0.488312
series.1	56214.41	56.74069
ADF	-5.6116	-3.41894
Q5	1338.818	2280.86
Q10	1768.675	3990.099
Q20	2346.429	6734.344
Hurst	0.764089	0.831612

In this paper, both RV and LnRV are modeled directly in the ARFIMA model specification in which the structure of the model is optimized using the AIC, BIC and HIC criterions. In regard of conditional heteroscedasticity and non-normal distribution, this paper combined the ARFIMA models with 3 GARCH models and 6 distributions in model specification. Hence, we constructs and tests 36 long memory models in this paper (table 2).

For obtaining MCS statistics and corresponding p-values, we set the block length d=2 and simulation times B=10000 as the control parameters during the Bootstrap procedure. Following Hansen, Lunde and Nason [11], we set the confidence level α=0.1, so that the model will be eliminated if its p-values <0.1, otherwise it will survive the MCS procedure.

From table 3 we can draw the following conclusions. (1) In the first 5 loss functions, models based on RV are eliminated completely which p-values <0.1, and only few models survived the 6th loss function. The results indicates that the logarithm-transformed realized volatility are outperformed than realized volatility, which is consist with Kotkatvuori-Örnberg (2016).

Table 2 ARFIMA models.

M	RV	M	LnRV
M1	RV-sGARCH-norm	M19	LnRV-sGARCH-norm
M2	RV-sGARCH-std	M20	LnRV-sGARCH-std
M3	RV-sGARCH-ged	M21	LnRV-sGARCH-ged
M4	RV-sGARCH-snorm	M22	LnRV-sGARCH-snorm
M5	RV-sGARCH-sstd	M23	LnRV-sGARCH-sstd
M6	RV-sGARCH-sged	M24	LnRV-sGARCH-sged
M7	RV-eGARCH-norm	M25	LnRV-eGARCH-norm
M8	RV-eGARCH-std	M26	LnRV-eGARCH-std
M9	RV-eGARCH-ged	M27	LnRV-eGARCH-ged
M10	RV-eGARCH-snorm	M28	LnRV-eGARCH-snorm
M11	RV-eGARCH-sstd	M29	LnRV-eGARCH-sstd
M12	RV-eGARCH-sged	M30	LnRV-eGARCH-sged
M13	RV-gjrGARCH-norm	M31	LnRV-gjrGARCH-norm
M14	RV-gjrGARCH-std	M32	LnRV-gjrGARCH-std
M15	RV-gjrGARCH-ged	M33	LnRV-gjrGARCH-ged
M16	RV-gjrGARCH-snorm	M34	LnRV-gjrGARCH-snorm
M17	RV-gjrGARCH-sstd	M35	LnRV-gjrGARCH-sstd
M18	RV-gjrGARCH-sged	M36	LnRV-gjrGARCH-sged

Table 3 MCS test results based on ARFIMA models.

	MSE		HMSE		MAE		HMAE		QLIKE		R^2LOG	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
M1	0	0	0	0	0	0	0	0	0	0	0	0
M2	0	0	0	0	0	0	0	0	0	0	0	0
M3	0	0	0	0	0	0	0	0	0	0	**0**	**0**
M4	0	0	0	0	0	0	0	0	0	0	0	0
M5	0	0	0	0	0	0	0	0	0	0	**0**	**0**
M6	0	0	0	0	0	0	0	0	0	0	**0**	**0**
M7	0	0	0	0	0	0	0	0	0	0	0	0

	MSE		HMSE		MAE		HMAE		QLIKE		R^2LOG	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
M8	0	0	0	0	0	0	0	0	0	0	0	0
M9	0	0	0	0	0	0	0	0	0	0	0	0
M10	0	0	0	0	0	0	0	0	0	0	0	0
M11	0	0	0	0	0	0	0	0	0	0	0	0
M12	0	0	0	0	0	0	0	0	0	0	0	0
M13	0	0	0	0	0	0	0	0	0	0	0	0
M14	0	0	0	0	0	0	0	0	0	0	0	0
M15	0	0	0	0	0	0	0	0	0	0	0	0
M16	0	0	0	0	0	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	0	0	0	**0**
M18	0	0	0	0	0	0	0	0	0	0	**1**	**1**
M19	0	0	0	0	0	0	0	0	**0**	**0**	0	0
M20	0	0	0	0	0	0	0	0	0	**0**	0	0
M21	0	0	0	0	0	**0**	0	0	0	**0**	0	0
M22	0	0	**0**	0	0	0	0	0	**0**	**0**	0	0
M23	0	0	**0**	**0**	0	0	0	0	**0**	**0**	0	0
M24	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	0	**0**	0	0
M25	0	0	0	0	0	0	0	0	0	**0**	0	0
M26	0	0	0	0	0	0	0	0	0	**0**	0	0
M27	0	0	0	0	0	0	0	0	0	**0**	0	0
M28	0	0	0	0	0	0	0	0	0	**0**	0	0
M29	0	0	0	0	0	0	0	0	0	**0**	0	0
M30	0	0	0	0	0	0	0	0	0	**0**	0	0
M31	0	0	0	0	0	0	0	0	0	**0**	0	0
M32	0	0	0	0	0	0	0	0	0	**0**	0	0
M33	**0**	**0**	0	0	**0**	**0**	**0**	**0**	0	**0**	0	0
M34	0	0	0	0	0	0	0	0	**0**	**0**	0	0
M35	0	0	0	0	0	0	0	0	**0**	**0**	0	0
M36	0	0	0	0	0	0	0	0	**0**	**0**	0	0

Table 4 MCS test results based on ARMA models.

	MSE		HMSE		MAE		HMAE		QLIKE		R^2LOG	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
M1	0	0	0	0	0	0	0	0	0	0	0	0

	MSE		HMSE		MAE		HMAE		QLIKE		R^2LOG	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
M2	0	0	0	0	0	0	0	0	0	0	0	0
M3	0	0	0	0	0	0	0	0	0	0	0	0
M4	0	0	0	0	0	0	0	0	0	0	**1**	**1**
M5	0	0	0	0	0	0	0	0	0	0	0	0
M6	0	0	0	0	0	0	0	0	0	0	0	0
M7	0	0	0	0	0	0	0	0	0	0	0	0
M8	0	0	0	0	0	0	0	0	0	0	0	0
M9	0	0	0	0	0	0	0	0	0	0	0	0
M10	0	0	0	0	0	0	0	0	0	0	0	0
M11	0	0	0	0	0	0	0	0	0	0	0	0
M12	0	0	0	0	0	0	0	0	0	0	0	0
M13	0	0	0	0	0	0	0	0	0	0	0	0
M14	0	0	0	0	0	0	0	0	0	0	0	0
M15	0	0	0	0	0	0	0	0	0	0	0	0
M16	0	0	0	0	0	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	0	0	0	0
M18	0	0	0	0	0	0	0	0	0	0	0	0
M19	0	0	0	0	0	0	0	0	0	0	0	0
M20	0	0	0	0	0	0	0	0	0	0	0	0
M21	0	0	0	0	0	0	0	0	0	0	0	0
M22	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	0	0	0
M23	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	0	0	0
M24	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0	0
M25	0	0	0	0	0	0	0	0	**1**	**1**	0	0
M26	0	0	0	0	0	0	0	0	0	0	0	0
M27	0	0	0	0	0	0	0	0	0	0	0	0
M28	0	0	0	0	0	0	0	0	0	0	0	0
M29	0	0	0	0	0	0	0	0	0	0	0	0
M30	0	**0**	0	0	0	0	0	0	0	0	0	0
M31	0	**0**	0	**0**	0	0	0	0	0	0	0	0
M32	0	**0**	0	0	0	0	0	0	0	0	0	0
M33	0	0	0	0	0	0	0	0	0	0	0	0
M34	**1**	**1**	**0**	**0**	**0**	**0**	0	0	0	0	0	0

	MSE		HMSE		MAE		HMAE		QLIKE		R^2LOG	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
M35	**0**	**0**	0	**0**	**0**	**0**	0	0	0	0	0	0
M36	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	0	0	0

(2) Compared the models based on different distribution, we can find that, the amount of survived models based on distributions ged and sged are more than the survived models based on the rest distributions. This results suggest that the symmetric and skewed generalized error distributions mostly approximate the actual distribution of the volatility series than normal and student-t distributions.

(3) In all 36 models, the models which survived the most loss functions is M24, namely LnRVsGARCH-sged. It survived the first 5 loss functions, and the following model is M33 (LnRV-gjrGARCHged). For the sake of robustness, this paper utilizes the short term ARMA model instead of ARFIMA in model specification to construct another 36 forecasting models for robust test. The MCS test results are presented in table4. The results are consistent with the results shown in table 3, and the model M24 is still outperformed.

4. Discussion and Conclusion

This paper utilizes the realized volatility and logarithm-transformed realized volatility to forecast the actual volatility of CSI300 stock index. We construct 36 long memory ARFIMA models for forecasting, and then applying the out-of-sample rolling time window forecasting combined with Model Confidence Set test to evaluate and compare the predictive ability of the models specified. For the sake of robustness, we conduct the same procedure to 36 short memory ARMA models and the empirical results are similar. The empirical results show that: (1) Both RV and LnRV series have a long memory due to both Hurst indexes are greater than 0.5 and smaller than 1. (2)The symmetric and skewed generalized error distributions ged and sged are employed more accuracy than normal and student-t distributions. (3) The model LnRV-sGARCH-sged is outperformed than the rest in the long memory model as well as in the short memory model.

References

1. Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International economic review, 885-905.
2. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579-625.
3. Barndorff-Nielsen, O. E., & Shephard, N. (2002). Estimating quadratic variation using realized variance. Journal of Applied econometrics, 17(5), 457-477.
4. Barkoulas, J. T., & Baum, C. F. (1997). Fractional differencing modeling and forecasting of eurocurrency deposit rates. Journal of Financial Research, 20(3), 355-372.
5. Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. Journal of applied econometrics, 20(7), 873-889.
6. Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. Econometrica, 79(2), 453-497.
7. Kotkatvuori-Örnberg, J. (2016). Measuring actual daily volatility from high frequency intraday returns of the S&P futures and index observations. Expert Systems with Applications, 43, 213-222.
8. Kanellopoulou, S., & Panas, E. (2008). Empirical distributions of stock returns: Paris stock market, 1980–2003. Applied Financial Economics, 18(16), 1289-1302.
9. Taylor, S. J., & Xu, X. (1997). The incremental volatility information in one million foreign exchange quotations. Journal of Empirical Finance, 4(4), 317-340.

Measuring the hidden economy and improving Moroccan GDP exhaustiveness through the labor matrix

Bahija Nali, Yattou Ait Khellou
National accountants, HCP, Morocco

Abstract

GDP or gross domestic product is an economic aggregate that measures the level of production and wealth achieved within a country or zone during a given period, typically the year or quarter. It is an aggregate that is widely used by national policy makers for the establishment of adequate economic policies, and by international organization for international comparison purposes. For this reason, national accountants are compelled to ensure the exhaustiveness of the measurement of the various activities included in this indicator according to the standards defined by the System of National Accounts 2008 (2008 SNA). To achieve exhaustiveness, national accountants must identify the entire productive universe and collect information on all activities that fall within the production boundary. However, the character extremely broad of this area, which is defined by the SNA, makes the task very difficult. It recommends that all economic activities, irrespective of their nature: formal or informal, legal or illegal and declared or unreported, be included in national accounts estimates. The approach we are presenting is based on a local reality, where the unregistered "activities" occupy an important place. In such a context, embracing the entire productive space across production units proves difficult, since these tend to conceal a part of their productions. We will use labour input method which aims to approach the productive world through the labour factor, which is the best known factor of production, as reported by households in Labour Force Survey (LFS). Our work involves the production of a Labour supply matrix (demographic matrix) to be confronted with another matrix that represents the use of labour by employers (economic matrix) to reach, finally, a unique matrix of labour input (jobs' matrix), able to put production in relation with the workforce that gave birth to it. The objective is to capture the employed population not observed or not traced by economic surveys and to assign an output to it.

Keyword

exhaustiveness of national accounts; hidden economy; supply and demand of labour; labour productivity by industry

1. Introduction

GDP or gross domestic product is an economic aggregate that measures the level of production and wealth achieved within a country or zone during a given period, typically the year or quarter. It is an aggregate that is widely used by national policy makers for the establishment of adequate economic policies, and by international organization for international comparison purposes. For this reason, national accountants are compelled to ensure the exhaustiveness of the measurement of the various activities included in this indicator according to the standards defined by the System of National Accounts 2008 (2008 SNA).

To achieve exhaustiveness, national accountants must identify the entire productive universe and collect information on all activities that fall within the production boundary. However, the character extremely broad of this area, which is defined by the SNA, makes the task very difficult. It recommends that all economic activities, irrespective of their nature: formal or informal, legal or illegal and declared or unreported, be included in national accounts estimates.

Nevertheless, information on these activities is generally lacking, and it is difficult to take into account because they are not covered by statistical data collection, and constitute "gray areas" to which the national accountant must shed light.

Despite the difficulty and delicacy of the task, the national accountant is forging the necessary means to capture the unrecorded statistics and clarify the "gray areas". Its objective is to measure all the production carried out in its economic territory and consequently to ensure the exhaustiveness of the aggregate which synthesizes this production which is the GDP.

But far from ensuring the exhaustiveness of the GDP, the unrecorded statistics hides behind it atypical forms of work, which are precarious and unprotected and do not guarantee decent work for those who exercise them, and to which attention must be paid. The development of national policies for the improvement of individual well-being should normally be part of a line that aims to target this segment of the population. To do so, we must first locate it, and then estimate its importance and its contribution to the national production.

Several researchers have attempted to estimate the unrecorded portion of GDP through macroeconomic models, but their results are too dependent on macroeconomic assumptions. And reassessing GDP on their bases would lead to inconsistencies that are difficult to justify by the economic system.

The estimation of the "shadow economy" by macro-modeling methods consists in the elaboration of models through explanatory variables. These are the models that use tax evasion rates, the circulation of money or the consumption of electricity. However, experience in developed and developing countries has shown the inefficiency of these models to clearly identify "areas

of darkness" and their inability to reconstruct the functioning of the economy by integrating their results, especially for developing countries, which are characterized by a very particular economic reality.

The search for exhaustiveness is an inherent feature of national accounts, which seeks rather to make as full use as possible of all available basic data and to use models only at the most disaggregated levels of the national accounts. It compares basic data with one another, interprets discrepancies, makes estimates of unobserved or unregistered activities, and uses specific surveys to measure their magnitude. Above all, it adopts an analytical approach which consists of identifying the various kinds of the non-observed economy and studying the extent to which some of these activities are already taken into account and how to evaluate the shares of those that are not yet taken into account by the usual methods.

The approach we are presenting is based on a local reality, where the unregistered "activities" occupy an important place. In such a context, embracing the entire productive space across production units proves difficult, since these tend to conceal a part of their productions. We will use labour input method which aims to approach the productive world through the labour factor, which is the best known factor of production, as reported by households in Labour Force Survey (LFS).

This approach is based on the most analytical work possible. First, it involves the production of a Labour supply matrix (demographic matrix) to be confronted with another matrix that represents the use of labour by employers (economic matrix) to reach, finally, a unique matrix of labour input (jobs' matrix), able to put production in relation with the workforce that gave birth to it. The objective is to capture the employed population not observed or not traced by economic surveys and to assign an output to it. The output thus calculated is then integrated into an iterative data analysis and arbitration process proposed by the central framework of the national accounts system in order to achieve an exhaustive and integrated measure of national GDP.

Our work consists in the compilation, for a benchmark year of national accounts, of a single jobs' matrix by status in employment (employee, employer, Own-account worker and contributing family worker) and by industry (International Standard Industrial Classification ISIC).

The development of this matrix is carried out in three stages: the first two levels correspond to statistical arbitrations between the supply and use of labour matrixes already prepared. And the last stage aims to provide the elements and indicators needed for the calculation of the output generated by hidden labour.

First stage

a) Labour supply matrix

Two sources are used to compile this matrix, the Labour Force Survey (LFS and Administrative data. Exploitation of the results of the first survey allowed us to elaborate a matrix of the economically active population by industry and status in employment while referring to other administrative data to approximate the total coverage of the employed population.

To ensure the consistency of the data set, we converted the occupied labor force matrix to a Full-time equivalent employment (FTE) matrix using data on number of hours worked for the main and secondary jobs of the employed workforce.

b) Labour demand matrix

It is a matrix developed from the statements of the firms on the number of jobs they generate. In Morocco, the labour demanded is reported in structural surveys realised by the statistical office, namely, surveys on economic structures carried out among organized companies and the survey on the informal sector. These two sources provided us with the information needed to set up the Labour demand matrix.

To produce comparable estimates of labour in the supply and use sides, the number of hours worked by permanent and non-permanent individuals given by structural surveys and the informal sector survey were used to establish the Labour demand matrix broken down by economic activity, status in employment and by formal and informal sector.

Second step

Given that the sampling method used in Labour Force Survey does not ensure the representativeness of the results at a very fine level of industries' classification. It should be noted that we have undertaken this step at an aggregate level that allows analysis by industry.

The second stage concerned the arbitration, analysis and comparison between Labour supply matrix and the Labour demand matrix. This work led to the detection of the discrepancies in the labour inputs reported by businesses in enterprise and informal surveys, and the labour inputs reported by individuals in Labour Force Survey and to the identification of activities showing these discrepancies. The differences in labour input, by industry, are normally the volume of undeclared work or not recorded by statistical surveys and to which an output must be assigned using the output per unit of labour input or value added per unit of labour input for the same activity.

Third step

Once the output is estimated, it is integrated into the central framework of national accounts system, to begin the iterative process of arbitration between the data, permitted by the use of the Supply and Use Table (SUT).

This process allows both to judge the quality of the output estimated and to situate it within the national economy.

2. Results of the study

Before presenting the results, it should be noted that this work concerns a simulation exercise conducted for the last base year of Moroccan accounts namely 2007. Currently, the process of rebasing our national account is in progress, and we propose to conduct the same exercise for the new benchmark year (2014) to ensure a better exhaustiveness of the Moroccan GDP.

Moreover, it important to note that the comparable component of the databases between supply and use of labour was the salaried jobs in non agricultural and non financial private sector.

a) Undeclared employees

	Undeclared salaried Employees jobs on FTE basis				
Industry	Labour Force Survey	Economic Structures Survey	Informal Sector Survey	Undeclared employees	share %
Fishing	62675	16809	10941	34925	4,3
Mining and quarrying	48689	30628		18060	2,2
Manufacturing	954227	488210	146501	319517	39,3
Construction	749176	284713	66154	398308	49,1
Services	1600636	1039912	520168	40556	4,9
TOTAL	3415403	1860272	743764	811367	100

Analysis, confrontation of estimate of number of employees in accordance with the household survey and the number of employees at enterprises and organizations reported by structural surveys and informal sector survey reveal that 811367 salaried jobs (expressed in full time equivalent basis) escape economic surveys. In other words, 23,8 % of employees are hidden or under-reported.

The construction sector is the largest contributor to the hidden work with a 49,1%, followed by manufacturing industries with 39,3%. Services contribute for 4,9%, the fishing sector for 4,3% and lastly, mining and quarrying sector represents 2,2% of the undeclared jobs.

b) Undeclared value added

The purpose of the labour input method is to estimate the input and the value added concealed using ratios of output and value added per unit of labour input. In our paper we propose to use two ratios: in the first case we

use the productivity estimates in small and non organized businesses and in the second case we use the productivity estimates in organized enterprises.

1st scenarios

Concealed value added (VA) in millions of Dirhams			
Industry	VA	**Concealed** VA	**reassessed** VA **(%)**
Fishing	5416	578	10,7
Mining and quarrying	12039	775	6,4
Manufacturing	73094	3569	4,9
Construction	36019	5529	15,4
Services	346559	588	0,2
TOTAL	**473127**	**11038**	**2,3**
AV total[1]	**576619**		**1,9**

1: total value added including agriculture, Administration and Financial services

With the assumption that one unit of hidden jobs produce as well as one labour unit in the unorganized enterprises, the 811367 undeclared jobs (FTEs) will generate 11038 MDH of value added with 5529MDH in construction activities, 3569MDH in manufacturing industries. The fishing, mining and services activities will create 1940MDH of value added.

In total, using this hypothesis the value added of all sectors including Agriculture, Administration and Financial services will be revised upwards by 1,9%. An analysis by activity shows that construction's value added will be reassessed by 15,4%, that of fishing activity by 10,7%. In addition, the value added of Mining and quarrying, manufacturing and services will be revised slightly upward by 6,4%, 4,9% and 0,2% respectively.

2nd scenarios

	Concealed values added(VA) in millions of Dirhams		
Industries	**VA**	**Concealed VA**	**VA reassessed (%)**
Fishing	5416	8477	156,5
Mining and quarrying	12039	5807	48,2
Manufacturing	73094	55094	75,4
Construction	36019	33570	93,2
Services	346559	5213	1,5
TOTAL	**473127**	**108161**	**22,9**
AV total	**576619**		**18,8**

Assuming that, in given activity, one unit of undeclared work produce as well as one labour unit in the organized enterprises, the Value Added created by undeclared jobs will reach 108161MDH. It is made up of 55094MDH in manufacturing industry, 33570MDH in construction and 19497 in the other industries.

Globally, in the second scenario, the total of the Value Added will be reviewed by 18,8%. It should be pointed out that it is in Fishing industry that the reevaluation will be the strongest with 156,5% while the services' Value Added will be reassessed only by 1,5%.

The tradeoff between the added values, in the first scenarios and in the second scenarios, depends largely on the arbitration work relating to the base year. Since the simulation exercise was developed after this work, it was difficult for us to choose between reassessed added values. Nevertheless, the two scenarios above represent the lower and upper revaluation limits that must not be exceeded during analysis and arbitration work.

c) GDP reassessment

	GDP reassessed in million of Dirhams		
	Initial	**reevaluation**	
		1st scenario	**2nd scenario**
GDP	647530	658568	755691
GDP reassessment rate (%)		1,7	16,7

Taking into account the hidden values added in the framework of Moroccan national accounts would have made it possible to reassess the GDP by 11038MDH in the first scenario and by 108116MDH in the second scenario. Thus, it would have successively passed to 658568MDH in the first case and 755691MDH in the second case instead of 647530MDH, that to say a reassessment of 1.7% and 16.7%.

Big Data

Md Shariful Islam
Department of Cypher, Ministry of Defense, Bangladesh

Abstract

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyse some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision. Big Data Analytics is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies. Big Data Analytics poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. Potential breakthroughs include new algorithms, methodologies, systems and applications in Big Data Analytics that discover useful and hidden knowledge from the Big Data efficiently and effectively.

Keywords

Batch Processing; Cassandra; Cloud Computing; NoSQL; ETL

1. Introduction

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large. It is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big to moves fast or it exceeds current processing capacity. Besides

Big Data refers to collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data process applications. Big Data Analytics and Data process tiles a platform to globalize the research by installing a dialogue between industries and academic organizations and knowledge transfer from research to industry. Big data can be characterized by 3vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be processed. Although the big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily. Because big data takes too much time and costs to money to load into a traditional relational database for analysis, new approaches to storing and analyzing data have e emerged that rely less on data schema and data quality. Although the demand for big data analytics is high, there is currently a shortage of data scientists and other analysis who have experience working with big data in a distributed, open source environment. In the enterprise, vendors have responded to this shortage by creating Hadoop appliances to help companies take advantages of the semi-structured and unstructured data they own. Big data can be contrasted with small data, another evolving term that's often used to describe data whose volume and format can be easily used for self-service analytics. A country quoted axiom is that "big data is for machines; small data is for people".

2. Methodology

Industry press is enamoured by the 4 V's of Big Data. These are Volume, Velocity, Variety and Veracity. Volume is referring to the size of the data. Velocity is referring to the speed of how data is collected and consumed. Variety referring to the different kinds of data consumed, from structured data, unstructured data and sensor data. Veracity is referring to the trustworthiness of the data.

The Methods of big Data can be described by the following characteristics:

1. Volume-The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration.
2. Variety- The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysis.
3. Velocity-The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.
4. Variability- This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown

by the data at times, thus hampering the process of being able to handle and manage the data effectively

5. Veracity- The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.
6. Complexity- Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to linked, connected and correlation in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as 'complexity' of Big Data.

3. Result of Big Data

Big Data applications designed to efficiently and easily collect, archive, display, transform, and process large sets of structured and unstructured data. Cloud-based strategies to provide access to data no matter where it resides inside or outside the organization. Analytics suites that contain automated capabilities, an open architecture, and a standards-based framework necessary to support activity-based intelligence. Dynamic capabilities that enable data integration from across the enterprise without the need for a lengthy process and costly rigid data warehouse integration. The ability to analyse big data provides unique opportunities for your organization as well. It shall be able to expand the kind of analysis can do. Instead of being limited to sampling large data sets, we can now use much more detailed and complete data for analysis. However, analyzing big data can also be challenging. Changing algorithms and technology, even for basic data analysis, often has to be addressed with big data.

4. Discussion and Conclusion

4.1 Big Data analytical tools

Analyzing Big Data can be very cumbersome and challenging. There is no particular software that can be used for the analysis. Different enterprises use different tools for Big Data analysis: However, the tool to use depends on the type of the data one needs to analyze. The choice of tools can also affect the quality of data one needs to analyze. The choice of tools can also affect the quality of data which can have a significant impact on analysis. Data can be analyzed both structural and unstructured data (Big Data) by different tools. Some tools are open wares while others are commercial and very expensive.

4.2 Open source Big Data Analysis Platforms and Tools

1 Hadoop

Without Hadoop no one can talk about big data. The Apache distributed data processing software is so pervasive that often the terms 'Hadoop' and 'big data' are used synonymously. The Apache distributed data processing software is so pervasive that often that often the terms 'Hadoop and 'Big Data' are used synonymously. The Apache Foundation also sponsors a number of related projects that extend the capabilities of Hadoop, and many of them are mentioned below. In addition, numerous vendors offer supported versions of Hadoop and related technologies. Operating system: Windows,
Linux, OS X,

2 MapReduce

Originally developed by Google, the Mapreduce website describes it as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes". It's used by Hadoop, as well as many other data processing applications. Operating System: OS Independent.

3 Gridgrain

Gridgrain offers an alternative to Hadoop's Mapreduce that is compatible with the Hadoop Distributed file system. It often in memory processing for fast analysis of real time data. One can Download the open source version from GitHub or purchase a commercially supported version from the link in operating System: Windows, Linux, OS X.

4 HPCC

Developed by LexisNexis Risk Solutions, HPCC is short for "high performance computing cluster." It claims to offer superior performance to Hadoop. Both free community versions and paid enterprise versions are available. Operating
System: Linux.

5 Storm

Owned by Twitter, Storm offers distributed real-time computation capabilities and is often described as the "Hadoop of real-time." It's highly

scalable, robust, and fault-tolerant and works with nearly all programming languages. Operating System: Linux.

6 Cassandra

Originally developed by Facebook, this NoSQL database is now managed by the Apache Foundation. It's used by many organizations with large, active datasets, including Netflix, Twitter, Urban Airship, Constant Contact, Reddit, Cisco and Digg. Commercial support and services are available through third-party vendors. Operating System: OS Independent.

4.3 The Benefit Big Data Analytics

Collecting and storing big data does not create business value. Value is created only when the data is analyzed and acted on. As the Starbucks, Chevron, and U.S. Xpress examples show, the benefits from big data analytics can be varied, substantial, and the basis for competitive advantage. Because of its potential benefits, some people add a fourth V to the characteristics of big data: high value. This value is realized, however, only when an organization has a carefully thought out and executed big data strategy.

Access to Big Data source and forging partnerships with other public and private organizations in order to work with big Data is becoming ever more important to national statistical systems (NSS) for fulfilling their mission in society. The National statistical systems (NSS) should collaborate rather than compete with the private sector, in order to advance the potential of official statistics. At the same time, the NSS should remain and impartial, and invest in communicating the advantages of exploiting the wealth of available digital data to the benefit of the people. Building public trust will be the key to success. The objectives of the task team are to facilitate access to Big Data sources for official statistics and facilitate forming partnerships with other public and private organizations in order to work with Big Data. The GWG Big Data Inventory is a catalog of Big Data projects that are relevant for official statistics, SDG indicators and other statistics needed for decision-making on public policies, as well as for management and monitoring of public sector programs/projects. This inventory is a joint product of the World Bank and the United Nations Statistics Division (UNSD) put together on behalf of the UN Global Working Group (GWG) on Big Data for Official Statistics.

4.4 Big Data and the Data Revolution

Big Data is not a single 'thing' - it is a collection of data sources, technologies and methodologies that have emerged from, and to, exploit the exponential growth in data creation over the past decade. Big data is a buzzword; used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. Data is a growing element of our lives. More and more data is being produced and becoming known in the popular literature as “big data”, its usage is becoming more pervasive, and its potential

for policy making and international development is just beginning to be explored.

4.5 A vision for Big Data and the 2030 Agenda

The very nature of Big Data requires new forms of inter-institutional relationships in order to leverage data resources, human talent, and decision-making capacity. The necessary capabilities enable the integration of big data into on-going policy decisions, thereby enabling its value to be continuity rather than one-time policy decisions, thereby enabling its value to be continually released and refined. Spaces will be needed in which technical, cultural, and institutional capabilities can commensurately develop. Given the variety and pervasiveness of the necessary capabilities to utilize big data to address big problems, collaborative spaces are needed to enhance the capacity of individuals, organizations, businesses and institutions to elucidate challenges and solutions in an interactive manner, strengthening a global culture of learning. Some elements of this neo ecosystem are already emerging. The UN Statistical commission established a Global Working Group (GWG) mandated to provide strategic vision, direction and coordination of a global program on Big Data for official statistics. The group found that nontraditional sources of data need to be leveraged and considered for adequacy to enrich the sources of official so that the data needs in new development areas can be satisfied and timely, detailed and spatially disaggregated data can be produced and made available to decision makers. This implies that the innovative and transformative power of information technology may be harnessed' from the collection stage to dissemination stage. The UN's Secretary General's independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG) is calling for action to mobilize the data revolution for sustainable. The recommendations out of this group relevant for better policy making have been taken as a basis for possible action steps that are given below.

4.6 Big Data and Cloud Computing

Big Data is an umbrella term which encompasses all sorts of data which exists today. From hospital records and digital data to the overwhelming amount of government paperwork which is archived – there is more to it than we officially know. It was the emergence of cloud computing which made it easier to provide the best of technology in the most cost effective packages. Cloud computing not only reduced costs, but also made a wide array of applications available to the smaller companies.

4.7 Big Data and e-commerce

In the digital age, the number Smartphone users are on a rise. With such a hike, the market base for e-commerce portals is also increasing. The number of digital buyers is expected to rise from 58.3% in 2016 to around 65% in 2021. With an increase in digital consumers, organizations look to leverage

technologies that can assist them in enhancing customer experience and increasing the number of customers retained. As modern-day companies hold enormous amounts of information, extracting actionable information from this data is crucial. Big data analytics helps companies in extracting information that can assist them in taking the right decisions. Big data in e-commerce holds the ability to support businesses in transforming their operations; here's how:

5. Conclusion

Big Data can play a key role in achieving Sustainable Development through the valuable insights that large groups of data can generate especially for improving urban and transportation planning in cities. The research findings and recommendation of Asia can provide insight in understanding changes in the urban population density and mobility. These findings can help urban planners and policy makers to create more sustainable cities and benefit from the cost savings associated with the new technologies instead of traditional less effective mechanisms to gather these data. The private sector offered access to historical and anonym zed mobile data to LIRNEasia. This was an opportunity to leverage Big Data using private sector's Data Philanthropy for public policy insights. It is also an opportunity for mobile and other companies to draw insight concerning the population they are servicing for commercial and profit making uses.

References

1. ESCAP. "Urbanization Trends in Asia Pacific" (2013) Accessible at http://www.unescapsdd.org/files/documents/SPPSFactsheet-urbanization-v5.pdf, LIRNEasia, Mobile network big data for urban and transportation planning in Colombo, SriLanka"Data for Policy conference, Presenters: Samarajiva,R. Lokanathan, (2015). Accessibleat:http://lirneasia.net/wpcontent/uploads/2013/09/Samarajiva_Cambridge_June15
2. UN GWG for Big Data – UNSD :UN Nation Big Data for Official statistics, 3.

Global hypothesis test to compare the predictive values of two diagnostic tests subject to a case-control study

José Antonio Roldán-Nofuentes[1], Saad Bouh Sidaty-Regad[2]
[1]Biostatistics, School of Medicine, University of Granada, Spain.
[2]Public Health and Epidemiology, School of Medicine, University of Nouakchott, Mauritania.

Abstract

The accuracy of a binary diagnostic test (BDT) is measured in terms of two fundamental parameters: sensitivity and specificity. The sensitivity is the probability of the result of the BDT being positive when the individual has the disease and the specificity is the probability of the result of the BDT being negative when the individual does not have the disease. Other fundamental parameters of a binary diagnostic test are the positive predictive value and the negative predictive value. The predictive values represent the clinical accuracy of the test, and they depend on the sensitivity and specificity of the diagnostic test and on the disease prevalence. The comparison of the predictive values of two binary diagnostic tests is a topic that has been the subject of different studies in the field of Statistics. In this work, we propose a global hypothesis test to compare the predictive values of two binary diagnostic tests subject to a case-control design, assuming for this purpose that there is an estimation of the disease prevalence. This global hypothesis test is based on the chi-squared distribution. The method proposed was applied to a real example.

Keywords

Chi-square distribution; Positive and negative predictive values; Type I binomial bivariate distribution

1. Introduction

The positive predictive value (PPV) is the probability of an individual having the disease when the result of the BDT is positive, and the negative predictive value (NPV) is the probability of an individual not having the disease when the result of the BDT is negative. The predictive values (PVs) represent the accuracy of the diagnostic test when it is applied to a cohort of individuals, and they are measures of the clinical accuracy of the BDT. The PVs depend on the sensitivity (Se) and the specificity (Sp) of the BDT and on the disease prevalence (p), i.e.

$$PPV = \frac{p \times Se}{p \times Se + (1-p) \times (1-Sp)} \quad \text{and} \quad NPV = \frac{(1-p) \times Sp}{p \times (1-Se) + (1-p) \times Sp}.$$

PVs quantify the clinical value of the BDT, since both the individual and the clinician are more interested in knowing how probable it is to have the disease given a BDT result.

The comparison of the performance of two binary diagnostic tests is a topic of special importance in the study of statistical methods for the diagnosis of diseases. This comparison is made through a paired-design or through a case-control design. Paired design consists of applying the two BDTs and the gold standard to all of the individuals in a single sample. Case-control design consists of applying the two BDTs to all of the individuals in two samples, one made up of individuals who have the disease (case sample) and another made up of individuals who do not have the disease (control sample). In this research, we study the comparison of the PVs of two BDTs subject to a case-control design. Subject to this type of design, the two BDTs are applied to all of the individual in two samples, one of n_1 individuals who have the disease (case sample) and another with n_2 individuals who do not have the disease (control sample). In a case-control design, the sample sizes n_1 and n_0 are set by the researcher. The sample of individuals that have the disease is extracted from a population of individuals that have the disease (e.g. registers of diseases), and the control sample is extracted from a population of individuals who are known not to have the disease. As the PVs depend on the disease prevalence and subject to a case-control design the quotient $n_1/(n_1+n_2)$ is not an estimator of the prevalence, in order to estimate and compare the PVs subject to this design it is necessary to have an estimation of the disease prevalence. This estimation can be obtained from health surveys or from previous studies. In Section 2, we study hypothesis tests to jointly and individually compare the PVs of two BDTs subject to case-control study. In Section 3 the results are applied to a real example of the diagnosis of Human African Trypanosomiasis.

2. The model

Let us consider two BDTs, Test 1 and Test 2, which are applied to all of the individuals in two samples, one of n_1 individuals who have the disease (case sample) and another of n_2 individuals who do not have it (control sample). Let T_1 and T_2 be two binary variables that model the results of each BDT, in such a way that $T_i=1$ when the result of the corresponding BDT is positive and $T_i=0$ when it is negative. In Table 1, we can see the probabilities associated to the application of both BDTs to both types of individuals (cases and controls), as well as the frequencies observed.

Table 1. Probabilities and observed frequencies subject to case-control design.

Probabilities							
Case				Control			
	$T_2=1$	$T_2=0$	Total		$T_2=1$	$T_2=0$	Total
$T_1=1$	ξ_{111}	ξ_{110}	Se_1	$T_1=1$	ξ_{211}	ξ_{210}	$1-Sp_1$
$T_1=0$	ξ_{101}	ξ_{100}	$1-Se_1$	$T_1=0$	ξ_{201}	ξ_{200}	Sp_1
Total	Se_2	$1-Se_2$	1	Total	$1-Sp_2$	Sp_2	1
Observed frequencies							
Case				Control			
	$T_2=1$	$T_2=0$	Total		$T_2=1$	$T_2=0$	Total
$T_1=1$	n_{111}	n_{110}	$n_{11\bullet}$	$T_1=1$	n_{211}	n_{210}	$n_{21\bullet}$
$T_1=0$	n_{101}	n_{100}	$n_{10\bullet}$	$T_1=0$	n_{201}	n_{200}	$n_{20\bullet}$
Total	$n_{1\bullet1}$	$n_{1\bullet0}$	n_1	Total	$n_{2\bullet1}$	$n_{2\bullet0}$	n_2

Using the conditional dependence model of Vacek (1987), the probabilities given in the table are written as

$\xi_{1jk}=Se_1^j\left(1-Se_1\right)^{1-j}Se_2^k\left(1-Se_2\right)^{1-k}+\delta_{jk}\varepsilon_1$ and

$\xi_{2jk}=Sp_1^{1-j}\left(1-Sp_1\right)^{j}Sp_2^{1-k}\left(1-Sp_2\right)^{k}+\delta_{jk}\varepsilon_0$

with $j,k=0,1$. The parameter ε_1 (ε_0) is the covariance between the two BDTs in (controls) cases, where $\delta_{jk}=1$ if $j=k$ and $\delta_{jk}=-1$ if $j\neq k$, and it is verified that

$0\le\varepsilon_1\le \text{Min}\left\{Se_1\left(1-Se_2\right),Se_2\left(1-Se_1\right)\right\}$ and

$0\le\varepsilon_0\le \text{Min}\left\{Sp_1\left(1-Sp_2\right),Sp_2\left(1-Sp_1\right)\right\}$. If $\varepsilon_1=\varepsilon_0=0$

then the two BDTs are conditionally independent from the disease status. In practice, the assumption of the conditional independence is not realistic, and therefore $\varepsilon_1>0$ and/or $\varepsilon_0>0$. In terms of the probabilities ξ_{ijk}, the sensitivities are written as

$$Se_1=\xi_{111}+\xi_{110} \text{ and } Se_2=\xi_{111}+\xi_{101},$$

and the specificities as

$$Sp_1=\xi_{201}+\xi_{200} \text{ and } Sp_2=\xi_{210}+\xi_{200}.$$

From the case sample, the estimators are

$$\hat{Se}_1=\frac{n_{11\bullet}}{n_1} \text{ and } \hat{Se}_2=\frac{n_{1\bullet1}}{n_1},$$

and from the control sample, the estimators are

$$\hat{Sp}_1=\frac{n_{20\bullet}}{n_2} \text{ and } \hat{Sp}_2=\frac{n_{2\bullet0}}{n_1},$$

and the estimators of their variances are $\hat{Var}\left(\hat{Se}_1\right)=\hat{Se}_1\left(1-\hat{Se}_1\right)/n_1$, $\hat{Var}\left(\hat{Se}_2\right)=\hat{Se}_2\left(1-\hat{Se}_2\right)/n_1$, $\hat{Var}\left(\hat{Sp}_1\right)=\hat{Sp}_1\left(1-\hat{Sp}_1\right)/n_2$ and $\hat{Var}\left(\hat{Sp}_2\right)=\hat{Sp}_2\left(1-\hat{Sp}_2\right)/n_2$. Therefore, the sensitivities and the specificities are estimated as proportions of marginal totals. In this way, in the case sample we are interested in the marginal frequencies $n_{11\bullet}$ and $n_{1\bullet 1}$, as these frequencies are the product of a type I bivariate binomial distribution (Kocherlakota and Kocherlakota, 1992). In an analogous way, the marginal frequencies $n_{20\bullet}$ and $n_{2\bullet 0}$ of the control sample are the product of a type I bivariate binomial distribution. In the case of individuals with the disease, the type I bivariate binomial distribution is characterized (Kocherlakota and Kocherlakota, 1992) by the two sensitivities (Se_1 and Se_2) and by the correlation coefficient (ρ^+) between T_1 and T_2. In an analogous way, in the case of individuals who do not have the disease the type I bivariate binomial distribution is characterized by Sp_1, Sp_2 and the correlation coefficient (ρ^-) between T_1 and T_2. In the case of the individuals with the disease (cases), the correlation coefficient between the two BDTs is

$$\rho^+=\frac{\xi_{111}-Se_1Se_2}{\sqrt{Se_1\left(1-Se_1\right)Se_2\left(1-Se_2\right)}}=\frac{\varepsilon_1}{\sqrt{Se_1\left(1-Se_1\right)Se_2\left(1-Se_2\right)}},$$

and in the case of the individuals who do not have the disease (controls), the correlation coefficient between the two BDTs is

$$\rho^-=\frac{\xi_{200}-Sp_1Sp_2}{\sqrt{Sp_1\left(1-Sp_1\right)Sp_2\left(1-Sp_2\right)}}=\frac{\varepsilon_0}{\sqrt{Sp_1\left(1-Sp_1\right)Sp_2\left(1-Sp_2\right)}},$$

with

$$\hat{\varepsilon}_1=\frac{n_1n_{111}-n_{11\bullet}n_{1\bullet 1}}{n_1^2} \text{ and } \hat{\varepsilon}_0=\frac{n_2n_{200}-n_{20\bullet}n_{2\bullet 0}}{n_2^2},$$

$$\hat{Cov}\left(\hat{Se}_1,\hat{Se}_2\right)=\left(\hat{\xi}_{111}-\hat{Se}_1\hat{Se}_2\right)/n_1=\hat{\varepsilon}_1/n_1$$

and

$$\hat{Cov}\left(\hat{Sp}_1,\hat{Sp}_2\right)=\left(\hat{\xi}_{200}-\hat{Sp}_1\hat{Sp}_2\right)/n_2=\hat{\varepsilon}_0/n_2 .$$

All of the other covariances are zero, since the two samples are independent. The estimators of ρ^+ and ρ^- are

$$\hat{\rho}^+=\frac{n_1n_{111}-n_{11\bullet}n_{1\bullet 1}}{\sqrt{n_{11\bullet}\left(n_1-n_{11\bullet}\right)n_{1\bullet 1}\left(n_1-n_{1\bullet 1}\right)}} \text{ and } \hat{\rho}^-=\frac{n_2n_{200}-n_{20\bullet}n_{2\bullet 0}}{\sqrt{n_{20\bullet}\left(n_2-n_{20\bullet}\right)n_{2\bullet 0}\left(n_2-n_{2\bullet 0}\right)}}.$$

Assuming that there is an estimation ρ of the disease prevalence, the estimators of the predictive values are

$$\hat{PV}_1=\frac{pn_2n_{11\bullet}}{pn_2n_{11\bullet}+qn_1\left(n_2-n_{20\bullet}\right)} \text{ and } \hat{NPV}_1=\frac{qn_1n_{20\bullet}}{pn_2\left(n_1-n_{11\bullet}\right)+qn_1n_{20\bullet}}$$

for Test 1, and

$$\hat{PPV}_2 = \frac{pn_2 n_{1\bullet 1}}{pn_2 n_{1\bullet 1} + qn_1\left(n_2 - n_{2\bullet 0}\right)} \text{ and } \hat{NPV}_2 = \frac{qn_1 n_{2\bullet 0}}{pn_2\left(n_1 - n_{1\bullet 1}\right) + qn_1 n_{2\bullet 0}}$$

for Test 2, when $q = 1 - p$. Let the variance-covariance matrixes be defined as

$$\sum_{\hat{Se}} = \begin{pmatrix} Var\left(\hat{Se}_1\right) & Cov\left(\hat{Se}_1, \hat{Se}_2\right) \\ Cov\left(\hat{Se}_1, \hat{Se}_2\right) & Var\left(\hat{Se}_2\right) \end{pmatrix} \text{ and } \sum_{\hat{Sp}} = \begin{pmatrix} Var\left(\hat{Sp}_1\right) & Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) \\ Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) & Var\left(\hat{Sp}_2\right) \end{pmatrix}.$$

Let $\boldsymbol{\theta} = \left(Se_1, Se_2, Sp_1, Sp_2\right)^T$ be a vector whose components are the sensitivities and the specificities, and let $\boldsymbol{\omega} = \left(PPV_1, PPV_2, NPV_1, NPV_2\right)^T$ be a vector whose components are the predictive values. The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\sum_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes \sum_{\hat{Se}} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes \sum_{\hat{Sp}},$$

where $\otimes$ is the product of Kronecker. Applying the delta method, the matrix of variances- covariances of $\hat{\boldsymbol{\omega}}$ is

$$\sum_{\hat{\boldsymbol{\omega}}} = \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}}\right) \sum_{\hat{\boldsymbol{\theta}}} \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}}\right)^T.$$

Then, we study the joint comparison and the individual comparison of the PVs of the two BDTs. In both cases, and as has been explained in Section 1, it is assumed that there is an estimation of the disease prevalence based on a health survey or other studies.

Global hypothesis test

The global hypothesis test to simultaneously compare the PVs of the two BDTs is

$$H_0 : \left(PPV_1 = PPV_2 \text{ and } NPV_1 = NPV_2\right) \text{ vs } H_1 : \text{at least one equality is not true,}$$

which is equivalent to the hypothesis test

$$H_0 : \mathbf{A}\boldsymbol{\omega} = \mathbf{0} \text{ vs } H_1 : \mathbf{A}\boldsymbol{\omega} \neq \mathbf{0},$$

where $\mathbf{A}$ is a complete range design matrix and a dimension 2×4, i.e.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & -1 \end{pmatrix}.$$

As the vector $\hat{\boldsymbol{\omega}}$ is distributed asymptotically according to a multivariate normal distribution i.e. $\sqrt{n_1 + n_2}\left(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}\right) \xrightarrow[n_1+n_2 \to \infty]{} N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\right)$, then the statistic for the global hypothesis test is

$$Q^2 = \hat{\boldsymbol{\omega}}^T \mathbf{A}^T \left(\mathbf{A} \hat{\sum}_{\hat{\boldsymbol{\omega}}} \mathbf{A}^T\right)^{-1} \mathbf{A}\hat{\boldsymbol{\omega}},$$

which is distributed asymptotically according to Hotelling's *T*-squared distribution with a dimension 2 and $n_1 + n_2$ degrees of freedom, where 2 is the dimension of the vector $\mathbf{A}\hat{\boldsymbol{\omega}}$. When $n_1 + n_2$ is large, the statistic Q^2 is distributed according to a central chi-square distribution with 2 degrees of freedom when the null hypothesis is true. To be able to calculate the global

test statistic, $Q^2 = \hat{\boldsymbol{\omega}}^T \mathbf{A}^T \left(\mathbf{A} \hat{\Sigma}_{\hat{\boldsymbol{\omega}}} \mathbf{A}^T \right)^{-1} \mathbf{A} \hat{\boldsymbol{\omega}}$, it is necessary for the matrix $\mathbf{A} \hat{\Sigma}_{\hat{\boldsymbol{\omega}}} \mathbf{A}^T$ to be non-singular.

Individual hypothesis tests

The hypothesis test to individually compare the two PPVs (NPVs) is

$$H_0 : PV_1 = PV_2 \text{ vs } H_0 : PV_1 \neq PV_2,$$

where PV is PPV or NPV. Based on the asymptotic normality of the estimators, the statistic for this hypothesis test is

$$z = \frac{\left| PV_1 - PV_2 \right|}{\sqrt{\hat{V}ar\left(PV_1 \right) + \hat{V}ar\left(PV_2 \right) - 2Cov\left(PV_1, PV_2 \right)}},$$

which is distributed asymptotically according to a normal standard distribution, and where the variances-covariances is obtained from the equation

Alternative methods to the global test

The global hypothesis test simultaneously compares the PPVs and the NPVs of the two BDTs. Some alternative methods to this global hypothesis test, based on the individual hypothesis tests, are: 1) Solving the tests $H_0 : PPV_1 = PPV_2$ and $H_0 : NPV_1 = NPV_2$, each one to an error α; 2) Solving the individual tests, $H_0 : PPV_1 = PPV_2$ and $H_0 : NPV_1 = NPV_2$, and applying a multiple comparison method such as Bonferroni's method (1936) or Holm's method (1979), which are methods that are very easy to apply based on the p-values. Bonferroni's method consists of solving each individual hypothesis test to an error $\alpha/2$; and Holm's method is a step-down method which is based on Bonferroni's method but is more conservative.

Simulation experiments were carried out to study the type I errors and the powers of the four methods proposed to solve the global hypothesis test: the hypothesis test based on the chi-square, the individual hypothesis tests each one to an error α, and the individual hypothesis tests applying Bonferroni's method and Holm's method. From the results obtained in these experiments, we propose the following method to compare the PVs of two BDTs subject to a case-control design: 1) Applying the hypothesis test based on the chi-square distribution to an error α, 2) If the global hypothesis test is not significant, the equality hypothesis of the PVs is not rejected; if the global hypothesis test is significant to an error α, the investigation of the causes of the significance is made by solving the individual tests and applying Bonferroni's method or Holm's method to an error α.

3. Example

The results obtained were applied to the study by Matovu et al (2010) on the diagnosis of Human African trypanosomiasis (HAT) in Uganda. HAT, also known as sleeping sickness, is a parasitic disease caused by protozoa belonging to the genus Trypanosoma, and it is transmitted to human beings by a bite from the tsetse fly (genus Glossina) infected by other people or animals that host human pathogenic parasites. In some rural areas of Africa, the disease prevalence may reach 50% in periods of epidemics, and is a significant cause of death. Matovu et al (2010) applied two diagnostic tests to a sample of 75 cases and another sample of 65 controls. For a prevalence value equal to 10%, the estimations of the PVs are $P\hat{P}V_1 = 0.540$, $P\hat{P}V_2 = 0.858$, $N\hat{P}V_1 = 0.978$ and $N\hat{P}V_2 = 0.982$, and the estimated variance and covariance matrix of the estimators of the PVs is

$$\hat{\Sigma}_{\hat{\omega}} = \begin{pmatrix} 0.01158 & 0.00562 & 0.00015 & 0.00005 \\ 0.00562 & 0.01456 & 0.00006 & 0.00006 \\ 0.00015 & 0.00006 & 0.00003 & 0.00001 \\ 0.00005 & 0.00006 & 0.00001 & 0.00002 \end{pmatrix}$$

The value of the test statistic for the test

$H_0 : \left(PPV_1 = PPV_2 \text{ and } NPV_1 = NPV_2\right)$ vs $H_1 :$ at least one equality is not true,

is $Q^2 = 6.954$ $\left(P - value = 0.031\right)$ and therefore null hypothesis of the global test is rejected. Solving the individual hypothesis tests it is found that the value of the test statistic for the $H_0 : PPV_1 = PPV_2$ is equal to 2.606 $\left(\text{two sided p-value} = 0.009\right)$, and that the value of the test statistic for the test $H_0 : NPV_1 = NPV_2$ is equal to 0.886 $\left(\text{two sided p-value} = 0.375\right)$. Applying Bonferroni's (Holm's) method the equality hypothesis of the negative predictive values is not rejected and the equality hypothesis of the two positive predictive values is rejected. The positive predictive value of the *NASBA-OC* test is significantly greater than that of the PCR-OC test (95% CI: 0.079 to 0.558).

Acknowledgements

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P.

References

1. Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3-62.
2. Holm, S. (1979). A simple sequential rejective multiple testing procedure. *Scandinavian Journal of Statistics*, **6**, 65-70.
3. Kocherlakota, S., Kocherlakota, K., 1992. *Bivariate discrete distributions*, Marcel Dekker INC.
4. Matuvo, E,. Mugasa, C.M., Ekangu, R.A., Deborggrave, S., Lubega, G.W., Laurent, T., Schoone, G.J., Schallig, H.D., Büscher, P., 2010. Phase II evaluation of sensitivity and specificity of PCR and NASBA followed by oligochromatography for diagnosis of human African trypanosomiasis in clinical samples from D.R. Congo and Uganda. *PLOS Neglected Tropical Diseases*, **4**, e737.
5. Vacek, P.M., 1985. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, **41**, 959-968.

Women's participation in Brazilian cinema over the last two decades: Evidences based on statistical analysis

Paula Alves de Almeida, Denise Britz do Nascimento Silva, José Eustáquio Diniz Alves, Antonio Etevaldo Teixeira Junior
National School of Statistical Sciences – ENCE/IBGE

Abstract
Cinema prevails in entertainment worldwide and grounds visual patterns that dictate fashion references, behaviours, lifestyles, and the way social representations are constructed. This paper presents an analysis of women's participation in key functions in the crews of Brazilian feature films released between 1996 and 2016. It uses statistical modelling techniques to investigate if there is an association between the sex of film directors and scriptwriters with those of the individuals working in other key functions (such as cinematographer, producer, protagonist) and with other film characteristics (such as genre). The presence of women in preeminent functions in Brazilian films has increased in the last decades, but is still low in comparison with men's participation. The results indicate there is an important relationship between women directors and writers with those in other key behind-the-scenes roles. The probability of a film being directed by a women increases when the film also has female producers and scriptwriters. Similarly, the probability of a film being written by a women increases when the film also has female directors and producers.

Keywords
Gender; Film labour market; Logistic regression

1. Introduction

Cinema carries the ideals and values of the social groups that are in charge of its production, and reflects the relations and hierarchies of the society which it belongs. The low representation of women in cinema would be a reflection and, at the same time, it would reinforce the existing gender inequalities in society.

As in other areas of the labour market, the number of women working in the film business is increasing nowadays, but an important question still to be answered is if women already hold decision-making positions in this area.

According to Martha Lauzen (2019), that develops the annual study *The Celluloid Ceiling*, in 2018 women accounted for 8% of directors working on the top 250 domestic grossing films in the USA, 3 percentage points below from 11% in 2017, and 1 percentage point below the 9% achieved in 1998.

Women fared best as producers (26%), followed by executive producers (21%), editors (21%), writers (16%), directors (8%), and cinematographers (4%).

In Brazil, according to Miranda (1990), in the 1970s and 1980s a strong increase in national film production, especially as a result of government incentives, favoured a noteworthy number of women filmmakers began to work, as demonstrated by Alves (2011). The resumption of Brazilian cinema after the severe crisis in the 1990s was marked by a strong female presence, as highlighted by Ottone (2005). In the 2000s, Brazilian audio-visual production grew again and a substantial number of women made their debut in the direction of feature films.

The involvement of women in key functions in Brazilian film production, such as direction, scriptwriting, production, cinematography, protagonism, has been experiencing impressive growth in the last decades – as demonstrated by Alves et al. (2017) for the period 1961-2010 – but it appears to have stabilized in the last 20 years. For this reason, this paper presents an analysis of women's participation in film direction and other key functions in the crews of Brazilian feature films released between 1996 and 2016. It investigates if there is an association between the sex of film directors and scriptwriters with the sex of protagonists, cinematographers, producers and other film characteristics, such as genre.

2. Methodology

We constructed a database for this work, collecting, organizing and merging information extracted from *ANCINE – Agência Nacional do Cinema* (National Cinema Agency), *Dicionário de filmes brasileiros: longa-metragem* (Silva Neto, 2009), *Dicionário de Cinema Brasileiro* (Baladi, 2013), from the fan websites *Filme B*, IMDB, *AdoroCinema, Academia Brasileira de Cinema* and others, from institutional websites of film production and distribution companies, cinema festivals catalogues, press kits and trailers, and the films itself. This database adds information to the one prepared and analysed by Alves et al. (2017) that included films produced until 2010.

We choose to work with films with length 60 minutes or more. The variables related to protagonists and film genre was classified by the authors. Sex of the behind-the-scenes employees was classified as: female, male or both (when women and men role together the films function). The information about producers, executive producers and production directors was aggregated into the category "producers"; documentary, docudrama or semi-documentary films were classified "documentary".

In order to estimate the probability of a film being directed by a woman (or written by a woman, when we considered as response variable the "scriptwriter is a woman"), we estimated models relating the response variable "director is

a woman" with the following auxiliary variables: sex of scriptwriter, sex of cinematographer, sex of producer, sex of protagonist and film genre.

The film genre "experimental" was aggregated into the category "fiction", and the film genre "animation" was not considered in the modelling procedure because there were very few of them in comparison with fiction and documentary films.

For the modelling procedure, we considered only two categories for the sex of the behind-the-scenes employers, "men" and "women", excluding the category "both", for estimating differential effects for men and women in movie key functions. Since the response variable is binary, we used a binomial logistic regression model that relates the odds in favour of the event (the director is a woman or the scriptwriter is a woman) with possible associated factors (Dobson, 2002), as follows:

$$Ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

p_i is the probability of the director of the i-th film is a woman

X_{ki} $(k = 1,\ldots, K)$ are the auxiliary variables (sex of producers, sex of protagonists, etc.)

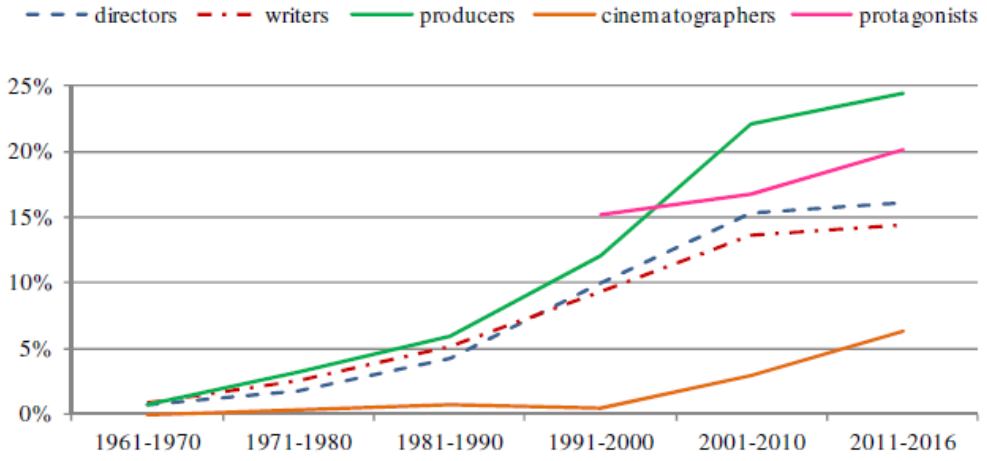
$[\ \beta_0, \beta_1, \beta_2, \ldots, \beta_k\ ]$ is the vector of unknown parameters

3. Main results

Female presence in selected behind-the-scenes roles in Brazilian film production from 1961 to 2016 is presented in Graph 1. We only have the information related with sex of protagonists for films produced after 1991. There is evidence that, for all decades, women's participation in these key functions is very low, although has been increasing over time.

Films with women taking the considered roles markedly increased after 1980, except for cinematographers. The female presence as director, writer and producer portrays a similar pattern. Women fared best as producers (24.5% in the last period, 2011-2016) and worst as cinematographers (only about 6.3% in the last period). It can be noted, when comparing data from 1991-2000 and 2001- 2010, an upsurge of women presence as producers, directors and scriptwriters. Female involvement in film direction seemed to have stabilized after 2001. Therefore, this paper focuses the statistical modeling procedure on data of the last decades.

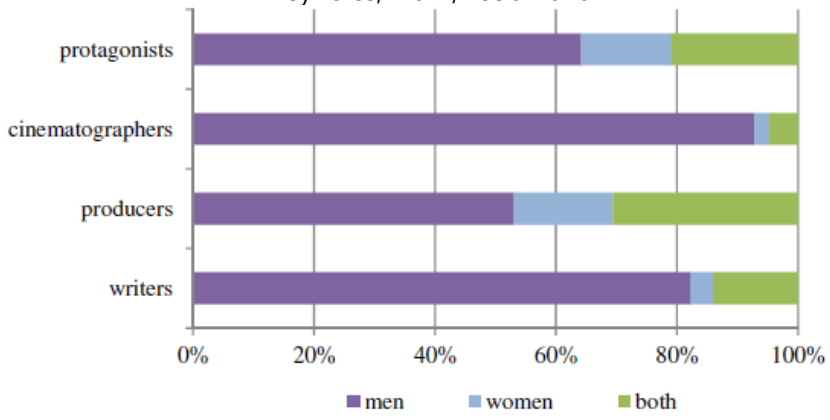
Graph 1 Percentage of feature films with women in key behind-the-scenes roles, by function and decade of production, Brazil, 1961-2016



Sources: ANCINE, 2018; Filme B, 2018; Silva Neto, 2009; Baladi, 2013.

Graphs 2 and 3 show a relationship between the sex of the directors and those in other key roles in Brazilian feature films released in the last decades, between 1996 and 2016.

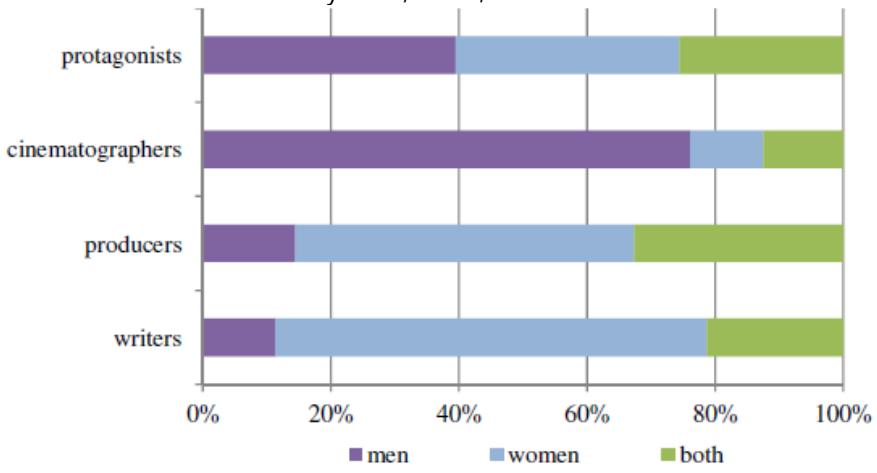
Graph 2 Percentage distribution of Brazilian feature films directed by men by sex of key roles, Brazil, 1996-2016



Sources: ANCINE, 2018; Filme B, 2018; Silva Neto, 2009; Baladi, 2013.
Both – films co-directed by men and women.

Women participation as writers, producers, cinematographers and protagonists is higher in films with women directors. On the other hand, the presence of men is greater working as these same functions in films directed by men. Men are also majority as cinematographers and protagonists in films directed by women. Moreover, the presence of men in all selected key functions in films directed by women is higher than the participation of women in all functions in films directed by men. Besides, there are more films in which the functions are performed by both men and women in films directed by women, than in films directed by men.

Graph 3 Percentage distribution of Brazilian feature films directed by women by sex of key roles, Brazil, 1996-2016



Sources: ANCINE, 2018; Filme B, 2018; Silva Neto, 2009; Baladi, 2013.

For the modelling procedure we considered as response variable "the director is a woman" and as auxiliary variables the following characteristics of films: sex of scriptwriter, sex of cinematographer, sex of producer, sex of protagonist, and film genre. The variables "sex of cinematographer" and "film genre" were not statistically significant in the logistic regression and were excluded from the model. The "sex of protagonist" was significant, but the model with this variable has a smaller value of the pseudo R2 and lower predictive power than the chosen model.

Table 1 indicates that when the scriptwriter is a woman, the odds ratio of the film's being directed by a woman is 115 times that for films with male scriptwriters. In addition, the odds ratio for a film's being directed by a woman is roughly 6.5 times when the producer is a woman then when the producer is a man.

Table 1 Estimated coefficients, corresponding standard errors and odds ratios to "sex of director"

Variables	Coefficient	Standard error	Sig.	Odds ratio
Intercept	-4.54	0.297	<0.001	0.011
Sex of writer (reference = Men)				
Women	4.75	0.308	<0.001	115.463
Sex of producer (reference = Men)				
Women	1.87	0.308	<0.001	6.488

The chosen model presents the best value for goodness-of-fit statistics and predictive power among the fitted models. According to the value of the pseudo R^2, the set of variables included in the model may explain 71% of the variability of the response variable. The model also correctly predicts about 86% of the outcomes when the film is directed by a woman and 96% of the

cases for films directed by a man and 95% of all cases, indicating a good predictive power.

We also considered as response variable "the scriptwriter is a woman" and as auxiliary variables the following characteristics of films: sex of director, sex of cinematographer, sex of producer, sex of protagonist, and film genre. The variables "sex of cinematographer" and "film genre" were not statistically significant in the logistic regression and were excluded from the model. The "sex of protagonist" was statistically significant, but the model with this variable has a smaller value of the pseudo R^2 and lower predictive power than the chosen model. Table 2 indicates that when the director is a woman, the odds ratio of the film has also a female scriptwriter is 115 times that for films with male directors. In addition, the odds ratio for a film's being written by a woman is roughly 2 times when the producer is a woman then for male producers.

Table 2 Estimated coefficients, corresponding standard errors and odds ratios to "sex of writer"

Variables	Coefficient	Standard error	Sig.	Odds ratio
Intercept	-3.43	0.202	<0.001	0.033
Sex of director (reference = Men)				
Women	4.75	0.308	<0.001	115.463
Sex of producer (reference = Men)				
Women	0.68	0.294	0.020	1.979

The chosen model was the one with the best value for the goodness-of-fit statistics and predictive power among the fitted models. According to the value of the pseudo R^2, the set of variables included in the model may explain 65% of the variability of the response variable. The model correctly predicts roughly 79% of the outcomes for films directed by a woman, 98% of the cases for films directed by a man and 95% of all cases.

As a result of the statistical model to investigate the association between the "sex of the director" and the other selected film characteristics, Table 3 displays the estimated probabilities of the occurrence of the event "a film is directed by one or more women" and the event "a film is directed by one or more men" when controlling by the explanatory variables. In addition, based on the statistical model to investigate the association between the "sex of the scriptwriter" and the other selected film characteristics, Table 4 presents the estimated probabilities of the occurrence of the event "a film is written by women" and the event "a film is written by men" according to the explanatory variables.

Table 3 Estimated probabilities to "sex of director" by explanatory variables

Explanatory variables		Estimated probability of	
Sex of writer	Sex of producer	Female director	Male director
Male	Male	0.011	0.989
Male	Female	0.065	0.935
Female	Male	0.552	0.448
Female	Female	0.999	0.001

Table 4 Estimated probabilities to "sex of writer" by explanatory variables

Explanatory variables		Estimated probability of	
Sex of director	Sex of producer	Female writer	Male writer
Male	Male	0.031	0.969
Male	Female	0.060	0.940
Female	Male	0.789	0.211
Female	Female	0.881	0.119

4. Discussion and Conclusion

Cinema expresses the intentions and preconceptions of those involved in its production, mirroring relations and hierarchies of the social contexts. The representations of social groups, gender, sexuality, race and ethnic relations, the issues addressed in films and other decisions are made by directors, producers and screenwriters, functions predominantly controlled by men as highlighted in this paper. The involvement of women in key functions in Brazilian film production, such as direction, scriptwriting, production, cinematography, protagonism, has been experiencing impressive growth in last decades, but is still low in comparison of male participation.

This paper shows that there are gender inequalities in command positions in the Brazilian film industry. Therefore, decisions regarding strategic planning, budget management and, consequently, the choices concerning the representations of men and women with their relations in the work-family environment and the propagation of social values are mostly in male hands.

The statistical model demonstrated an association related to gender issues regarding the key functions in the films considered on the scope of this study. The estimated probability of a film's being directed by a woman increase when the film also has women as producers and writers. The same way, the estimated probability of a film's being written by a woman increase when the film also has women as directors and producers. We did not choose the models with the variable sex of the protagonists, but it was also statistically significant to both response variables, indicating association also between the gender of directors and writers with the gender of the protagonists.

The importance of image and sound in contemporary society is increasing steadily. Controlled by men, it constitutes a potent tool to keep them at the

centre of decision and power. The achievement of real gender equality inevitably leads to an increase in women's presence in command positions in all fields, including, and especially in art, media and culture, alongside with the spread of positive and non-stereotyped representations of women – two inseparable and complementary paths.

References

1. Adoro Cinema. Retrieved from: <<http://www.adorocinema.com/filmes/>>.
2. Agência Nacional do Cinema – ANCINE. Observatório Brasileiro do Cinema e do Audiovisual – OCA. Listagem Completa dos Filmes com os Mecanismos de Incentivo. Retrieved from:<<http://oca.ancine.gov.br/producao_.htm>>.
3. Alves. P. (2011). O cinema brasileiro de 1961 a 2010 pela perspectiva de gênero (Master's thesis, Escola Nacional de Ciências Estatísticas, Rio de Janeiro, Brazil). Retrieved from:<<http://www.ence.ibge.gov.br/images/ence/doc/mestrado/dissertacoes/2011/Dissertacao_2011_Paula_Alves.pdf >>.
4. Alves. P., Teixeira Junior. A. E..& Silva D. B. N. (2017). The use of statistical modelling to analyse women's participation in Brazilian cinema. In: 61st World Statistics Congress - ISI2017, 2017, Marrakesh. Proceedings of the 61st World Statistics Congress - ISI2017.
5. Baladi. M. (2013). Dicionário de Cinema Brasileiro: filmes de longa-metragem produzidos entre 1909 e 2012. São Paulo. Brazil. Martins Fontes.
6. Dobson, A. J. (2002). An introduction to generalized linear models – 2nd edition. London, England, Chapman and Hall.
7. Filme B. Database Brasil. Retrieved from: <<http://www.filmeb.com.br/database>>.
8. Lauzen, M. (2019). The Celluloid Ceiling. San Diego, USA, Women in Film and Television Institute.
9. Miranda, L. F. A. (1990). Dicionário de Cineastas Brasileiros. São Paulo, Brazil, Art Editora.
10. Ottone, G. (2005). Terra Brasil 95-05. El Renacimiento del cine brasileño. Madri, Spain, T&B Editores.
11. Silva Neto, A. L. (2009). Dicionário de filmes brasileiros: longa-metragem – 2ª edição revista e atualizada. São Bernardo do Campo, Brazil, Ed. do Autor.

A method of bias correction when response rate follows linear function

Hee Young Chung, Key-Il Shin
Hankuk University of Foreign Studies, Yongin, Rep. of Korea

Abstract

In recent sample surveys, the accuracy and precision of estimates are decreasing due to non-responses. In particular, there are cases where non-response is affected by the variables of interest and if we apply some commonly used non-response treatment methods to those cases, then we may have bias in estimation. Recently, a method has been proposed to improve the accuracy of estimation by appropriately reducing the bias occurred in the case where the response rate is an exponential function of the variable of interest. In this study, we propose a method to increase the accuracy of estimation when the response rate is a linear function of variable of interest and the distribution of errors included in the super population model follows normal distribution. Simulation results show the superiority of the proposed method. We also suggest the optimal number of substrata that can be used in practice based on the simulation results.

Keywords

linear inclusion probability, sample distribution, regression model, sample weight

1. Introduction

In recent sample surveys, the importance of proper treatment of non-response is increasing. The non-response rate becomes significantly higher, resulting in insufficient number of final survey data, which increases sampling error. Of course, this problem is already well known and several treatments are developed. However, there are some cases where the rate of non-response or response depends on the value of the variable of interest and we need to apply a proper method to those cases. Especially if we have a super population model and a corresponding response rate model like the informative sampling technique, we can calculate the magnitude of bias and so we can correct the bias caused by non-response.

Chung and Shin (2017) studied the case that the super population model is a simple regression model and the response rate model is exponential. They showed that the suggested method improved the accuracy of estimation by correcting the bias. In this paper, we study the case where the response rate is a linear function and the super population model is a simple regression

model. We also suggest the optimal number of substrata that can be used in practice based on the simulation results.

2. Estimation of bias using informative sampling technique

a. Review of the exponential response rate function

i. Review of informative sampling technique

The informative sampling is a sampling design in which there is a sample selection mechanism which is the inclusion probability is influenced by the values of the variable of interest and there is a super population model which is the model between the variable of interest and the auxiliary variable. Pfeffermann et al. (1988) showed that under the informative sampling, we have $f_s(y_i|\theta^*, x_i) = f(y_i|i \in s, x_i) = \frac{\Pr(i \in s|y_i, x_i)\, f_p(y_i|\theta, x_i)}{\Pr(i \in s|x_i)}$ where θ^* is a function of θ. Also with $\Pr(i \in s|y_i, x_i) = E_p(\pi_i|y_i, x_i)$ and $\Pr(i \in s|x_i) = E_p(\pi_i|x_i)$, we have the following relationship,

$$f_s(y_i|x_i) = \frac{E_p(\pi_i|y_i, x_i) f_p(y_i|x_i)}{E_p(\pi_i|x_i)} \tag{1}$$

where $f_p(y_i|x_i)$ is the population distribution, $f_s(y_i|x_i)$ is the sample distribution and $E_p(\pi_i|y_i, x_i)$ is the probability that the datum will be included in the sample when x_i, y_i are given. Whenever $E_p(\pi_i|y_i, x_i) = E_p(\pi_i|x_i)$, the population distribution is the same as the sample distribution. When the super population model is a simple regression model and the response rate is exponential, we have the following results:

$$f_p(y_i|x_i) = N(\beta_0 + \beta_1 x_i, \sigma^2) \tag{2}$$

$$E_p(\pi_i|y_i, x_i) = exp(a_0 + a_1 y_i) \tag{3}$$

where $f_p(y_i|x_i)$ is the population distribution. Now substituting (2) and (3) into equation (1) yields the following sample distribution.

$$f_s(y_i|x_i) = N(\beta_0 + a_1\sigma^2 + \beta_1 x_i, \sigma^2) \tag{4}$$

Therefore, by comparing (2) and (4), we can confirm the $a_1\sigma^2$ of bias.

ii. Parameter estimation of the exponential response rate function

The magnitude of the bias calculated in equation (4) is $a_1\sigma^2$. Since informative sampling uses known a_0, a_1, the magnitude of bias can be calculated by estimating σ^2 form the regression model which is made by the variable of interest and auxiliary variable. However, in the case of the response rate model, a_1 should be estimated. For this, Chung and Shin (2017) estimated a_1 by a method dividing the given stratum or population into equally spaced

substrata. That is, dividing the stratum by auxiliary variable into substrata gives the value of the weight w_i. Also using $E_s(w_i|y_i,x_i) = \frac{1}{E_p(\pi_i|y_i,x_i)}$ and $E_s(w_i|y_i,x_i) \approx w_i$ from Pfeffermann and Sverchkov (2003) and w_i obtained by substrata, we can construct the following model.

$$\log\left(\frac{1}{w_i}\right) = a_0 + a_1 y_i + \eta_i \tag{5}$$

Finally, a_1 can be estimated using (5).

3. Estimation of bias on a linear non-response rate model

a. Sample distribution and bias estimation

In this study we consider the error of the super population model follows the normal distribution and the population distribution is given by (2). Also, we consider the linear response rate model as follows.

$$E_p(\pi_i|y_i,x_i) = b_0 + b_1 y_i \tag{6}$$

Then simply we have the following result.

$$\begin{aligned} E_p(\pi_i|x_i) &= E\left(E_p(\pi_i|y_i,x_i)\right) = E_p(b_0 + b_1 y_i|x_i) \\ &= b_0 + b_1 E_p(y_i|x_i) \end{aligned} \tag{7}$$

Now, using the equations (6) and (7), we get the following result.

$$\frac{E_p(\pi_i|y_i,x_i)}{E_p(\pi_i|x_i)} = \frac{b_0 + b_1 y_i}{b_0 + b_1 E_p(y_i|x_i)} \tag{8}$$

Substituting equation (8) into equation (1), the following result is obtained.

$$f_s(y_i|x_i) = \frac{b_0 + b_1 y_i}{b_0 + b_1 y_i E_p(y_i|x_i)} f_p(y_i|x_i)$$

$$= \frac{b_0}{b_0 + b_1 y_i E_p(y_i|x_i)} f_p(y_i|x_i) + \frac{b_1}{b_0 + b_1 y_i E_p(y_i|x_i)} y_i f_p(y_i|x_i)$$

Now let $f_p^*(y_i|x_i)$ be a distribution of $y_i f_p(y_i|x_i)$. Then it becomes simply the following form.

$$\begin{aligned} f_s(y_i|x_i) = &\frac{b_0}{b_0 + b_1 y_i E_p(y_i|x_i)} f_p(y_i|x_i) \\ &+ \frac{b_1 \mu_i}{b_0 + b_1 y_i E_p(y_i|x_i)} f_p^*(y_i|x_i) \end{aligned} \tag{9}$$

where $\mu_i = E_p(y_i|x_i) = \beta_0 + \beta_1 x_i$. Thus, the sample distribution $f_s(y_i|x_i)$ is a linear combination of population distributions $f_p(y_i|x_i)$ and $f_p^*(y_i|x_i)$. Therefore, we have the following result:

$$E_s(y_i|x_i) = \frac{b_0}{b_0 + b_1\mu_i}\mu_i + \frac{b_1\mu_i}{b_0 + b_1\mu_i}\frac{1}{\mu_i}\left(\mu_i^2 + \sigma^2\right) = \mu_i + \frac{b_1\sigma^2}{b_0 + b_1\mu_i}$$

Now, if we use $\frac{b_1\sigma^2}{b_0 + b_1\mu_i} \approx \frac{b_1\sigma^2}{b_0 + b_1\mu_i^{(s)}}$ to simplify the calculation and let $E_s(y_i|x_i) = \mu_i^{(s)}$, then the corrected estimator is as follows.

$$E_p(y_i|x_i) = \mu_i = \mu_i^{(s)} - \frac{b_1\sigma^2}{b_0 + b_1\mu_i^{(s)}} \tag{10}$$

b. Parameter estimation of linear response rate model

The parameters of linear response rate model are estimated by using the following model similarly used in the exponential response rate parameter estimation.

$$\frac{1}{w_i} = b_0 + b_1 y_i + \eta_i \tag{11}$$

Therefore, b_0, b_1 can be estimated whenever we have the weight w_i and the data y_i obtained from the substrata. Here it is assumed that η_i is independent and identically distributed.

c. The proposed estimator for a given stratum mean

(1) Simple mean estimator

Since the weight of a given stratum is constant, $w_h = w = \frac{N}{n}$ is used and the following equation is obtained for mean estimator

$$\hat{\bar{Y}}_s = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w y_i = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_i = \bar{y} \tag{12}$$

(2) Stratified weighted mean estimator

Since the weights of the substrata are different, the following mean estimator is used.

$$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h y_{hi} \tag{13}$$

(3) Bias corrected estimator

The expected value $\mu_i^{(s)}$ obtained from $\mu_i^{(s)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $\hat{b}_0, \hat{b}_1$ obtained from the response rate model (11) are used. The proposed estimator is as follows.

$$\hat{\bar{Y}}_{inf}^{L} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_h \left(\mu_i^{(s)} - \frac{b_1\sigma^2}{b_0 + b_1\mu_i^{(s)}}\right) \quad (14)$$

4. Simulation

a. Simulation design

In this section, we perform simulation studies in order to confirm the theoretical results and compare the proposed estimator to the existing estimators.

Auxiliary variable x_i in the population are generated with $x_i = 100 + r_i, i = 1, \cdots, N$ where $r_i \sim^{iid} U(0,100)$. Population data is generated by the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim^{iid} N(0,\sigma^2)$. Here we use $\beta_0 = 10, \beta_1 = 5, \sigma^2 = 400$, $N = 10{,}000$. We select n samples by simple random sample from N population data, where $\mathrm{n} = 50, 100, 150, 200, 250, 300, 400, 500$.

Let π_y^{min} be the response rate at the minimum value of y_i and π_y^{max} at the maximum value of y_i. Then we calculate b_0, b_1 by using $(\pi_y^{min}, \pi_y^{max}) = (0.9, 0.7), (0.7, 0.9)$ and given y_i respectively. Finally we can calculate $\pi_i = b_0 + b_1 y_i,\ \pi_i \in [0,1]$ as the response rate applying to the selected n samples. Note that in this study we divide the population or given stratum into L substrata using quantiles based on the auxiliary variable x_i. Finally, aforementioned three estimators are compared using the comparison statistics which are Bias, Absolute bias and Root mean squared error (RMSE) defined by

$$\text{Bias}: \frac{1}{R}\sum_{r=1}^{R}(\hat{\bar{Y}}_r - \bar{Y}_r),\ \text{Abias}: \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\bar{Y}}_r - \bar{Y}_r\right|,\ \text{RMSE}: \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\bar{Y}}_r - \bar{Y}_r)^2}$$

and the number of iteration, R is 3,000.

b. Simulation result

We tabulate the results. Table 1 and Table 2 show that the proposed estimator has better performance based on comparison statistics. Table 3 and Table 4 show that the optimal number of substrata is not affected by the population size, N. Also, Table 5 suggests that the optimal number of sample size in substrata.

i. Results of comparison statistics with uniform distribution

Table 1. Comparison result of $\mathrm{U}(0,100)$ with $\mathrm{n} = 50$

π_y^{min}	π_y^{max}	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$
0.9	0.7	40	4	-8.178	-0.804	-0.661	19.309	5.645	5.603	24.256	8.059	8.019
			5	-8.178	-0.730	-0.608	19.309	5.091	5.012	24.256	9.337	9.283
			6	-8.178	-0.984	-0.880	19.309	5.083	4.975	24.256	12.282	12.211
1	1	50	4	0.167	-0.043	-0.038	16.114	4.870	4.858	20.300	6.089	6.084
			5	0.167	-0.004	-0.017	16.114	4.263	4.207	20.300	6.610	6.558
			6	0.167	0.139	0.118	16.114	3.804	3.721	20.300	5.428	5.323
0.7	0.9	40	4	8.804	0.541	0.382	19.467	5.620	5.594	24.358	8.010	7.981
			5	8.804	0.452	0.276	19.467	5.012	4.929	24.358	8.217	8.136
			6	8.804	0.012	-0.164	19.467	4.826	4.723	24.358	9.774	9.715

Table 2. Comparison result of $\mathrm{U}(0,100)$ with $\mathrm{n} = 500$

π_y^{min}	π_y^{max}	r	L	Bias			Abias			RMSE		
				$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$	$\hat{\bar{Y}}_s$	$\hat{\bar{Y}}_{st}$	$\hat{\bar{Y}}_{inf}^L$
0.9	0.7	400	28	-8.485	-0.147	-0.006	9.322	0.848	0.811	11.065	1.059	1.013
			30	-8.485	-0.150	-0.006	9.322	0.849	0.817	11.065	1.057	1.013
			40	-8.485	-0.165	0.033	9.322	0.863	0.820	11.065	1.305	1.250
1	1	500	28	0.066	0.014	0.009	5.061	0.733	0.714	6.303	0.913	0.892
			30	0.066	0.009	0.002	5.061	0.732	0.716	6.303	0.912	0.892
			40	0.066	0.017	0.008	5.061	0.731	0.700	6.303	0.908	0.870
0.7	0.9	400	28	8.609	0.192	0.033	9.380	0.857	0.805	11.194	1.067	1.011
			30	8.609	0.180	0.027	9.380	0.852	0.805	11.194	1.059	1.007
			40	8.609	0.158	0.002	9.380	0.867	0.810	11.194	1.262	1.208

ii. Optimal number of substrata with various N

Table 3. Result of optimal number of substrata with $\mathrm{U}(0,100)$ and $\mathrm{n} = 50$

π_y^{min}	π_y^{max}	r	N	RMSE	L
0.9	0.7	40	500	6.376	5
			1000	6.710	4
			10000	8.019	4
			50000	6.935	4
1	1	50	500	4.994	5
			1000	4.955	5
			10000	5.323	6
			50000	5.033	5
0.7	0.9	40	500	6.396	5
			1000	6.715	4
			10000	7.981	4
			50000	6.836	4

Table 4. Result of optimal number of substrata with $\mathrm{U}(0,100)$ and $\mathrm{n} = 300$

π_y^{min}	π_y^{max}	r	N	RMSE	L
0.9	0.7	240	5000	1.331	20
			10000	1.375	20
			30000	1.377	20
			50000	1.401	21
1	1	300	5000	1.149	27
			10000	1.185	27
			30000	1.197	27
			50000	1.218	24
0.7	0.9	240	5000	1.333	20
			10000	1.362	20
			30000	1.371	20
			50000	1.395	21

Table 5. Optimal number of substrata with $\mathrm{U}(0,100)$

Population \ r	40	50	80	100	160	200	240	300	400	500
5000	4	6	9	9	12	18	20	21	30	45
10000	4	6	9	10	12	15	20	27	28	40
30000	4	5	9	10	12	20	20	27	28	28

5. Conclusion

When the response rate is a function of the variable of interest, the magnitude of the bias of the non-response can be calculated based on the relationship information. The functions that can be used for the response rate model are various. Especially exponential and linear models are commonly used when the error distribution of the super population model is normal. In this study, the magnitude of the bias caused by non-response is estimated using the linear response rate model. Simulation studies show that the suggested estimator show better performance to estimate the population mean by correcting the bias. Also, through simulation studies, we suggest the optimal number of substrata. Overall, the optimal number of substrata is not significantly affected by the number of populations N. On the other hand, the total sample size affects greatly the results.

References

1. Baillargeon, S. and Rivest L.-P. (2011). The construction of stratified designs in R with the package stratification, Survey Methodology, 37, 53–65.
2. Chung, H. Y. and Shin, K. I. (2017). Estimation using informative sampling technique when response rate follows exponential function of variable of interest, Korean Journal of Applied Statistics, 30, 993–1004.
3. Pfeffermann, D. Krieger, A. M., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling, Statistica Sinica, 8, 1087–1114.

Birth order and birth weight in Uganda: a multilevel analysis of DHS data

Leonard K. Atuhaire
School of Statistics and Planning Makerere University Kampala, Uganda

Abstract

The 'well-known' influence of birth order on birth weight has, to our knowledge, not been established in Uganda. In fact a recent study concluded that birth order had no effect on (low) birth weight. DHS data, with thousands of records that include both the birth order and the birth weight, with matching mother records that include many maternal variables that are known to influence birth weight, provide invaluable material to investigate the influence of birth order on birth weight while adjusting for the effect of maternal factors. However, the nature of DHS data is such that each woman record is associated with several child records. Any analysis that includes both mother and child factors has to take into account this hierarchical structure of the data. A natural approach is to apply multi-level models, with the child as the first level and the mother as the second level. This paper examines the relationship between birth order and birth weight using the 2016 Uganda DHS. A multilevel linear model for birth weight in kilograms and a multilevel binary logistic model for the binary outcome (low birth weight) are fitted. The results show that after adjusting for mother characteristics and other child characteristics, increasing birth order is associated with increasing birth weight as well as reduced incidence of low birth weight. The results also show justification for multilevel modelling.

Keywords

Multilevel linear regression; Multilevel logistic regression; Random effects

1. Introduction

The 'well-known' influence of birth order on birth weight (Seidman et al., 1988; Diamond et al., 2001; Côté et al., 2003) has to our knowledge not been established in Uganda. In fact a recent study (Bayo et al., 2016) concluded that birth order had no effect on (low) birth weight. DHS data, with thousands of records that include both the birth order and the birth weight, with matching mother records that include many maternal variables that are known to influence birth weight, provide invaluable material to investigate the influence of birth order on birth weight while adjusting for the effect of maternal factors.

However, the nature of DHS data is such that each woman record is associated with several child records. Any analysis that includes both mother

and child factors has to take into account this hierarchical structure of the data. A natural approach is to apply multilevel models (Fitzmaurice, Laird, and Ware; 2004), with the child as the first level and the mother as the second level.

This paper uses multilevel models to examine the relationship between birth order and birth weight using the 2016 Uganda DHS data. Studies of the factors influencing birth weight variously use birth weight in kilogrammes (e.g. Diamond et al., 2001; Côté et al.,2003) or low birth weight (<2.5 Kg) (e.g. Gathimba et al., 2017; Ngwira 2015)as the dependent variable. In this paper we use both measures.

2. Methodology

The data used are from the 2016 Uganda Demographic and Health Surveys which collected information on a nationally representative sample of women in child-bearing age (15-49) (Uganda Bureau of Statistics (UBOS) and ICF, 2017). The survey collected a large number of indicators for the respondent, her partner, the household she resides in, and her children who were born within the five years preceding the survey. This study is based on 10,429 children whose weights at birth are available. These belong to 7562 women.

Two models were fitted:

(i) a multilevel linear regression model for birth weight in kilograms,

$$y_{ij} = \mathbf{x}^{t}_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij} \quad (1)$$

$$u_i \sim N(0, \sigma^2_u)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

(ii) a multilevel binary logistic regression model for the binary outcome (low birth weight),

$$\ln\{p(y_{ij}<2.5)/(1-p(y_{ij}<2.5))\} = \mathbf{x}^{t}_{ij}\boldsymbol{\beta} + u_i \quad (2)$$

$$u_i \sim N(0, \sigma^2_u)$$

In both (1) and (2) yij is the weight of the jth child of the ith woman, xtij is a row of covariates, β is the vector of coefficients, ui are the mother level random effects, and εij are the residual errors. Both models were fitted with stata version 13 (StataCorp, 2013).

3. Result

Table 1 below gives the number of children, birth weight and percentage with low birth weight by the child and mother characteristics considered in the study. Some important variables like mother's BMI were not considered because of large numbers of missing values.

Table 1: Number of Children, Birth Weight and Percentage with Low Birth Weight by Child and Mother Characteristics

Mother/Child Characteristic	Number of children	Birth Weight		% with low birth weight
		Mean	St. Dev.	
Overall	10429	3.34	0.85	9.76
Child's birth order				
1	2541	3.21	0.82	12.32
2-3	3585	3.34	0.84	9.18
4-5	2226	3.42	0.85	8.40
6+	2077	3.43	0.89	9.10
Child's sex				
Male	5297	3.41	0.87	8.40
Female	5132	3.27	0.83	11.17
Child is twin				
Yes	322	2.56	0.84	45.03
No	10107	3.37	0.84	8.64
Mother's age at first birth				
<20	6730	3.37	0.88	10.01
20-34	3684	3.30	0.80	9.28
35-49	15	3.12	0.76	13.33
Mother smokes				
Yes	294	3.27	0.79	10.20
No	10135	3.34	0.85	9.75
Mother's education level				
No education	1,227	3.39	0.91	9.37
Incomplete prim.	4,598	3.34	0.88	11.07
Complete prim.	1,487	3.37	0.85	8.81
Incomplete sec.	2,210	3.33	0.82	8.73
Complete sec.	148	3.31	0.72	4.73
Higher	759	3.25	0.68	8.30
Wealth Quintile				
Poorest	2636	3.27	0.85	11.49
Poorer	1983	3.38	0.89	10.29
Middle	1842	3.40	0.89	9.99
Richer	1834	3.37	0.87	8.56
Richest	2134	3.33	0.77	7.97
Place of residence				
Urban	2426	3.30	0.77	8.79
Rural	8003	3.36	0.88	10.08
Region	Not shown			
Marital status				
Ever in union	9920	3.35	0.86	9.64
Never in union	509	3.17	0.77	12.18

Table 2 below shows the results of fitting models (1) and (2) to the Uganda DHS 2016 data.

Table 2: Results of multilevel models

	Multilevel linear regression model for birth weight in Kilogrammes			Multilevel logistic regression model for low birth weight (<2.5 Kg)		
Mother/Child Characteristic	β	se	z	β	Se	Z
Child's birth order	0.027	0.004	6.92	-0.117	0.020	-5.99
Child's sex Male	0.127	0.016	8.18	-0.362	0.079	-4.57
Child is twin	-0.807	0.052	-15.55	2.770	0.190	14.55
Mother's age at first birth	-0.009	0.003	-3.14	-0.001	0.014	-0.10
Mother smokes	-0.019	0.060	-0.32	0.046	0.272	0.17
Mother's education level No education	Ref.					
Incomplete prim.	-0.081	0.033	-2.49	0.060	0.152	0.39
Complete prim.	-0.037	0.039	-0.94	-0.302	0.187	-1.61
Incomplete sec.	-0.060	0.039	-1.56	-0.328	0.183	-1.79
Complete sec. Higher	-0.023	0.083	-0.27	-1.249	0.493	-2.53
	-.079	0.051	-1.57	-0.361	0.244	-1.48
Wealth Quintile						
Poorest	Ref.					
Poorer	0.022	0.029	0.76	-0.101	0.128	-0.79
Middle	0.008	0.031	0.25	-0.093	0.141	-0.66
Richer	-0.018	0.032	-0.55	-0.303	0.151	-2.01
Richest	-0.001	0.039	-0.02	-0.331	0.185	-1.80
Place of residence						
Urban	-0.053	0.026	-2.08	-0.035	0.122	-0.29
Region	Significant regional differences: Not shown			Only one region significantly differs from the other 13.		
Marital status: Ever in union	0.109	0.039	2.77	-0.159	0.178	-0.90
σ2u	0.248	0.012	20.53	1.483	0.311	4.76

Both models (1) and (2) show a very strong influence of birth order on birth weight even after adjusting for mother and other child factors. In summary, from model (1) higher birth weight is associated with higher birth order, being male, not being a twin, lower mother's age at first birth, rural residence, and mother ever having been in union. Mother's smoking status, wealth quintiles do not appear to be associated with birth weight, while the results for education are rather bizarre.

From model (2) low birth weight is associated with lower birth order, being female, being a twin, and not being in the 'richer' quintile. Mothers's age at first birth, mother's smoking status, place of residence, and mother's marital status do not appear to be associated with low birth weight. The results for education are again bizarre.

4. Discussion and Conclusion

We have established that increasing birth order is associated with increasing birth weight as well as reduced incidence of low birth weight. We have also identified the maternal and other child factors that influence birth weight. Our results generally agree with what has been has been observed elsewhere in the sub-Saharan Africa region (Bililign et al., 2018; Gathimba et al., 2017; Kayode et al., 2014; Mwabu, 2008; Ngwira 2015), apart from the effect of urban residence. The significance of the variance of the mother level random effects highlights the importance of using multilevel models in such situations.

References

1. Bayo, L., Buyungo, S., Nakiwala, M., Nabimba, R.., Luyinda E., Nsubuga, T., Namagembe, I., Kasangaki, A. and Banura C. (2016) Prevalence and Factors Associated with Low Birth Weight among Teenage Mothers in New Mulago Hospital: A Cross Sectional Study. J Health Sci (El Monte),4: 192–199.doi: 10.17265/2328-7136/2016.04.003
2. Bililign, N., Legesse, M., Akibu, M. (2018) A Review of Low Birth Weight in Ethiopia: Socio-Demographic and Obstetric Risk Factors. Glob J Res Rev Vol.5 No.1:4
3. Côté, K., Blanchard, R., Lalumière, M.L. (2003) The influence of birth order on birth weight: does the sex of preceding siblings matter? J Biosoc Sci., 35(3):455-62.
4. Diamond, G., Zalzberg, J., Inbar, D., Zvi Cohen, Z., Laks, Y., Geva, D., Grossman, T. and Cohen, H. J.(2001)Birth order, birth weight and later patterns of growth. Ambulatory Child Health 7: 259–267
5. Fitzmaurice, G., Laird, N., Ware, J. (2004) Applied Longitudinal Analysis. New Jersey: John Wiley & Sons, Inc.
6. Gathimba, N. W., Wanjoya, A., Kiplagat, G. K., Mbugua, L., Kibiwott, K. (2017) Modeling Maternal Risk Factors Affecting Low Birth Weight Among Infants in Kenya. American Journal of Theoretical and Applied Statistics. Vol. 6, No. 1, 2017, pp. 22-31.doi: 10.11648/j.ajtas.20170601.13
7. Kayode, G.A., Amoakoh-Coleman, M., Agyepong, I.A., Ansah, E., Grobbee, D.E., Klipstein-Grobusch, K. (2014) Contextual Risk Factors for Low Birth Weight: A Multilevel Analysis. PLoS ONE 9(10): e109333. https://doi.org/10.1371/journal.pone.0109333

8. Mwabu,G. (2008) The Production of Child Health in Kenya: A Structural Model of Birth Weight. Economic Growth Center,Yale University:Center Discussion Paper No. 963
9. Ngwira, A., Stanley, C.C. (2015) Determinants of Low Birth Weight in Malawi: Bayesian Geo-Additive Modelling. PLoS ONE 10(6): e0130057. https://doi.org/10.1371/journal.pone.0130057
10. Seidman, D.S., Ever-Hadani, P., Stevenson, D.K., Slater, P.E., Harlap, S., Gale, R. (1988) Birth order and birth weight reexamined. Obstetrics and Gynecology, 72(2):158-162. (PMID:3260664)
11. StataCorp (2013) Stata Statistical Software: Release 13.1. College Station, TX: Stata Corporation.
12. Uganda Bureau of Statistics (UBOS) and ICF (2017) Uganda Demographic and Health Survey 2016: Key Indicators Report. Kampala, Uganda: UBOS, and Rockville, Maryland, USA: UBOS and ICF.

Research on the digital economic index system and evaluation in Qingdao

Wei Gang, Liu Wei
Qingdao Statistical Institute, China

Abstract

With the rapid development of information and communication technology, digital economy has increasingly become an important driving force for economic and social development and progress. The development of digital economy has been taken by various regionsas an important engine for implementing the new development concept and building a modern economic system. This paper, on the basis of earnestly learning from and reference to the experience and practice of relevant research institutes and advanced municipalities such as Chongqing, aims to define the implication of digital economy, establish statistic standards for digital economy, try to construct a digital economy evaluation index system, and analyze the status quo and existing shortcomings of the digital economy development in Qingdao in a quantitative manner in an effort to provide reference for macro decision-making, formulation of plans to promote digital economy development, acceleration of the conversion from traditional economic development momentum to new economic development momentum, enhancement of high-quality development, and optimization of business environment.

Keywords

Digital economy; Evaluation index system; Macro decision-making; Development plan

1. Introduction

At present, with the improvement of production efficiency and transaction efficiency brought about by the innovation, integration and diffusion of information and communication technology, as well as the continuous emergence of new industries, new formats, new models and new technologies, the digital economy, as a new economic form, has become an important driving force for transformation and upgrading, and a commanding height for a new round of industrial competition.

In 1995, Don Tapscott, a Canadian master of business strategy, published *The Digital Economy*, which gave detailed discourse upon the impact of Internet on economy and society. He is regarded as one of those who first proposed the concept of "digital economy", and is honored as "the father of

digital economy". In 1997, Japan's MITI began to use the term "digital economy". In 1998, U.S. Department of Commerce issued the report *The Emerging Digital Economy*. Since then, it has continued to concern about the "new economy" phenomenon closely related to Internet technology, and has released a number of annual research results under the theme of "digital economy". China also attaches great importance to the promotion on economy and society brought by information and communication technology, but didn't put forward the concept until recent years. In the *Report on the Work of the Government* of 2015, "Internet +" was first proposed, to accelerate further development of "Internet +" by promoting the role of Internet in integration and innovation. Additionally, the requirement to accelerate the growth of digital economy was also put forward therein. In 2018, the central government issued *Outline of Digital Economy Development Strategy*, and Qingdao also issued *Digital Shandong Development Plan (Exposure Draft)*.

A, Digital economy emerged at the end of the 20th Century and obtained competitive development in various countries. Especially in recent years, because of the deep integration of information technology and network technology, digital economy has boomed and become an important engine for the development of "new economy."

(I) Implicationof digital economy

1. Definition. This paper considers that digital economy mainly refers to a series of economic and social activities that take digital knowledge and information as the key production factors, take the innovation of digital technology as the core driving force, take modern information network as the important carrier, to create new industries, continuously improve the digitalization and intellectualization levels of traditional industries, and accelerate the reconstruction of economic and social development and government governance modelsthrough the in-depth integration of digital technology and real economy.
2. Main contents. Digital economy covers digital industrialization and industrial digitalization. Digital industrialization is also known as the basic part of digital economy, which means that the information and communication industry, including electronic information production, telecommunications, software and information technology services, the Internet and related services, need to be digitized. Industrial digitalization is also known as the integrated part of digital economy, which means the increase in production quantity, quality and production efficiency brought about by the application of digital technology in traditional industries, and its new output constitutes an important part of the digital economy, including "digital+" primary industry, "digital+" second industry, and "digital+" tertiary industry,

such as intelligent manufacturing, e-commerce platform, Internet finance, and online retail. Based on the above concepts, we define the statistical standards for digital economy, mainly involving 11 departments, 32 categories, and 188 subcategories.

(II) Characteristics of digital economy

Metcalfe's Law, Moore's Law and Davidow's Law are the three internationally recognized laws on digital economy, and are constantly improving. Under this premise, the following five basic understandings or characteristics can be refined: the construction of information infrastructure such as the Internet is a necessary support for digital economy development; convenience and publicity are important characteristics of the rapid development of digital economy; inclusiveness is the fundamental attribute of digital economy development; accelerating industrial integration is the path of digital economy development; promoting the construction of a modern economic system is one of the purposes of digital economy development.

2. Methodology

At present, National Bureau of Statistics has not yet clarified the statistical standards of digital economy, nor has it established the statistical reporting system for digital economy. All municipal statistical departments are in the stage of exploring and discussing. In order to reflect the digital economy development in Qingdao, we tried to establish the digital economic index system and conducted trial evaluation.

(I) Construction and data source of the evaluation index system for digital economy development

According to the development characteristics of digital economy, we select the following digital economy evaluation indexes by combining with the characteristics and rules of Qingdao's digital economy development:

1. Digital infrastructure: fixed Internet penetration, mobile Internet penetration, number of mobile phone base stations, number of 4G base stations, MAN outlet bandwidth, average speed of fixed broadband ports, and optical cable length per square kilometer.
2. Digital industrialization: the proportion of new generation information technology to industrial output; the proportion of information transmission, software and information technology service industry to GDP; internal expenses in R&D expenditures for computers, communications and other electronic equipment manufacturing industries, and software business income.
3. Industrial digitalization: the index for the integration and development of industrialization and informatization, the growth rate of transaction volume via e-commerce platform, the proportion of

online retail sales to the total retail sales of social consumer goods, the number of smart phones, and Internet finance (the proportion of added value in the finance industry to GDP).

4. Digital government affairs: the number of government websites, the online handling rate of administrative licensing, and the online handling rate of public services.
5. Digital livelihood: digital medical care (person-times of appointments for treatment), total number of social insurance cards, and coverage of digital campus construction.

(II) Evaluation method and weight determination of the digital economy evaluation index system

In this paper, the comprehensive index method is adopted for evaluation. Since the selected digital economic indexes have different contributions to the overall goal, the weight of each individual index needs to be determined before evaluation. The weight is a quantitative value determined by the degree of influence of each evaluation index therein on the evaluation system. Since the influences of evaluation indexes on the overall goal of digital economy development vary, it is very important to determine the weight scientifically and reasonably. At present, there are many ways to determine the weight, which can be roughly divided into two categories: subjective assignment method and objective assignment method.

In this paper, the entropy method in the objective assignment method is adopted to determine the weight. The entropy method is to determine the index weight based on the amount of information provided by the observed value of each index. It is generally believed that the entropy method can profoundly reflect the utility value of the entropy value of index information, and the index weight given thereby has high credibility.

3. Results

This paper adopts empirical analysis and comparative analysis of relevant cities to find out the rules and shortcomings of digital economy development in Qingdao.

(I) Empirical analysis

We analyzed the measured data of Qingdao from 2013 to 2017:

1. Weight analysis of indexes. Among primary indexes, the digital infrastructure had a maximum weight of 0.3027, which had the greatest impact on digital economy, followed by industrial digitalization, digital industrialization, digital government affairs and digital livelihood in sequence. Among secondary indexes, index weights such as the growth rate of transaction volume via e-

commerce platform, the proportion of online retail sales to the total retail sales of social consumer goods, the number of government websites and mobile Internet penetration were relatively great, playing a decisive role in the development of digital economy.

2. Analysis of the overall score. On the whole, the overall digital economy development in Qingdao from 2013 to 2017 showed an accelerating trend. The score of 0.2696 in 2017 reached the highest level in the five years; the lowest score of 0.1472 in 2013 was only about half of that in 2017. From the perspective of key fields, all the five major fields showed positive growth, and the growth rates were accelerating. "Digital industrialization" witnessed the highest development rate, doubling in five years, while "industrial digitalization" had the lowest development rate, with only 50% growth. "Digital infrastructure" accounted for the largest proportion, up to 30.8% of the digital economy in 2017, while "digital livelihood" accounted for the smallest proportion, only 12.6% of the digital economy in 2017. "Digital infrastructure" made the greatest contribution, up to 35.5% of contribution rate in 2017, while "industrial digitalization" made the smallest contribution, only 4.0% of contribution rate in 2017.

(II) Comparison among cities

This paper compares the relevant indexes of some cities in China. Since we can't get all the data of other cities, we only make a comparative analysis of software business income with scientific and technological innovation.

1. Income of software business

As a typical industry of digital industrialization, the software industry is of great significance to reflect the development of digital economy. In terms of the number of enterprises, there were 1,640 software companies in Qingdao in 2017, ranking third in the sub-provincial cities, which was 862 and 96 lower than Wuhan (2,502) and Jinan (1,736) respectively. From the perspective of software business income, the total volume in Qingdao was RMB180.21 billion, ranking the seventh. But it had a large gap to other cities, only accounting for 35% of Shenzhen, 51% of Nanjing, 55% of Hangzhou, and was RMB1.78 billion lower than the average level of sub-provincial cities. In short, the number of software enterprises in Qingdao is large, but the average income is low, and the growth rate of business income continues to decline. The overall development lags behind such cities as Shenzhen and Hangzhou, and shows a certain gap to Nanjing and Chengdu.

2. Scientific and technological innovation

Original innovation is one of the key factors to develop digital economy. Compared with municipalities such as Beijing, Shanghai, Guangzhou, Shenzhen and Hangzhou, Qingdao still has a big gap in scientific and

technological innovation. This paper selects six most representative indexes in scientific and technological innovation to compare with Shenzhen. The results show that indexes such as R&D personnel, internal R&D expenditures, and patent applications in Qingdao are only about 30% of those in Shenzhen, far lower than about 50% of GDP. The investment intensity of scientific research funds in the whole society is about 60% of that in Shenzhen. Compared with some key cities in China, the R&D investment intensity of Qingdao are respectively 2.86%, 0.42%, 0.36% and 0.27% lower than that of Beijing (5.64%), Hangzhou (3.20%), Hefei (3.14%) and Nanjing (3.05%), and 0.65% higher than the national average. So it is necessary to increase the investment intensity.

As a whole, Qingdao's digital economy has achieved certain results, but still faces some bottlenecks. Among them, there is not only a problem of weak foundation in digitalization on the demand side, but also insufficient platform support capacity on the supply side, as well as the need for the improvement of environment. Those are mainly manifested in the following: the institutional mechanism needs to be adjusted and improved; the development of digital inclusion needs to be further improved; the support for the industrial Internet platform is insufficient; the enterprise innovation capability needs to be strengthened; the effort of research on new digital industries and formats is insufficient; talents in the digital economy industry are insufficient; the sharing of data information resources is insufficient, etc.

4. Discussion and Conclusion

In light of the problems in the development of digital economy in Qingdao, this paper puts forward opinions and suggestions for accelerating the orderly development of digital economy in Qingdao.

(I) Take digital economy as the starting point of promoting the conversion from traditional economic development momentum to new economic development momentum, as well as optimizing and upgrading industries, and break through the high-quality development with the efficient flow of "total factor" in the market.

It is recommended that we should take digital economy as the paramount project in Qingdao, and guide various market factors such as labor, capital, knowledge, technology and management with big data, to form an economic and social ecosystem with orderly and free flow of human resources, logistics, information, capital and even the complete supply chain.

(II) Take determining the responsibilities of leading agencies in the digital economy as the key to coordinate, design and lead digital economy development

In view of the urgent need for the adjustment and improvement of institutional mechanism construction, upgrade the municipal power in promoting the coordination leading group for big data development, clarify

the departments, refine the responsibilities, and coordinate the promotion of Qingdao's digital economy work. Enhance the top-level design, study and formulate the development plan of digital economy of Qingdao as soon as possible.

(III) Take broadening the construction of digital economy foundation as the support to narrow the "digital gap" in hardware environment.

In light of the decline in investment growth, the lack of high-tech investment, and the low materialization of good projects in recent years, the construction of information infrastructure should be enhanced to expand network coverage, improve the speed, quality and efficiency and reduce fees to resolve imbalance between urban and rural development and narrow the "digital gap".

(IV) Take enhancing scientific and technological innovation as the strategic guide to optimize the soft environment construction of digital economy development

In response to the lack of Internet platform support, weak enterprise innovation capability and intention, and insufficient scientific research talents, enhance scientific and technological innovation as the strategic guide and innovation investment engine. In addition, it is also necessary to strengthen the safeguard construction for talents, funds, data security and other factors in digital economy, to create a good environment for data development.

(V) Take lowering the digital economy threshold and opening the market as the necessity to widely introduce the competitive and active elements of development

Deepen the reform of streamlining administration, delegating more powers to lower-level governments and society, improving regulation and optimizing services, open the digital market, accurately carry out the "investment attraction & recruitment of talents and introduction of knowledge", increase the recruitment for digital enterprises and talents, stimulate local digital enterprises to participate in competition, and promote the healthy development of digital economy in Qingdao.

(VI) Take sharing the results of digital economy development as a goal, to provide enterprises and people with practical convenience

Based on the principle of sharing, and with non-sharing as an exception, actively promote the data interconnection between departments and enterprises, to improve the efficiency of data development and utilization. Actively take digital means to achieve online operation of government affairs services, improve work efficiency and service efficiency, reduce regulatory costs, and digitize government affairs services and people's livelihood services.

With the advancement of Digital China strategy, Qingdao is actively building an international metropolis. As the digital economy will usher in a new period of development opportunities, Qingdao Municipal Bureau of

Statistics will establish the digital economy statistics monitoring system, clarify the scope of statistics, statistical indexes, statistical standards and calculation methods, improve the evaluation index system, and conduct statistics survey on digital economy by combining the overall development thought and specific action plan of the municipal party committee and the municipal government on the basis of the current connotation of digital economy, so as to reflect Qingdao's development status in digital economy more comprehensively, accurately and scientifically, and promote the high-quality development of Qingdao's economy.

References

1. *The Digital Economy*, published by Don Tapscott, a Canadian master of business strategy, in 1995
2. "Digital economy", first used by Japan's MITI in 1997
3. *The Emerging Digital Economy*, issued by U.S. Department of Commerce in 1998
4. *Report on the Work of the Government*,2015
5. , *Outline of Digital Economy Development Strategy*, issued by China in 2018
6. *Digital Shandong Development Plan (Exposure Draft)*, issued by China Shandong Province in 2018
7. *China Statistical Yearbook*, *Shandong Statistical Yearbook*, *Qingdao Statistical Yearbook* and exchange data.

Multivariate approach to dimension reduction based on the enhanced Scatter Search – Composite I-distance indicator (eSS-CIDI) approach: The case of the Sustainable Society Index (SSI)

Milica Maricic, Veljko Jeremic, Milica Bulajic
University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia

Abstract

Sustainability and sustainable development goals have become a major topic of the world's policy agenda. Nations worldwide are making efforts to create sustainable societies for their citizens and future generations. However, the issue of measuring the level of sustainability emerges. So far, composite indicators have been used with success to provide decision-makers and wider public the information regarding the achieved level of sustainability. Nevertheless, frameworks of such composite indicators can be complex as they incorporate indicators which measure different aspects of sustainability. Therefore, herein we propose the application of the enhanced Scatter Search – Composite I-distance indicator (eSS-CIDI) approach to reduce the number of dimensions of a composite indicator. As a case study we chose the acknowledged Sustainable Society Index (SSI). Our results show that the SSI framework could be modified. The presented approach and obtained results can be a foundation for further research on dimension reduction procedures and composite indicators.

Keywords

Dimension reduction, Composite index, Multivariate analysis, eSS-CIDI approach, Sustainable Society Index

1. Introduction

In the recent years composite indicators have become a valuable source of information for policy makers, decision makers and the wider public (Greco et al., 2018; Saisana et al., 2011). The OECD (2013) defines composite indicators as metrics "formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multi-dimensional concept that is being measured". From this definition various questions arise (Nardo et al., 2005) such as which indicators to include in the framework, whether to normalize the data or no, how to decide on the importance of individual indicators, and so on.

An important obstacle of composite indicators is that they usually aim to measure a multidimensional phenomenon which cannot be measured with a sole indicator (Decancq & Lugo, 2013). Therefore, the issue arises how to

decide on the indicators which will be used in the composite indicator framework. One study direction which has been recently developing is the dimension reduction analysis of composite indicators (Marozzi, 2009). The goal of such analysis is to exclude indicators used to rank entities and thus simplify the composite indicator framework. This study aims to propose a novel hybrid multivariate statistical approach for dimension reduction which also improves the stability of the metric.

We begin the study with a short literature review on dimension reduction techniques which have been used in the field of composite indicators. The Section 3 sees the presentation of the methodologies. We first present in brief the methodological framework of the Sustainable Society Index (SSI), the composite index which was here used as a case study. Next, we introduce the basics of the enhanced Scatter Search – Composite I-Distance Indicator (eSS-CIDI) approach which we used to reduce the dimensionality of the SSI. The obtained results are provided in Section 4, while the concluding remarks are given in the final chapter.

2. Dimension reduction

The issue of dimension reduction is a topic of high interest for researchers, but also for policy makers. So far different approaches have been suggested. Namely, Fodor (2002) in his detailed literature review on dimension reduction listed Principal Component Analysis (PCA), Factor Analysis (FA), Projection pursuit (PP), Independent component analysis (ICA), Non-linear principal component analysis, Random projections and other non-linear methods and extensions as major dimension reduction techniques. Herein, we will place our attention on the methods which have been used in the field of composite indicators.

One of the most common dimension reduction techiniques is the Principal component analysis (PCA) which was initialy proposed by Carl Pearson. The idea behind the PCA is to find a linear combination of variables which accounts for as much variation in the original variables as possible (Tabachnick & Fidell, 2013). The benefits of this analysis have been acknowledged by composite indicator creators. Namely, the OECD Handbook on creating composite indicators suggests to perfrom PCA to define the dimensionality of the composite indicator and to define weights (Nardo et al., 2005). Just one of the examples of researches in the field of composite indicators which employ the PCA are Kotzee & Reyer (2016).

In his research Marozzi (2009) proposed a four-step algorithm for dimension reduction based on the Spearman correlation coefficient. The first step is to create the composite indicator using all indicators and obtain the rank R_k. In the following step h indicators are excluded where $h \in \{1, 2, ..., k\}$,

where k is the number of indicators in the composite indicators. After each exclusion, the rank of entities R_{k-h} is obtained. In the third step the Spearman correlation coefficients between the R_k and all R_{k-h} are obtained. If the Spearman correlation coefficient is close to 1 that indicates that an indicator can be removed. As a stopping rule Marozzi (2009) suggests when the value of Spearman correlation coefficient drops below 0.9 and 0.8.

For example, Markovic et al. (2016) proposed a post hoc I-distance approach to reduce the number of indicators which make the OECD Better life index. Their approach is based on the I-distance method, a multivariate statistical analysis which is able to synthetize indicators without assigning weights (Ivanovic, 1977). The method stands out as attempts to minimize the duplicity of information (Jeremic et al.). The post hoc I-distance undermines the application of I-distance and removal of the least important variable for the ranking process after each iteration. The importance of the variable for the ranking process was measured through the coefficient of determination.

This literature review should indicate that the dimension reduction and framework reduction is a topic of high interest in the field of composite indicators and that statistical methodologies have so far been employed with a lot of success to solve the issue.

3. Methodology

3.1 Sustainable Society Index (SSI)

Sustainable Society Index (SSI) is a multi-layered composite indicator consisted of 21 indicators divided into seven categories which make three dimensions: Human wellbeing, Environmental wellbeing, and Economic wellbeing. In our research, we focused solely on indicator data. The list of indicators which are used to compute the SSI are listed in Table 1. For more information on the description of the indicators please consult the official framework description (SSI, 2018).

The structure of the SSI is based on three pillars of sustainability as suggested in the Brundtland report (1987): Human, Economic, and Environmental pillar. Kaivo-oja et al. (2014) observe that the structure of the SSI is therefore quite conventional. The Joint Research Centre (JRC) conducted an audit of the SSI in 2012 (Saisana & Philippas, 2012) and stated that there is conceptual coherence of the structure of the SSI, that there are few imbalances within categories, that the marginal weights do not differ too much and that the ranking is robust. On the other hand, there are studies which showed that the SSI could be modified. Maricic et al. (2014) used the I-distance method and provided an alternative ranking of the countries based on the SSI indicators which is free from weights. Savic et al. (2016) attempted to revise the number of indicator within the SSI using the post hoc I-distance. They

suggested an eight-indicator structure. These results indicate that further research on the SSI structure and its weighting scheme could be conducted.

The next steps of the framework methodology which should be more closely observed are the aggregation and the weighting approach applied. In the latest edition of the SSI geometric average was used as the aggregation method, while the assigned weights were equal on each level. In our case study, we will observe the hypothetical case when the aggregation method is arithmetic. In such case, it would be valuable to explore the effective weight of each indicator. Namely, the fact that equal weighting is used on each level of the indicator does not imply that all indicators have the same importance (Greco et al., 2018). The effective weights are obtained as the product of weights assigned to the indicator on each level (Nardo et al., 2005). The effective weights of the indicators of the SSI are given in Table 1. As it can be observed, the weights range from 3.70% (for example indicators *Sufficient Food* and *Sufficient to Drink*) to 8.33% (indicators *Organic Farming* and *Genuine Savings*). Therefore, we can conclude that there is difference in the importance of indicators for the ranking process.

Table 1: Weights assigned to indicators on each levels of the SSI and their effective weights

Dimension	Category	Indicator	Weight within indicators (a)	Weight within categories (b)	Weight within dimension (c)	Effective weight $(a \cdot b \cdot c)$
Human wellbeing	Basic needs	Sufficient Food	33.33%	33.33%	33.33%	3.70%
		Sufficient to Drink	33.33%	33.33%	33.33%	3.70%
		Safe Sanitation	33.33%	33.33%	33.33%	3.70%
	Personal development & Health	Education	33.33%	33.33%	33.33%	3.70%
		Healthy Life	33.33%	33.33%	33.33%	3.70%
		Gender Equality	33.33%	33.33%	33.33%	3.70%
	Well-balanced society	Income Distribution	33.33%	33.33%	33.33%	3.70%
		Population Growth	33.33%	33.33%	33.33%	3.70%
		Good Governance	33.33%	33.33%	33.33%	3.70%
Environmental wellbeing	Natural resources	Biodiversity	33.33%	50.00%	33.33%	5.56%
		Renewable Water Resources	33.33%	50.00%	33.33%	5.56%
		Consumption	33.33%	50.00%	33.33%	5.56%
		Energy Use	25.00%	50.00%	33.33%	4.17%

Dimension	Category	Indicator	Weight within indicators (a)	Weight within categories (b)	Weight within dimension (c)	Effective weight $(a \cdot b \cdot c)$
Economic wellbeing	Climate & energy	Energy Savings	25.00%	50.00%	33.33%	4.17%
		Greenhouse Gases	25.00%	50.00%	33.33%	4.17%
		Renewable Energy	25.00%	50.00%	33.33%	4.17%
	Transition	Organic Farming	50.00%	50.00%	33.33%	8.33%
		Genuine Savings	50.00%	50.00%	33.33%	8.33%
	Economy	GDP	33.33%	50.00%	33.33%	5.56%
		Employment	33.33%	50.00%	33.33%	5.56%
		Public Debt	33.33%	50.00%	33.33%	5.56%

3.2 Enhanced Scatter Search – Composite I-distance indicator (eSS-CIDI) approach

To scrutinize the weighting scheme of the SSI and potentially to reduce the number of indicators which are used in its computation we propose the recently devised enhanced Scatter Search – Composite I-distance Indicator (eSS-CIDI) approach (Maricic, 2018). The idea of the approach is to obtain a data-driven weighting scheme which will produce the most stable rankings of entities if the sensitivity analysis is conducted for the weighting scheme. The stability of the ranks is measured using standard deviations of relative contributions (Dobrota et al., 2016; Murias et al., 2008). Relative contribution v_{ie} of an indicator $i, i \in \{1,2,...,k\}$ to the overall composite index of entity $e, e \in \{1,2,...,n\}$ is the percentual share of the weighted indicator in the overall composite index. Therefore, if the contribution of indicator i of all observed entities varies that indicates that the stability of the results and ranks is low (Dobrota et al., 2016; Savic et al., 2016). Accordingly, the goal is to propose a weighting scheme which will minimize the sum of standard deviations of relative contributions of all indicators which are used in the framework.

The eSS-CIDI approach is conducted in three steps. The first step is to conduct the bootstrap CIDI to devise bounds within which the novel weighting scheme will be chosen from. The bootstrap CIDI has already been used to restrain GAR DEA model (Data Envelopment Analysis) (Radojicic et al., 2018). The bootstrap CIDI consists of performing the Composite I-distance Method (CIDI) on *m out of n samples without replacement* where m is the subsample size and n is the sample size. Namely, after each iteration, a novel CIDI weighting scheme will be obtained. As the subsample size we chose $0.632 \cdot n$ as suggested by De Bin and associates (2016), and for the number of bootstrap

iterations we choose 500. The second step is to choose weight bounds. Herein, we chose minimum and maximum bootstrap CIDI weight. Finally, the optimization problem is solved using the enhanced Scatter Search metaheuristics (Egea et al., 2009). For more details regarding the eSS-CIDI please consult Maricic (2018).

The eSS-CIDI is therefore a data-driven approach for devising weights which is based on acknowledged statistical and optimization methods. So far, the eSS-CIDI was used to devise a weighting scheme of a novel composite indicator. Maricic (2018) used the approach in the creation of a European Index of Life Satisfaction (EILS). Herein we aim to extend the application of the method to its use in dimension reduction.

4. Results

The dataset needed for the analysis contained all 21 indicator values for 154 countries for the year 2016. The dataset is available online on the official website of the SSI (Sustainable Society Foundation, 2018). The dataset was originally normalised, so the next step in our analysis was to inspect whether all the indicators correlate positively with the I-distance. Namely, if some indicators correlate negatively with the final value of the I-distance that indicates that their direction should be changed and that reciprocated values of the indicator should be used. In our case, five variables correlated negatively: *Energy Use* (-0.671), *Greenhouse Gasses* (-0.616), *Consumption* (-0.587), *Renewable Energy* (-0.460) and *Public Debt* (-0.028). Therefore, prior to conducting the eSS-CIDI approach we computed the reciprocated values of the above listed indicators.

The first step in the eSS-CIDI algorithm is the bootstrap CIDI. We performed 500 bootstrap replications with the sample size of 97 as $154 \cdot 0.632 = 97.328 \approx 97$. The obtained results are given in Table 2. The bootstrap CIDI intervals suggest to give more importance to indicators *Sufficient Food, Sufficient to Drink, Safe Sanitation, Education, Healthy life, Gender Equality, Population Growth, Good Governance, GDP.* In cases of other indicators, they suggest lower weights or the bootstrap interval covers the official weight. It is of interest to observe that the lower bound of three indicators is 0 - *Renewable Water Resources*, *Employment* and *Public Debt.* This could indicate that these indicators might not be valuable for the ranking process and in some bootstrap samples they were insignificant. In the final step, the optimization model was solved using eSS.

The optimal weighting scheme suggested by the eSS-CIDI approach is given in Table 2 and it provides interesting insights. Firstly, several indicators have been awarded with visibly higher weights than the official ones (e.g. *Good Governance* from 3.70% to 8.2%), while some were given visibly lower weights (e.g. *Genuine Savings* from 8.33% to 2.2%). Importantly, the approach

suggested to exclude two indicators: *Renewable Water Resources* and *Employment.* The provided results show that the framework of the SSI can be simplified. The exclusion of these two indicators can be interpreted two-ways. First, that the remaining indicators completely cover the information provided by the two which are therefore redundant, and second, that the two indicators decrease the stability of the SSI.

The exclusion of the two indicators is a notable result, as the algorithm could have excluded all three indicators whose lower bound was 0. Namely, *Public Debt* was given the weight 2.2% when it could have been given weight 0. The stability of the SSI is also improved. If the official weighting scheme is used the sum of standard deviations of relative contributions is 0.34210233 while when the optimized weighting scheme is used it is 0.259332032.

Table 2: Effective weights of the indicators of SSI, min and max bootstrap CIDI weights, and the optimal weight suggested by the eSS-CIDI approach

Indicator	Effective weight	$w_{\min i}$	$w_{\max i}$	w_i
Sufficient Food	3.70%	4.8%	6.0%	6.0%
Sufficient to Drink	3.70%	5.7%	6.7%	6.7%
Safe Sanitation	3.70%	5.9%	6.9%	5.9%
Education	3.70%	6.5%	7.5%	7.5%
Healthy Life	3.70%	6.7%	7.7%	7.7%
Gender Equality	3.70%	4.4%	6.6%	6.6%
Income Distribution	3.70%	2.1%	4.6%	2.1%
Population Growth	3.70%	4.3%	6.1%	4.3%
Good Governance	3.70%	7.1%	8.2%	8.2%
Biodiversity	5.56%	2.2%	4.3%	2.2%
Renewable Water Resources	5.56%	0.0%	2.3%	0.0%
Consumption	5.56%	4.3%	6.3%	6.3%
Energy Use	4.17%	4.0%	6.2%	6.2%
Energy Savings	4.17%	2.3%	4.9%	2.3%
Greenhouse Gases	4.17%	3.3%	5.7%	5.7%
Renewable Energy	4.17%	2.1%	5.5%	3.7%
Organic Farming	8.33%	5.2%	7.0%	5.2%
Genuine Savings	8.33%	2.2%	4.2%	2.2%
GDP	5.56%	7.0%	7.9%	7.0%
Employment	5.56%	0.0%	1.6%	0.0%
Public Debt	5.56%	0.0%	4.0%	4.0%

5. Discussion and Conclusion

The issue of dimension reduction has been attracting the attention of researchers of different expertise. In the field of composite indicators, it is an important step in the validation or scrutinization of the final metric. Dimension

reduction has several benefits. First, policymakers and composite index users will be given a less complex theoretical framework. Second, the index creators might speed up the data collection process as less data is needed to be acquired. Finally, a complex structure does not guarantee that the final composite index will effectively measure the desired phenomenon. In some cases, adding indicators decreases the quality of the metric (Van der Maaten et al., 2009).

Herein we proposed the application on the novel hybrid weighting approach, the eSS-CIDI, to devise a novel weighting scheme and to reduce the number of indicators within a composite indicator. As a case study, we scrutinized the acknowledged Sustainable Society Index (SSI). The results indicated that two indicators can be excluded from the SSI framework to simplify its structure and to improve the stability of the composite indicator. The results also show that the eSS-CIDI can be successfully used in the process of dimension reduction. The future directions of the study could be two-fold. One direction could be the modification of the eSS-CIDI algorithm (different approach to choosing subsample size, number of resamples, or weight bounds). The other direction could be towards the inclusion of expert opinion in defining final weight bounds.

We hope that the presented approach and the obtained results can be a foundation for further research on dimension reduction procedures and composite indicators.

References

1. Brundtland, G. (1987). *Our Common Future*.
2. De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A. L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, *72*(1), 272–280. https://doi.org/10.1111/biom.12381
3. Decancq, K., & Lugo, M. A. (2013). Weights in Multidimensional Indices of Wellbeing: An Overview. *Econometric Reviews*, *32*(1), 7–34. https://doi.org/10.1080/07474938.2012.690641
4. Dobrota, M., Bulajic, M., Bornmann, L., & Jeremic, V. (2016). A new approach to the QS university ranking using the composite I-distance indicator: Uncertainty and sensitivity analyses. *Journal of the Association for Information Science and Technology*, *67*(1), 200–211. https://doi.org/10.1002/asi.23355
5. Egea, J. A., Balsa-Canto, E., García, M.-S. G., & Banga, J. R. (2009). Dynamic Optimization of Nonlinear Processes with an Enhanced Scatter Search Method. *Industrial & Engineering Chemistry Research*, *48*(9), 4388–4401. https://doi.org/10.1021/ie801717t
6. Fodor, I. (2002). *A survey of dimension reduction techniques*.
7. Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2018). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Social Indicators Research*, 1–34. https://doi.org/10.1007/s11205-017-1832-9
8. Ivanovic, B. (1977). *Teorija klasifikacije*. Institut za ekonomiku industrije, Beograd.
9. Jeremic, V., Bulajic, M., Martic, M., & Radojicic, Z. (2011). A fresh approach to evaluating the academic ranking of world universities. *Scientometrics*, *87*(3), 587–596. https://doi.org/10.1007/s11192-011-0361-6
10. Kaivo-oja, J., Panula-Ontto, J., Vehmas, J., & Luukkanen, J. (2014). Relationships of the dimensions of sustainability as measured by the sustainable society index framework. *International Journal of Sustainable Development & World Ecology*, *21*(1), 39–45. https://doi.org/10.1080/13504509.2013.860056
11. Kotzee, I., & Reyers, B. (2016). Piloting a social-ecological index for measuring flood resilience: A composite index approach. *Ecological Indicators*, *60*, 45–53. https://doi.org/10.1016/j.ecolind.2015.06.018
12. Maricic, M. (2018). Assessing the quality of life in the European Union: The European Index of Life Satisfaction (EILS). *Statistical Journal of the IAOS*, 1–7. https://doi.org/10.3233/SJI-180481
13. Maricic, M., Jankovic, M., & Jeremic, V. (2014). Towards a framework for evaluating Sustainable Society Index. *Romanian Statistical Review*, *62*(3), 49–62.

14. Marković, M., Zdravković, S., Mitrović, M., & Radojičić, A. (2016). An Iterative Multivariate Post Hoc I-Distance Approach in Evaluating OECD Better Life Index. *Social Indicators Research*, *126*(1), 1–19. https://doi.org/10.1007/s11205-015-0879-8
15. Marozzi, M. (2009). A composite indicator dimension reduction procedure with application to university student satisfaction. *Statistica Neerlandica*, *63*(3), 258–268. https://doi.org/10.1111/j.1467-9574.2009.00422.x
16. Murias, P., de Miguel, J. C., & Rodríguez, D. (2008). A Composite Indicator for University Quality Assesment: The Case of Spanish Higher Education System. *Social Indicators Research*, *89*(1), 129–146. https://doi.org/10.1007/s11205-007-9226-z
17. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on constructing composite indicators*. *OECD Statistics Working Papers*. https://doi.org/10.1787/533411815016
18. OECD. (2013). Glossary ofstatistical terms. Retrieved November 9, 2018, from https://stats.oecd.org/glossary/detail.asp?ID=6278
19. Radojicic, M., Savic, G., & Jeremic, V. (2018). Measuring The Efficiency Of Banks: The Bootstrapped I-Distance Gar Dea Approach. *Technological and Economic Development of Economy*, *24*(4), 1581–1605. https://doi.org/10.3846/tede.2018.3699
20. Saisana, M., D'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, *40*(1), 165–177. https://doi.org/10.1016/j.respol.2010.09.003
21. Saisana, M., & Philippas, D. (2012). *Sustainable Society Index (SSI): Taking societies' pulse along social, environmental and economic issues*.
22. Savic, D., Jeremic, V., & Petrovic, N. (2016). Rebuilding the Pillars of Sustainable Society Index: A Multivariate Post Hoc I-Distance Approach. *Problemy Ekorozwoju – Problems of Sustainable Development*, *12*(1), 125–134.
23. Sustainable Society Foundation. (2018). Sustainable Society Index. Retrieved November 20, 2018, from http://www.ssfindex.com/
24. Tabachnick, B., & Fidell, L. (2013). *Using Multivariate Statistics, 6th Edition*. Pearson.

On complex seasonal SSA based forecasting

Winita Sulandari[1,2]; Subanar[1]; Suhartono[3]; Herni Utami[1]; Muhammad Hisyam Lee[4]

[1]Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]Universitas Sebelas Maret, Surakarta, Indonesia
[3]Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[4]Universiti Teknologi Malaysia, Johor, Malaysia

Abstract

Modifications in the forecasting model based on SSA have been an interesting subject among researchers in recent years. SSA is considered as a powerful method in decomposing complex time series and a flexible model can be constructed based on SSA decomposition. In this paper, a wellknown monthly accidental deaths in USA is used as the experimental study. We compare the performance of the hybrid SSA-model for the deaths series with other methods proposed in the literature. The results show that the hybrid SSA-NN model yields better forecasting accuracy than other methods discussed in the literature.

Keywords

SSA; Time series; Accidental; Hybrid; NN

1. Introduction

The discussion of singular spectrum analysis (SSA) and its application has been conducted by researchers from various fields (see Afshar & Bigdeli, 2011; Hassani, Soofi, & Zhigljavsky, 2010; Hassani, Webster, Silva, & Heravi, 2015; Mahmoudvand, Konstantinides, & Rodrigues, 2017; Nong, 2012; Suhartono et al., 2018). This method can be helpful in extracting and identifying the trend, harmonics and noise of a series (Hassani, 2007). SSA with linear recurrent formula (LRF) discussed by Golyandina & Korobeynikov (2014) is more appropriate for handling deterministic forecasting problem. Recently, modelling each component of SSA by stochastic models and then combining the results can be an alternative to SSA-LRF (Liu, Zhang, & Zhang, 2015).

By employing the hybrid method, the weakness of one method can be handled by another method such that the accuracy performance of the forecasting values usually improves. Neural network (NN) is a powerful method in solving complex forecasting problems (G. P. Zhang & Qi, 2005). Tseng, Yu, & Tzeng, (2002) has successfully combined NN with ARIMA and applied this hybrid method to forecast seasonal time series data.

In this study, we consider that the combination between SSA and NN can be a powerful method in handling complex series. The capability of SSA in extracting the deterministic component and the capability of NN in capturing the nonlinearity and uncertainty in the data can improve the accuracy performance of the forecast values. Sulandari, Subanar, Suhartono, & Utami (2017) have provided examples of successful application of this method to the trend and seasonal time series.

In this paper, we present the methodology of hybrid SSA-NN and apply the method to the well-known monthly accidental deaths series. We compare the hybrid SSA-NN with other methods in literature in term of forecast accuracy.

2. Methodology

The hybrid SSA-NN is a method that consists three steps in modelling a complex series. A brief discussion on the methodology of the hybrid SSA-NN method is presented below. Assumed that the original series $\{Y_t, t = 1, 2, \dots, N\}$ is divided into two parts. The first part is for the training data set that consists of N_{train} observations and the second one is the testing data set that consist of N_{tes} observations, where $N_{tes} = N - N_{train}$.

Step 1: obtaining the trend and harmonics components by SSA decomposition

In decomposing the series, SSA has two stages, decomposition and reconstruction. In decomposition step, we need to set a certain positive integer value of window length (L) that usually proportional to the period of the original series but less than or equal to $N/2$ (see Golyandina, 2010). The original series $\{Y_t, t = 1, 2, \dots, N_{train}\}$ is decomposed via its trajectory matrix using singular value decomposition method. In the second steps, we obtained several groups of matrices that are separable each other and do the reconstruction to transform them into several separable series. The strength of the separability between components can be measured by weighted correlation values. How to find the values was discussed in Elsner & Tsonis (1996), Golyandina & Zhigljavsky (2013), and Golyandina, Nekrutkin, & Zhigljavsky (2001).

Step 2: obtaining the deterministic function for the trend and harmonics

Consider that the original series is decomposed into m components, including the trend, harmonics, and noise. SSA decomposition help us in identifying the deterministic function of the hybrid SSA-NN model, especially for defining the trend and harmonic function. In general, the deterministic function can be written as

$$D_t = \sum_{i=1}^{n_c} C_t^{(i)} + \varepsilon_t$$

where $C_t^{(i)} = f_i(t)g_i(t)$, and ε_t is the noise. In this case, we consider that $f_i(t)$ is the polynomial function and $g_i(t)$ is the sinusoid function where $f_i(t) = \sum_{j=0}^{n_i} a_{ij}\, t^j$ and $g_i(t) = \gamma_i + \alpha_i \cos(\frac{2\pi}{P} t) + \beta_i \sin(\frac{2\pi}{P} t)$. Notation $n_c = m - 1$ and n_i presents the total number of trend and harmonic components and the order of the polynomial function, respectively. When the series shows a trend pattern then the function $g_i(t) = 1$. The parameters of polynomial function can be estimated by ordinary least square (OLS) method while the parameters of sinusoid function can be estimated by iterative OLS method as in Sulandari, Subanar, Suhartono, & Utami (2018). When $f_i(t)$ and $g_i(t)$ are present at the same time then the sinusoid will show a time-varying amplitude.

Step 3: defining the hybrid SSA-NN model

We need to determine the irregular component $\{I_t,\ t = 1, 2, \dots, N_{train}\}$ by subtracting the original series with the forecast values obtained from deterministic function, that is

$$I_t = Y_t - \hat{D}_t = Y_t - \sum_{i=1}^{n_c} \hat{C}_t^{(i)}.$$

where $\hat{C}_t^{(i)}$ is the predicted value for the ith component at time t. The irregular series I_t is then approximated by NN model with n_{input} inputs, i.e. (I_{t-1}, I_{t-2}, … , $I_{t-ninput}$), n_{hidden} hidden nodes and one output. Several number of inputs (n_{input}) and the number of hidden nodes (n_{hidden}) are combined and the best approximation is the NN that yields the smallest RMSE. We use a tangent sigmoid function for all the hidden nodes and purelin for the output node as the activation function. Here, the network is trained by Levenberg-Marquardt training algorithm. References related to the study of NN for time series forecasting can be found in Adhikari & Agrawal (2012), Saini & Soni (2002), Zhang, Patuwo, & Hu (1998), and Zhang, Patuwo, & Hu, (2001).

3. Results

A well-known monthly accidental death in USA displayed in Figure 1 is used to illustrate the application of the hybrid SSA-NN method. In this study we evaluate the performance of the hybrid SSA-NN by comparing the results with those discussed in the previous literature. Brockwell and Davis (2000) have discussed the implementation of SARIMA model, ARAR model, and Holt-Winter algorithm to the death series. The same series were also discussed in Hassani (2007) to show the capability of SSA in extracting the series into several components and forecasting. As in the previous literature, we use the first seventy two observations (January 1973 to December 1978) as the training data set and the last six observations (January 1979 to June 1979) as the testing data. The calculations and the figures in this works are obtained by Matlab 2015.

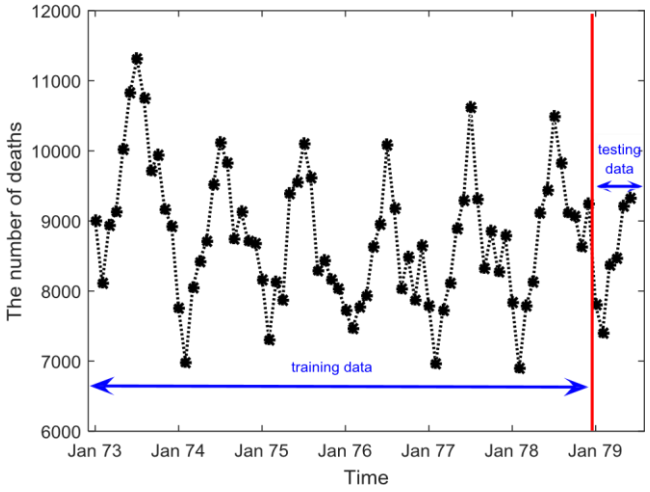


Figure 1: The monthly accidental deaths in USA for January 1973 – June 1979

The first step is decomposing the series into several components. As in Hassani (2007), the window length L is set to be 24, proportional to the period of the series. Based on the w-correlation matrix shown in Figure 2, we separate the series into four components: a trend, two harmonics, and a noise. Trend is reconstructed from the first eigentriple while the first harmonics is obtained by grouping the second and the third eigentriple and the second harmonic is reconstructed from the fourth and fifth eigentripe. The rest is the noise.

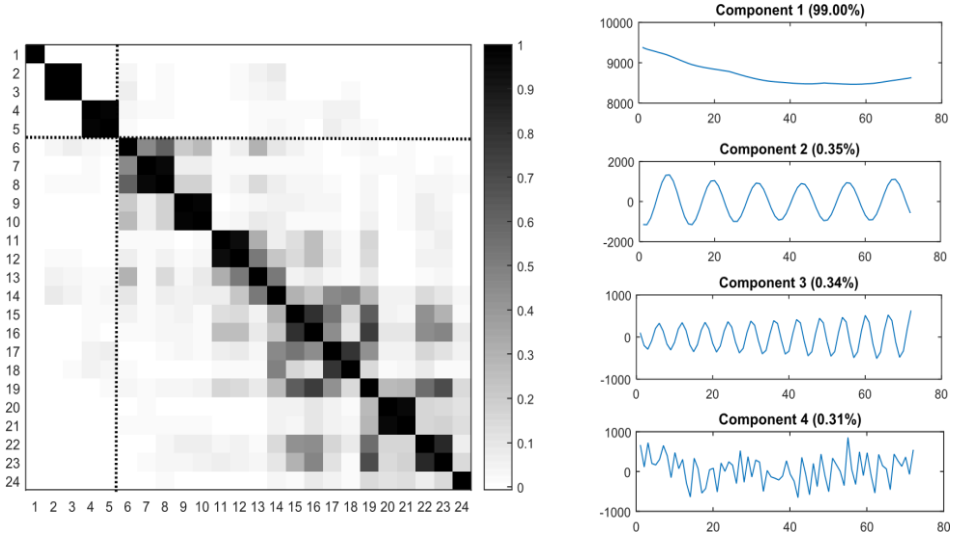


Figure 2: The w-correlations matrix for the accidental death series with L = 24 (left) and the components (right)

The second step is modelling each component trend and harmonics by using function of time. In this case we find that quadratic function is the best one for approximating the trend series. The function is

$$\hat{C}_t^{(1)} = 9439.14 - 38.00\, t + 0.37\, t^2$$

where $\hat{C}_t^{(1)}$ is the forecast value at time t for the first component. The second and the third component can be approximated by oscillatory function with time-varying amplitude. The best function for the second component is quadratic amplitude modulated sinusoid function and can be represented as

$$\hat{C}_t^{(2)} = 30.78 - 0.62t + 0.01t^2 - 1067.95\cos\left(\frac{2\pi}{11.98}t\right) - 1116.13\sin\left(\frac{2\pi}{11.98}t\right) + 21.66t\cos\left(\frac{2\pi}{11.98}t\right) + 22.63t\sin\left(\frac{2\pi}{11.98}t\right) - 0.25t^2\cos\left(\frac{2\pi}{11.98}t\right) - 0.27t^2\sin\left(\frac{2\pi}{11.98}t\right)$$

where $\hat{C}_t^{(2)}$ is the forecast value at time t for the second component. The third component tend to have linear amplitude so that the best function for the series is linear amplitude modulated sinusoid function that can be written as

$$\hat{C}_t^{(3)} = 3.34 + 0.04t + 295.74\cos\left(\frac{2\pi}{6.03}t\right) - 14.82\sin\left(\frac{2\pi}{6.03}t\right) + 3.43t\cos\left(\frac{2\pi}{6.03}t\right) - 0.17t\sin\left(\frac{2\pi}{6.03}t\right)$$

where $\hat{C}_t^{(3)}$ is the forecast value at time t for the third component. At last, the deterministic function for the accidental death series can be obtained from those three functions and can be written as

$$\begin{aligned}\widehat{D}_t = \sum_{i=1}^{3}\hat{C}_t^{(i)} &= 9473.26 - 38.58t + 0.38t^3 - 1067.95\cos\left(\frac{2\pi}{11.98}t\right) - 1116.13\sin\left(\frac{2\pi}{11.98}t\right) \\ &+ 21.66t\cos\left(\frac{2\pi}{11.98}t\right) + 22.63t\sin\left(\frac{2\pi}{11.98}t\right) - 0.25t^2\cos\left(\frac{2\pi}{11.98}t\right) \\ &- 0.27t^2\sin\left(\frac{2\pi}{11.98}t\right) + 295.74\cos\left(\frac{2\pi}{6.03}t\right) - 14.82\sin\left(\frac{2\pi}{6.03}t\right) + 3.43t\cos\left(\frac{2\pi}{6.03}t\right) \\ &- 0.17t\sin\ \frac{2\pi}{6\,03}t\ .\end{aligned}$$

In the third step, we can define the irregular component (I_t) by subtracting the original series with the forecast value for the deterministic component, that is

$$I_t = Y_t - \widehat{D}_t.$$

The irregular series is then approximated by NN method. NN method is considered to handle the uncertainty and the nonlinearity in the data. In this work we have trained the network by combination of a certain input nodes and a number of nodes in hidden layer vary from 1 to 10. We choose six and twelve nodes for the input corresponding to the period 12 and 6. For this case, network with six input nodes and eight nodes in the hidden layer, denoted by NN(6-8-1), produces the smallest root mean square error (RMSE) among the networks whose residuals are random.

In order to evaluate the performance of the models, we use four measures. The four measures are RMSE, mean absolute error (MAE), mean absolute performance error (MAPE), and mean realative absolute error (MRAE). Comparison results for the forecasting accuracy for period January 1979 to June 1979 with those results presented in (Brockwell & Davis, 2002) and (Hassani, 2007) are resumed in Table 1 and Figure 3.

Table 1: Comparison of RMSEs, MAEs, MAPEs, and MRAEs for the testing data of accidental death obtained by hybrid SSA-NN and other methods in literature

Model	RMSE	MAE	MAPE	MRAE
Naïve	789.63	624.00	7.61%	1
ARIMA(0,1,1)(0,1,1)$_{12}$ (Brockwell and Davis, 2002)	582.67	523.50	6.11%	2.72
Subset ARIMA(0,1,[1,6,12,13])(0,1,0)$_{12}$ (Brockwell and Davis, 2002)	500.50	415.00	4.81%	2.22
Seasonal Holt-Winter (Brockwell and Davis, 2002)	401.26	351.33	4.23%	1.27
ARAR (Brockwell and Davis, 2002)	253.20	226.50	2.77%	0.77
SSA-LRF (Hassani, 2007)	278.20	179.67	2.13%	0.81
Hybrid SSA-NN(6-8-1)	16.74	11.32	0.14%	0.06

Based on Table 1 and Figure 3 (bottom) we can see that the hybrid SSA-NN produce the smallest RMSE, MAE, MAPE, and MRAE among others. Furthermore, Figure 3 (top) shows that the forecasting values for January 1979 until June 1979 obtained by SSA-NN red line) are the closest one to the actual values (black line).

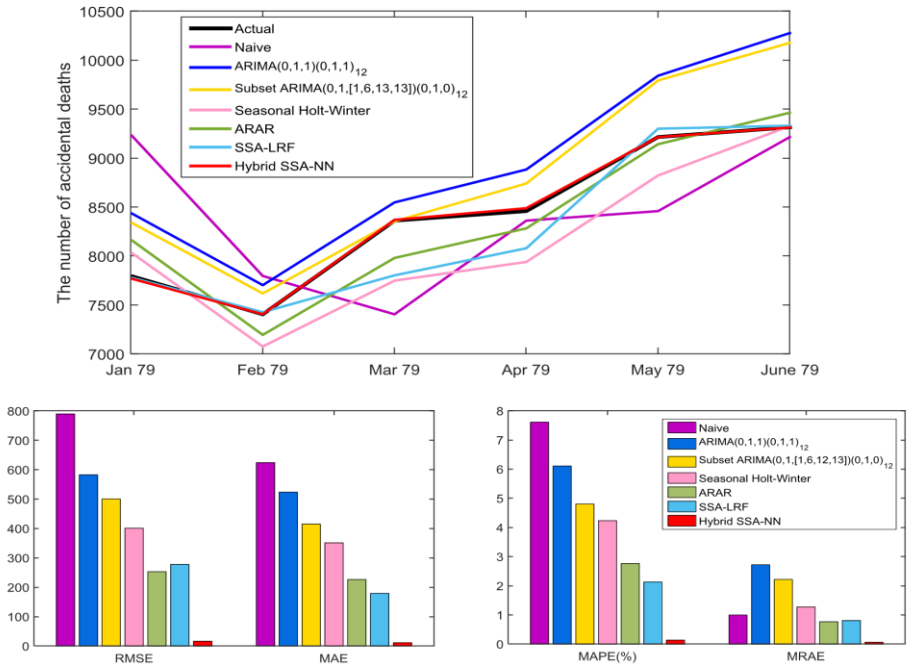


Figure 3: Comparison of actual and the forecast values obtained by several methods (top) and comparison of the forecast accuracy between them (bottom).

4. Discussion and Conclusion

The combination between SSA and NN on monthly accidental deaths series produces more accurate forecast values rather than Naïve, SARIMA, Subset ARIMA, ARAR, and SSA-LRF. The SSA method decomposed the series into four components, where the first three components can be approximated by the deterministic functions. We model each component individually and then combine them. In this step, we need to be careful in identifying and determining the best fit function for the component series. By this technique,

we get more flexible model since the functions can be adjusted to the behaviour of the series.

In the next step, we apply NN to model the irregular component. NN is a powerful method in handling the nonlinearity and uncertainty found in the series. We observe several nodes in the hidden layer varying from 1 to 10, and combine with a number of inputs that proportional to the period. NN with the smallest RMSE and random residuals will be the chosen one.

Furthermore, results show that the hybrid SSA-NN(6-8-1) yields the best performance in comparison with other methods in the mentioned literature. Its MAE and MAPE are even smaller than those obtained by the hybrid method based on local linear neuro-fuzzy model and optimized singular spectrum analysis, named OSSA-LLNF (see Abdollahzade, Miranian, Hassani, & Iranmanesh, 2015). The methodology discussed in this paper can be applied in other cases.

References

1. Abdollahzade, M., Miranian, A., Hassani, H., & Iranmanesh, H. (2015). A new hybrid enhanced local linear neuro-fuzzy model based on the optimized singular spectrum analysis and its application for nonlinear and chaotic time series forecasting. *Information Sciences*, *295*, 107–125.
2. Adhikari, R., & Agrawal, R. K. (2012). Forecasting strong seasonal time series with artificial neural networks. *Journal of Scientific and Industrial Research*, *71*(October 2012), 657–666.
3. Afshar, K., & Bigdeli, N. (2011). Data analysis and short term load forecasting in Iran electricity market using singular spectral analysis (SSA). *Energy*, *36*(5), 2620–2627.
4. Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (2nd ed.). Springer-Verlag.
5. Elsner, J. B., & Tsonis, A. A. (1996). *Singular Spectrum Analysis A New Tool in Time Series Analysis*. Springer Science & Business Media.
6. Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Stat Interface*, *3*(3), 259–279.
7. Golyandina, N., & Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*, *71*, 934–954.
8. Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and related techniques* (Vol. 90). Chapman & Hall/CRC, Boca Raton, FL.
9. Golyandina, N., & Zhigljavsky, A. (2013). *Singular Spectrum Analysis for time series*. Springer Science & Business Media.
10. Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science*, *5*, 239–257.
11. Hassani, H., Soofi, A. S., & Zhigljavsky, A. (2010). Predicting Daily Exchange Rate with Singular Spectrum Analysis Data. *Nonlinear Analysis: Real World Applications*, *11*(3), 2023– 2034.
12. Hassani, H., Webster, A., Silva, E. S., & Heravi, S. (2015). Forecasting US tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, *46*, 322–335.
13. Liu, G., Zhang, D., & Zhang, T. (2015). Software reliability forecasting: singular spectrum analysis and ARIMA hybrid model. In *Theoretical Aspects of Software Engineering (TASE), 2015 International Symposium on* (pp. 111–118). IEEE.
14. Mahmoudvand, R., Konstantinides, D., & Rodrigues, P. C. (2017). Forecasting mortality rate by multivariate singular spectrum analysis. *Applied Stochastic Models in Business and Industry*, *33*(6), 717–732. https://doi.org/10.1002/asmb.2274

15. Nong, J. (2012). An Ensemble Technique to Daily Rainfall Forecasting Based on SSA. In *Computational Sciences and Optimization (CSO), 2012 Fifth International Joint Conference on* (pp. 5–9). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/6274666/
16. Saini, L. M., & Soni, M. K. (2002). Artificial neural network based peak load forecasting using Levenberg–Marquardt and quasi-Newton methods. *IEE Proceedings-Generation, Transmission and Distribution*, *149*(5), 578–584.
17. Suhartono, S., Isnawati, S., Salehah, N. A., Prastyo, D. D., Kuswanto, H., & Lee, M. H. (2018). Hybrid SSA-TSR-ARIMA for water demand forecasting. *International Journal of Advances in*
18. *Intelligent Informatics*, *4*(3), 238–250. https://doi.org/10.26555/ijain.v4i3.275
19. Sulandari, W., Subanar, S., Suhartono, S., & Utami, H. (2017). Forecasting Time Series with
20. Trend and Seasonal Patterns Based on SSA. In *2017 3rd International Conference on Science in Information Technology (ICSITech) "Theory and Applicattion of IT for Education, Industry and Society in Big Data Era"* (pp. 694–699). Bandung, Indonesia: IEEE.
21. Sulandari, W., Subanar, S., Suhartono, S., & Utami, H. (2018). An Empirical Study of Error Evaluation in Trend and Multiple Seasonal Time Series Forecasting Based on SSA. *Pakistan Journal of Statistics and Operation Research*, *14*(4), 945–960.
22. Tseng, F.-M., Yu, H.-C., & Tzeng, G.-H. (2002). Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change*, *69*(1), 71–87.
23. Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62.
24. Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (2001). A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers & Operations Research*, *28*(4), 381–396.
25. Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*(2), 501–514

Fighting innumeracy with TV

Jo Røislien[1,2]
[1]Faculty of Health Sciences, University of Stavanger, Norway
[2]Department of Research, Norwegian Air Ambulance Foundation, Norway

Abstract

Introduction Stories of statistical misconceptions in the public domain are plentiful, and statistics is generally considered a difficult topic to disseminate in understandable layman's terms. General interest in STEM subjects is in decline worldwide, yet statistical literacy has never been more important in a world increasingly fuelled by statistical analyses of quantitative data. In an attempt to change the public's view on maths and stats a large-scale mass communication project funded by national Norwegian broadcaster NRK was initiated in Norway in 2009.

Methodology TV production company Teddy TV teamed up a young stats professor with a director of international music videos and commercials. Through innovative use of everyday objects, contemporary high-end cinematography and pop-cultural know-how the team invented a series of visual demonstrations explaining basic and advanced topics from maths and stats. The end result was ten half hour prime time TV-episodes aimed at the general public.

Results The series premiered fall 2011 on national broadcaster NRK to massive ratings, enthralling more than 600,000 viewers weekly, in a country of only 5 million people. The series rocketed host Prof Jo Røislien into the Norwegian mainstream, as a celebrity alongside rockstars and actors. The TV series has since ended up in classrooms throughout Scandinavia as part of the regular maths and stats education. Clips from the series has hundreds of thousands of views on YouTube, and has been the blueprint for Scandinavian science communication for years.

Discussion and conclusion Film are an underexplored medium for explaining statistics. As scientists have limited know-how of large scale public communication, teaming statisticians up with professional communicators with pop-cultural know-how opens up for more effective communication of important basic statistical competence to large audiences.

Keywords

Mass communication, attention, TV, music videos, pop-culture

1. Introduction

Stories of statistical misconceptions in the public domain are plentiful, and statistics is generally considered a difficult subject to grasp, and a difficult topic to disseminate in understandable terms to laymen outside the scientific community.

Statistical literacy has never been more important in a world increasingly fueled by and run off statistical analyses of quantitative data. At the same time general interest in STEM subjects is in steep decline worldwide. Attention span among the public appears to be decreasing (1), and people spend increasingly more time in front of screens, browsing the internet for content. And film is taking center stage in what people consume online. It is estimated that by 2021 80% of all material consumed online will video (2). Even when watching longer film sequences and drama series attention span is still in the seconds. Psychological research into how people who watch TV and film has unveiled a high demand for continuous gratification to avoid people changing channels.

In an attempt to help change the public's view on maths and stats in Norway a large-scale mass communication project was initiated by commercial Scandinavian TV production company Teddy TV in Norway 2009. The project was eventually given the green light and funded by national broadcaster NRK. The aim of the project was to turn mathematics and statistics into binge-worthy television for a large general audience.

2. Methodology

The television producers teamed up young mathematician and professor of medical statistics Jo Røislien, and director of music videos and commercials Christian Holm-Glad, the latter with international mega-popstars like Calvin Harris and Kygo to his name, alongside films for international brands like Apple and Netflix. Journalists and commercial text writers were added to the team, alongside high-end cinematographers and film editors.

Through innovative use of everyday objects and real-life settings, contemporary high-end cinematography and pop-cultural knowhow, the team invented a series of visually experiments and demonstrations explaining both basic and advanced topics from maths and stats. The spoken language was, just like the imagery, contemporary and non-academic, using only words well-known to everyone. The end result of this science communication experiment was ten half hour TV-episodes aimed at the general public, on prime time national television.

3. Results

The result of the profession-crossing collaboration was a maths and stats series free of blackboards, books, universities and expert interviews. The series premiered fall 2011 on national Norwegian broadcaster NRK to massive

ratings, enthralling more than 600,000 viewers weekly, in a country of only 5 million people, not including reruns and streaming.

A central part of the series was the numerous previously unseen and creative ways of presenting and demonstrating various topics from statistics. The difference between people's perception of what randomness should look like and actual randomness was demonstrated by throwing hundreds of bright yellow rubber ducks into a frozen diving pool (figure 1) highlighting the difference between 'evenly distributed' and 'unpredictable'. The topic of polls was demonstrated using plastic Playmobil figurines and a cement mixer as a randomizing machine (figure 2), and the fundamental idea of regression analyses explained in a 60 second animation short, likening it to "walking the dog" (figure 3). The animation made the rounds on internet, ending up in among others the opinion pages of The New York Times (3). The animation has since been copied by National Geographic Channel in their 2014 remake of *Cosmos* hosted by internationally renowned science communicator Neil deGrasse Tyson.

The maths and stats series rocketed host Prof Jo Røislien into the Norwegian mainstream, becoming a household name appearing alongside rockstars, actors and other celebrities on talk shows and other TV and radio shows, and magazine interviews. The TV series was sold to Sweden, Finland and Denmark, and has since ended up in classrooms throughout Scandinavia, as well as scientific conferences worldwide. Clips from the series has hundreds of thousands of views on YouTube, and has been the blueprint for Scandinavian science communication for years.

4. Discussion and Conclusion

Film is an underexplored medium for explaining statistics. In the current media landscape where film is the dominating form of communication in the public domain this cannot be ignored. Through well thought-out visuals even complicated topics from statistics and quantitative research methodology can be turned into engaging films, teaching both school kids and the general public how to read and interpret quantitative information and statistical analyses.

There is nothing people care more about than other people. This easy-to-grasp phenomenon is often presented as a problem when discussing the dissemination of statistics, which deals with populations and groups rather than individuals. However, the need for repetition, the need for $n>1$ in order for something to even *be* statistics, can easily be turned into an asset rather than a hindrance. In the rubber ducks demonstration of randomness (Figure 1), the Playmobil figurines for demonstrating polls (Figure 2) and the animation short on regression analyses (Figure 3) the need for $n>1$ is actually at the core of what drives the film sequences.

The maths and stats series was an eye-opener for Scandinavian science communicators in general and teachers, lecturers and communicators of maths and stats in particular. Røislien received in 2013 the
Swedish award 'Statistics Promoter of the Year' alongside the famous Hans Rossling.

Prof Røislien and director Holm-Glad have since experimented further with using pop cultural short films for communicating maths and stats, even using highly commercial outlets, creating among others a promotional film for telecom operator Telia based on scientific, quantitative research into the association between music and physiology and psychology (Figure 4), in order to help push quantitative scientific thinking into the mainstream.

Experts in popcultural communication allows for a level of outreach scientists are unable to match, having only a limited know-how of how to perform mass communication on a large scale to the public. By allowing professional communicator with contemporary pop-cultural competence and know-how into the communication of statistics it is possible to reach out to – and teach – the general public stats topics which the stats community is woefully underskilled to do on its own. Teaming statisticians up with professional communicators opens up for more effective communication of important basic statistical competence on a large scale

References

1. McSpadden, K (2015, May 14). You now have a shorter attention span than a goldfish [Web log post]. Retrieved January 28, 2019, from http://time.com/3858309/attention-spans-goldfish/
2. Cisco (2017, June 8). Cisco visual networking index predicts global annual IP traffic to exceed three zettabytes by 2021 [Web log post]. Retrieved January 28, 2019, from https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1853168 3. Revkin, AC (2012, January 11). Can better communication of climate science cut climate risks? [Web log post]. Retrieved January 28, 2019, from https://dotearth.blogs.nytimes.com/2012/01/11/canbetter-communication-of-climate-science-cut-climate-risks/
3. Voldheim D (2017, June 18). TELIA – music freedom [Web log post]. Retrieved January 28, 2019, from https://vimeo.com/222101890

A spatial rank-based EWMA chart for monitoring linear profiles

Longcheen Huwang, Jian-Chi Lin, and Li-Wei Lin
Institute of Statistics National Tsing Hua University Hsinchu, Taiwan

Abstract

Profile monitoring has been recently considered as one of the most promising areas of research in statistical process monitoring (SPM). It is a technique for monitoring the stability of a functional relationship between a dependent variable and one or more independent variables over time. The monitoring of linear profiles is the most popular one because the relationship between the dependent variable and the independent variables is easy to describe by linearity, in addition to its flexibility and simplicity. Furthermore, almost all existing charting schemes for monitoring linear profiles assume that error terms are normally distributed. In some applications, however, the normality assumption of error terms is not justified. This makes the existing charting schemes not only inappropriate but also less efficient for monitoring linear profiles. In this article, based on the spatial rank-based regression, we propose a charting method for monitoring linear profiles where the error terms are not normally distributed. The charting scheme applies the exponentially weighted moving average (EWMA) to the spatial rank of the vector of the Wilcoxon-type rank-based estimators of regression coefficients and a transformed error variance estimator. Performance properties of the proposed charting scheme are evaluated and compared with an existing charting method based on multivariate sign in terms of the in-control (IC) and out-of-control (OC) average run length (ARL). Finally, a real example is used to demonstrate the applicability and implementation of the proposed charting scheme.

Keywords

Average run length; Out-of-control; Profile Monitoring; Spatial rank EWMA; Wilcoxon rank estimators.

1. Introduction

As the progress in sensing and information technologies, automated quality data collection has been commonly used in many manufacturing industries. Consequently, SPM based on large amounts of quality data has become more and more important. Sometimes, the quality of a process can be best characterized by a relationship between a dependent variable and one or more independent variables and this relationship is called a profile. SPM for changes of profile is called profile monitoring. The methods of profile

monitoring, in general, can be divided into parametric and non-parametric approaches. In the parametric approach, the dependent and independent variables are assumed to satisfy a known, either linear or non-linear, model and then the charting statistics are developed based on the estimated model parameters from the profile data to monitor whether the functional relationship has changed or not. On the other hand, in the non-parametric approach, the relationship between the dependent and independent variables is not assumed to be known. Some nonparametric methodologies are employed to estimate such a relationship and thus the charting statistics are constructed using the estimated relationship to monitor the stability of the unknown functional relationship.

For monitoring simple linear profiles, there are many studies in the literature. See, for example, Kang and Albin (2000), Kim, Mahmoud, and Woodall (2003), Mahmoud and Woodall (2004), Zou, Zhang, and Wang (2006), Gupta, Montgomery, and Woodall (2006), Mahmoud, Parker, Woodall, and Hawkins (2007), Zou, Zhou, Wang, and Tsung (2007), among several others. Multiple linear profile monitoring has been studied by Zou, Tsung, and Wang (2007), Mahmoud (2008), Eyvazian, Noorossana, Saghaei, and Amiri (2011), Zou, Ning, and Tsung (2012), Huwang, Wang, Xue, and Zou (2014), Amiri, Zou, and Doroudyan (2014), Kazemzadeh, Amiri, and Kouhestani (2016), Zhang, Shang, Gao, and Wang (2017), and Ghashghaei and Amiri
(2017). Although monitoring linear profiles is an important task, in many practical applications profiles cannot be adequately represented by linear models. Non-linear profile monitoring has been investigated by Walker and Wright (2002), Woodall, Spitzner, Montgomery, and Gupta (2004), Williams, Woodall, and Birch (2007), Colosimo and Pacella (2007), Williams, Birch, Woodall, and Ferry (2007), Yu, Zou, and Wang (2012), Maleki, Amiri, and Taheriyoun (2017), Maleki, Amiri, Taheriyoun, and Castagliola (2017), Esmaeeli, Sadegheih, Amiri, and Doroudyan (2017), Maleki, Castagliola, Amiri, and Khoo (2018), Fotuhi, Amiri, and Maleki (2018), Menafoglio, Grasso, Secchi, and Colosimo (2018), and Khosravi and Amiri (2018). There are many works in the literature that aim to monitor generalized linear model-based regression profiles. See, for example, Yeh, Huwang, and Li (2009), Shang, Tsung, and Zou (2011), Koosha and Amiri (2013), Shadman, Mahlooji, Yeh, and Zou, (2015), Amiri, Koosha, Azhdari, and Wang (2015), Amiri, Yeh, and Asgari (2016), Qi, Wang, Zi, and Li (2016), and Izadbakhsh, Noorossana, and Niaki (2018). Recently, monitoring profiles based on non-parametric regression models has been developed by Zou, Tsung, and Wang (2008), Qiu, and Zou (2010), Qiu, Zou, and Wang (2010).

In this talk, we focus on monitoring profiles which can be represented by linear models. Our study will concentrate on the on-line phase II monitoring. Traditionally, studies on monitoring linear profiles assume that the error terms

of the profiles are normally distributed which is reasonable for most of situations. For example, two different approaches to monitor linear profiles where the error terms are assumed to be normally distributed have been proposed by Zou, Tsung, and Wang (2007) and Huwang, Wang, Xue, and Zou (2014), individually. However, in many applications, the error terms of linear profiles do not follow normal distributions. The non-normality of the error terms makes the charting schemes, that automatically assume the normality of the error terms, inappropriate and inefficient for monitoring linear profiles.

In the non-parametric multivariate SPM, the fact that the performance of traditional control charts, which perform well for monitoring mean vector and/or covariance matrix under normal assumption, has been greatly affected when the process distributions are not normal has been investigated by Qiu and Hawkins (2001), Qiu (2008), Zhou, Zou, Zhang, and Wang (2009), Zou and Tsung (2011), and Li, Zou, Wang, and Huwang (2013). Various non-parametric control charts for monitoring the mean vector and/or the covariance matrix of non-normal processes have also been developed by these authors at the same time. However, based on our knowledge, researches on monitoring linear profiles under the situation that the error distribution is not normal are limited. A distribution-free robust method which uses a rank-based regression for monitoring linear profiles under non-normal assumption of error terms has been proposed by Zi, Zou, and Tsung (2012). Firstly, the so-called Wilcoxon-type rank-based estimators were used to estimate regression coefficients and then the multivariate sign EWMA method was applied to these Wilcoxon-type rank-based estimators to develop their charting scheme. In addition, based on the multivariate EWMA method to the trimmed least squares estimators of regression coefficients, control charts for monitoring linear profiles when the error terms have contaminated normal distributions have been investigated by Huwang, Wang, and Shen (2014). In this talk, the aforementioned Wilcoxon-type rank-based estimators of regression coefficients and a transformation of the error variance estimator will be adopted. Then, the multivariate EWMA method to the spatial rank of the vector of these Wilcoxon-type rank-based estimators and the transformed error variance estimator will be applied to develop the proposed charting scheme. Since the spatial rank extracts more information from multivariate data than the multivariate sign, it is expected that the proposed control chart is more effective than the multivariate sign chart for monitoring linear profiles when the error terms do not follow normal distributions.

In many applications, to collect a large number of IC profiles from Phase I study may not be available. As a result, the charting scheme based on the Phase I data may not have its actual (true) IC ARL (denoted by ARL_0) equal to the nominal ARL_0, and this causes the problem that it is difficult to have a fair

effective comparison among different charting schemes. The question that how many IC Phase I profiles are needed to make the charting scheme based on them to have the actual ARL_0 equal to the nominal level has been raised. To partially answer this question, a commonly accepted criterion will be employed to find the minimum number of IC Phase I profiles that causes the charting scheme based on these profiles to have its actual ARL_0 close to the nominal level. The rest of the talk is organized as follows. A brief introduction of spatial rank-based EWMA charts for monitoring multivariate data is given in Section 2. The proposed method that uses the spatial rank-based regression for monitoring linear profiles is developed in Section 3. The performance comparisons for the proposed multivariate EWMA chart using the spatial rank and the competing counterpart based on the multivariate sign for monitoring linear profiles are provided in Section 4. An example used to illustrate the applicability and implementation of the proposed chart is presented in Section 5. Discussions and conclusions are included in the last section.

Likelihood ratio tests for Lorenz Dominance

Mike S. Chang; Michelle Liou; Philip E. Cheng
Institute of Statistical Science Academia Sinica, Taipei, Taiwan, Republic of China

Abstract

The notion of Lorenz dominance (LD) between two Lorenz curves (LC; Lorenz, 1905) is useful for evaluating welfare redistribution toward decreasing the inequality or ranking income distributions based on expected utility. Early studies of estimation and testing for the LD property were mostly discussed based on a grid of finite points on the unit interval. Extending the inference from a finite grid to the entire unit interval, consistent tests for the LD hypothesis were recently developed upon functions of the empirical Lorenz processes. The asymptotic distributions of the test statistics depend on the unknown distributions and theoretical test *p*-values were empirically assessed using the bootstrap method. In view of the intersection (once) property of a pair of crossing Lorenz curves discussed in the 1980s, elementary properties of crossing and dominant Lorenz curves will be examined in this study. It is found that distinct patterns of the difference curve between a pair of curves can be used to characterize the LD and the crossing Lorenz curves (CLC) conditions through inequalities of their quantile functions. We will use these patterns to construct likelihood ratio (LR) tests tailored for the LD and the CLC hypotheses, respectively. The proposed LR tests are consistent with respect to the standard test levels and critical regions based on approximate chi-square distributions. A simulation study is conducted using Log-normal, Pareto and Weibull distributions, and reliable performance of the LR tests are obtained. An exceptional difficult case is found under the LD hypothesis when two lognormal distributions with close standard deviation parameters are rather close to each other, for which larger sample sizes are useful for making a test decision. An empirical study is carried out for certain pairs of real GDP annual data of 133 countries from "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988".

Keywords

Crossing Lorenz curves; Likelihood ratio test; Lorenz dominance

1. Introduction

The Lorenz curve provides a graph of overall income proportions shared by cumulative proportions of people acquiring incomes from low to high levels. The notion of Lorenz dominance (LD) between two curves (one curve is

everywhere above the other) was useful for evaluating welfare redistribution toward decreasing the inequality through a progressive transfer, or for ranking income distributions according to expected utility (Atkinson, 1970). Inference for the LD property has been extensively discussed in the literature since 1980s. Asymptotic normality of estimating Lorenz curve vector ordinates and testing LD by significant differences among pairwise comparisons of Lorenz ordinates were essentially based on a grid of finite points on the unit interval. Extending the inference of LD from a finite grid to the entire unit interval, a consistent test was developed for the LD hypothesis based on functions of the empirical Lorenz processes (Barrett et al, 2014). The theoretical test *p*-values were empirically assessed using the bootstrap method because the asymptotic distributions of the test statistics depend on the unknown distributions.

The goal of this study is to investigate a distribution-free test scheme for the LD hypothesis. This begins with an elementary fact that the difference curve of a pair of Lorenz curves is a combination (or continuation) of concave and convex curves depending on the condition of crossing Lorenz curves, denoted by CLC, or dominant curves, the LD case. Distinct aspects of the difference curve expressed by inequality patterns of the quantile functions naturally characterize the CLC and LD conditions. Thus, the quantile inequality patterns can be used to construct likelihood ratio (LR) tests under the CLC and the LD hypotheses, respectively. The LR test for the LD hypothesis is by design a reduced form of that for the CLC hypothesis, it is convenient to first test the latter when it is applicable, otherwise, test only the former; and the separate test results can be summarized to support a decision. The proposed LR tests are consistent with respect to the critical regions and test levels based on approximate chi-square distributions. The proposed LR tests are examined using a simulation study and a real GDP per capita data analysis.

The proposed test design is laid out in two sections. In Section 2, it is illustrated that distinct inequality patterns between paired quantiles of the two distribution under comparison are exhibited under the CLC and LD hyptheses. From these patterns, LR tests for the CLC and LD hypotheses can be separately constructed by comparing the sample quantiles against the expected quantiles under the hypotheses. In Section 3, a simulation study of pairs of Lorenz curves from a few common distribution families is conducted to evaluate the effectiveness of the proposed LR tests under the CLC and LD hypotheses. The proposed LR tests are also used to investigate potential CLC and LD conditions among a few real GDP per capita data of 133 countries recorded across a few years in the Penn World Table (Summers and Heston, 1991). A pair of yearly data was tested to exhibit the CLC hypotheses and another pair exhibited the LD property, the proposed LR tests were effectively used and conclusive. The bootstrap method was also applied to testing the same two paired data, and found not quite satisfactory with the CLC case, and

not satisfatory with the LD case when two curves were quite close to each other. It is worth noting that the proposed LR tests derived from count statistics of bivariate data are applicable to random samples of any two distributions whether they are independent or not. We conclude this study with a brief explanation that the LR tests are developed along with the rationale of a nonparametric sign test for the difference curve between two Lorenz curves.

2. Methodology

The Lorenz curve:

Assume that a non-negative random variable X has an absolutely continuous distribution with continuous probability density f, finite mean $\mu_F = \mathrm{E}X$ and variance $\sigma_F^2 = \mathrm{Var}X$. Let $\{X_i,\ i = 1, \cdots, n\}$ and $\{Y_j,\ j = 1, \cdots, m\}$ denote random samples of variables X and Y with cumulative distribution functions (cdf) F and G, and probability density functions (pdf) f and g, respectively. The empirical cdf of the $\{X_i\}$ is defined by $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$, $0 \le x < \infty$; the quantile function and the empirical quantile function are defined by $F^{-1}(t) = \inf\{x: F(x) \ge t\}$ and $\hat{F}^{-1}(t) = \inf\{x: \hat{F}(x) \ge t\}$, $0 \le t \le 1$, respectively. Similar notations for the variable Y are defined by analogy.

Definition 1. The Lorenz curve of the random variable X is, for $0 \le\ t\ \le 1$,

$$L_F(t) = \frac{1}{\mu_F}\int_0^t F^{-1}(u)du,$$

and the empirical Lorenz curve is

$$\hat{L}_F(t) = \frac{1}{\hat{\mu}_F}\int_0^t \hat{F}^{-1}(u)du, \qquad (2.1)$$

where $\hat{\mu}_F = \frac{1}{n}\sum_{i=1}^{n} X_i\ (= \bar{X})$ is the sample mean.

Likelihood ratio (LR) test statistics:

Consider the null hypothesis H_0 of CLC between distribution functions F and G, and the alternative hypothesis H_1 of LD. Specifically, let $H_0 = H_{01} \cup H_{02}$ present two distinct cases of CLC: H_{01} denotes the case of a single intersection, say, L_F crosses L_G once from above; and H_{02} denotes the situation of having two or more intersection points, which rarely occurs in practice.

Case 1. Under H_{01}, there exists a unique $t^* \in (0, 1)$ such that $D(t) \equiv L_F(t) - L_G(t) > 0$, $t \in (0,\ t^*)$; $D(t^*) = 0$; and $D(t) < 0$, $t \in (t^*,\ 1)$. By convexity of the Lorenz curve, $D(t)$ is concave on $[0,\ t^*)$, and convex on $(t^*, 1]$. That is,

$$D''(t) = \frac{1}{\mu_F f(F^{-1}(t))} - \frac{1}{\mu_G g(G^{-1}(t))} < 0, \qquad (2.2)$$

for $t \in (0,\ t^*)$, $D''(t^*) = 0$ and $D''(t) > 0$ for $t \in (t^*,\ 1]$. By mean-value theroem, there are $\{t_1, t_2\}$ such that

$$D'(t_1) = 0 = D'(t_2),\ 0 < t_1 < t^* < t_2 < 1,$$

where $D'(t) = F^{-1}(t)/\mu_F$. The maximum and minimum of $D(t)$ occur at t_1 and t_2, respectively, and

$$\frac{F^{-1}(t)}{\mu_F} > \frac{G^{-1}(t)}{\mu_G},\ 0 < t < t_1\ ;\ t_2 < t < 1$$

and

$$\frac{F^{-1}(t)}{\mu_F} < \frac{G^{-1}(t)}{\mu_G},\ t_1 \le t \le t_2. \tag{2.3}$$

The sample analog of $\{t_1, t_2\}$ are solutions to the equations

$$\frac{\hat{F}^{-1}(\hat{t}_i)}{\bar{X}} - \frac{\hat{G}^{-1}(\hat{t}_i)}{\bar{Y}} = 0, \qquad i = 1, 2;$$

and the next two inequalities hold

$$\frac{\hat{F}^{-1}(t)}{\bar{X}} > \frac{\hat{G}^{-1}(t)}{\bar{Y}},\ 0 < t < \hat{t}_1\ ;\ \hat{t}_2 < t < 1$$

and

(2.4)

$$\frac{\hat{F}^{-1}(t)}{\bar{X}} < \frac{\hat{G}^{-1}(t)}{\bar{Y}},\ \hat{t}_1 \le t \le \hat{t}_2.$$

Case 2. The hypothesis H_{02} was rarely discussed in the literature, because it rarely occurs in practice. For ease of exposition, this Case 2 will not be discussed because it would not make any significant effect of inference when it would occur with negligibly small probability in application.
Under H_1, it follows from the discussions above that

$$\frac{\hat{F}^{-1}(t)}{\bar{X}} > \frac{\hat{G}^{-1}(t)}{\bar{Y}},\ 0 < t < \hat{t}_1$$

and (2.5)

$$\frac{\hat{F}^{-1}(t)}{\bar{X}} < \frac{\hat{G}^{-1}(t)}{\bar{Y}},\ \hat{t}_1 < t < 1.$$

To summarize the analysis, it is seen that under H_{01}, the inequality set (2.4) is satisfied by locating $\hat{t}_1$ and $\hat{t}_2$, and the observed counts of ordered sample quantile pairs are recorded in Table 1.1 below.

	Counts of unequal sample t^{th} quantile pairs in (2.4)		
Sample quantiles	$t < \hat{t}_1$	$\hat{t}_1 \le t < \hat{t}_2$	$\hat{t}_2 \le t < 1$
$\hat{F}^{-1}(t)/\bar{X} > \hat{G}^{-1}(t)/\bar{Y}$	n_{11}	$n_{12}+2$	n_{13}
$\hat{F}^{-1}(t)/\bar{X} < \hat{G}^{-1}(t)/\bar{Y}$	$n_{21}+1$	n_{22}	$n_{23}+1$

Table 1.1 Counts of unequal sample quantile pairs under H_{01}

The theoretical analog of Table 1.1 can be derived from the inequality set (2.3) as follows.

	Expected counts of t^{th} quantile pairs in (2.3)		
Quantiles	$t < t_1$	$t_1 \le t < t_2$	$t_2 \le t < 1$
$F^{-1}(u)/\mu_F > G^{-1}(u)/\mu_G$	$\min(m_1, n_1)$	2	$\min(m', n')$
$F^{-1}(u)/\mu_F < G^{-1}(u)/\mu_G$	1 (or 2)	$\min(m_2, n_2)$	1 (or 2)

Table 1.2. Expected counts of unequal quantile pairs under H_{01}

Here, $m_1 (= mt_1)$ is the expected counts of $F^{-1}(t)/\mu_F$ before the t_1^{th} quantile, $m_2 (= m(t_2 - t_1))$ is the expected counts between t_1^{th} and t_2^{th} quantiles, and $m' = m - (m_1 + m_2)$, the remaining counts among those larger than the t_2^{th} quantile, in the smoothed $F^{-1}(t)/\mu_F$. By analogy, n_1 is the expected counts of $G^{-1}(t)/\mu_G$ sample less than its t_1^{th} quantile, n_2 is the expected counts between the t_1^{th} and t_2^{th} quantiles, and $n' = n - (n_1 + n_2)$, the remaining quantiles. The set of integers {1, 2, 1} in Table 1.2 is accounted for the numbers of boundary ties at $t_i,\ i = 1, 2$, which makes little effect on the asymptotic distribution. It follows that a LR test statistic between Tables 1.1 and 1.2 can be defined as

$$T_0 = 2\begin{bmatrix} n_{11} log\left(\frac{n_{11}}{\min(m_1,n_1)}\right) + (n_{12}+2) log\left(\frac{n_{12}+2}{2}\right) + n_{13} log\left(\frac{n_{13}}{\min(m',n')}\right) \\ +(n_{21}+1) log\left(\frac{n_{21}+1}{1}\right) + n_{22} log\left(\frac{n_{22}}{\min(m_2,n_2)}\right) + (n_{23}+1) log\left(\frac{n_{23}+1}{1}\right) \end{bmatrix}, \quad (2.6)$$

which approximates the chi-square distribution with 2 *df*, or χ_2^2 under H_{01}.

Comparable to Tables 1.1 and 1.2, the following Tables 2.1 and 2.2 are obtained under H_1:

	Counts of sample quantile pairs in (2.5)	
Sample quantiles	$t < t_1$	$t_1 \le t < 1$
$\hat{F}^{-1}(t)/\bar{X} > \hat{G}^{-1}(t)/\bar{Y}$	n_{11}	$n_{12}+1$
$\hat{F}^{-1}(t)/\bar{X} < \hat{G}^{-1}(t)/\bar{Y}$	$n_{21}+1$	n_{22}

Table 2.1. Counts of unequal sample quantile pairs under H_1

	Expected counts of quantile pairs in (2.5)	
Quantiles	$t \le t_1$	$t_1 \le t < 1$
$F^{-1}(u)/\mu_F > G^{-1}(u)/\mu_G$	$\min(m_1, n_1)$	1 (or 2)
$F^{-1}(u)/\mu_F < G^{-1}(u)/\mu_G$	1 (or 2)	$\min(m_2, n_2)$

Table 2.2. Expected counts of unequal quantile pairs under H_1

Comparable to formula (2.6), the LR test Under H_1 is defined as

$$T_1 = 2\left[\begin{array}{l} n_{11}log\left(\frac{n_{11}}{\min(m_1,n_1)}\right) + (n_{12}+1)log\left(\frac{n_{12}+1}{1}\right) \\ +(n_{21}+1)log\left(\frac{n_{21}+1}{1}\right) + n_{22}log\left(\frac{n_{22}}{\min(m_2,n_2)}\right) \end{array}\right]. \quad (2.7)$$

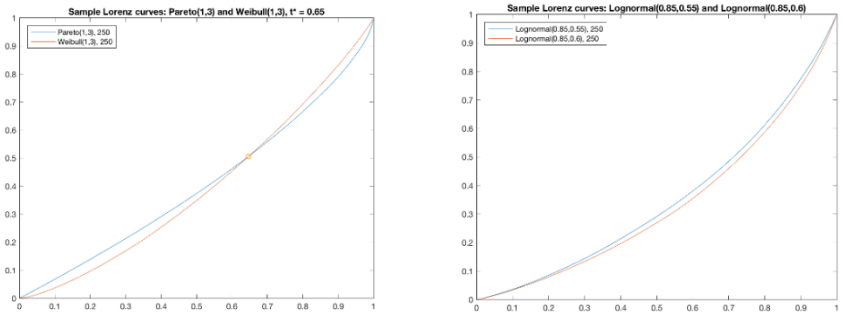
3. Results in Practice

In the simulation study, both test statistics T_0 and T_1 are examined under the CLC condition (H_{01}) and the LD condition (H_1), respectively, exemplified in Plot 1. The average rejection rates of the test statistics are recorded from 1000 replicates of the paired distributions between log-normal, Pareto and Weibull distributions, as listed in Tables 1 and 2 below.

1st Distribution	**Sample size**	**2nd Distribution**	Sample size	Test T_0	Test T_1
Pareto (1.0, 3.0)	250	Lognormal (0, 0.09)	250	0.005	0.999
	250		1000	0.002	1.000
	1000		250	0.0006	1.000
	1000		1000	0.000	1.000
Pareto (1.0, 3.0)	250	Weibull (1.0, 3.0)	250	0.050	1.000
	250		1000	0.014	1.000
	1000		250	0.047	1.000
	1000		1000	0.000	1.000

Weibull (1.0, 1.2)	250	Lognormal (0, 1.0)	250	0.265	0.992
	250		1000	0.167	0.994
	1000		250	0.127	1.000
	1000		1000	0.005	1.000

Table 1. Rejection rates in 1,000 replicates of the LR test (2.6) and (2.7) under H_{01}



Plot 1. CLC case: Pareto (1.0, 3.0) and Weibull(1.0, 3.0), with equal sample size 250 (Left). LD case: Log-normal pair (0.85, 0.3025) and (0.85, 0.36) with equal sample size 250 (Right).

1st Distribution	**Sample size**	**2nd Distribution**	Sample size	Test T_0	Test T_1
Pareto (1.0, 6.0)	250	Pareto (1.0, 3.0)	250	0.986	0.022
	250		1000	0.991	0.011
	1000		250	0.990	0.006
	1000		1000	0.999	0.000
Lognormal (0.85, 0.09)	250	Lognormal (0.85, 0.36)	250	0.960	0.059
	250		1000	0.982	0.024
	1000		250	0.965	0.039
	1000		1000	1.000	0.000
Lognormal (0.85, .3025)	250	Lognormal (0.85, 0.36)	250	0.851	0.807
	250		1000	0.898	0.683
	1000		250	0.736	0.789
	1000		1000	0.831	0.646
	10,000		10.000	1.0	0.186
	30.000		30.000	1.0	0.109
	50.000		50.000	1.0	0.080

Table 2. Rejection rates in 1,000 replicates of the LR tests (2.6) and (2.7) under H_1

The consumption data of 133 countries from the Penn World Tables is employed (Summers and Heston, 1995). The data were measures of real GDP per capita, in constant dollars adjusted for changes in terms of trade (1985 international prices for domestic absorption and current prices for exports and imports) for the years 1970, 1975, 1980 and 1985. The yearly paired data from (1970, 1980) were tested for CLC and those of (1975, 1985) were tested for LD. In (1970, 1980).

Hypothesis	LR test statistic	p - value	Bootsrap rejection rate
H_{01}	$T_0 = .001$	1.00	0.527
H_1	$T_1 = 36.843$	< 0.001	0.999

Table 3. Data pair (1970, 1980) of 133 countries; $t^* = 0.74, t_1 = 0.45$ *and* $t_2 = 0.90$

Hypothesis	LR test statistic	p - value	Bootsrap rejection rate
H_{01}	$T_0 = 58.224$	< 0.001	0.940
H_1	$T_1 = 1.673$	0.196	0.893

Table 4. Data pair (1975, 1985) of 133 countries; $t_1 = 0.77$

4. Discussion and Conclusion

In this study, the idea of nonparametric sign test is employed to test the LD hypothesis between two Lorenz curves, bypassing the standard analysis based on the empirical Lorenz process. It leads to testing the crossing Lorenz curves (CLC) hypothesis H_0 against the alternative LD H_1. The decomposition of H_0 into $H_{01} \cup H_{02}$ suggests that it is primary to test H_{01}, the case of "crossing exactly once". Whereas, H_{02}, the rare case of two or more crossing points, would be tested for its validity only when both H_{01} and H_1 are evidently significant, without a clue for conclusion.

In the simulation study, there is a difficult condition, where two sample Lorenz curves can be very close to (or entangled with) each other over the unit interval such that both LR tests for H_{01} and H_1 are significant. In this rare situation, it usually occurs that the rejection rates under H_1 are in general much less than that under H_{01}, indicating a fair support for the hypothesis H_1. Similar conditions in a simulation study would be less difficult to treat using increased sample sizes.

In summary, this study proposes useful LR tests by analogy with a sign test. The LR tests are however developed for practical testing effect, not designed for theoretical analysis of power and specificity in the classical testing hypotheses.

References

1. Atkinson, A. B. (1970), On the measurement of inequality, *Journal of Economic Theory*, 2, 244-263.
2. Barrett, G. F., Donald, S. G. & Bhattacharya, D. (2014). Consistent nonparametric tests for Lorenz dominance, *J. Business & Economic Statistics*, 32, 1-13.
3. Lorenz, M. O. (1905). Methods of measuring the concentration of wealth, *J. American Statistical Association,* 9, 209-219.
4. Summers, R. and Heston, A. (1995). The Penn World Tables, Version 5.6, NBER, Cambridge, MA. http://www.nber.org/pub/pwt56/.

Flipping the online classroom in a multivariate data analysis course

Christina Andersson[1], Gerald Kroisandt[2]
[1]Dept. of Computer Science and Engineering, Frankfurt University of Applied Sciences, Germany
[2]School of Engineering, htw saar, Germany

Abstract
Teaching statistics to non-statistics students at university level, we often face a lot of problems, e.g. lack of motivation to study statistics, anxiety of the subject and insufficient preknowledge in mathematics. To reduce the impact of such obstacles, one approach is to incorporate active learning components in the course, e.g. the flipped classroom strategy. In this paper, we present how the classroom was flipped in an advanced multivariate data analysis course, which is taught completely as an online course, i.e. without any on-site sessions at the university. We describe the teaching framework of the course and discuss the lessons learned from the first teaching experience.

Keywords
Flipped classroom; active learning; multivariate data analysis; e-learning

1. Introduction

A statistics course for non-statistics students can be a real challenge for both lecturers and students: The students often show a low motivation for learning the subject, not seldom combined with anxiety and lack of prerequisite knowledge, e.g. poor mathematical skills (Gal and Ginsburg, 1994; Dillon, 1982; Forte, 1995; Schutz et al., 1998, Townsend et al., 1998; Yilmaz, 1996; Väisanen et al., 2004; Onwuegbuzie, 2003). One way to overcome such obstacles and to improve the existing statistics courses can be to focus on student-centered learning (Roseth et al, 2008; Prins, 2009; Sciutto, 1995). To encourage the students to act as active learners and to engage them in the learning process can be an important component for enhancement of the courses and the students' learning environment (Prince, 2004; Freeman et al., 2014; Bonwell and Eison, 1991; Chickering and Gamson, 1987). This seems also to be the case for the application of active learning in statistics courses (Carlson and Winquist, 2011; Dolinksy, 2001; Gnanadesikan et al., 1997; Knypstra, 2009; Kvam, 2000; Dierker et al., 2018). One approach to use active learning is to flip the classroom. A short definition of the flipped classroom would be to say that those activities that in traditional teaching took place within the classroom now take place outside the classroom and vice versa (Lage et al., 2000). However, the flipped classroom strategy is more than an

interchange of the activities taking place in different locations. We can express the flipped classroom as an approach, where initial out-of-class activities include individual, often computer-based preparatory activities, followed by in-class group-centered interactive actions (Bishop and Verleger, 2013). This means that a kind of more passive knowledge acquisition takes place outside the classroom and hands-on related activities are dominating inside the classroom (Lockwood and Esselstein, 2013). The flipped classroom strategy has been described and evaluated, mostly with a positive result, in many reports (Strayer, 2012; Bergman and Sams, 2012; Roehl, 2013), including the application to undergraduate statistics courses (Wilson, 2013; Touchton, 2015; Cilli-Turner, 2015; Winquist and Carlson, 2014; Chen et al., 2015, Phillips and Phillips, 2016; Vidic and Clark, 2016).

Usually, the flipped statistics classroom has been applied in such a context that the out-of-class activities take place without a physical classroom, whereas the in-class activities take place in a real, physical classroom, where the students and lecturers meet face-to-face (Wilson, 2013; Touchton , 2015; Cilli-Turner, 2015; Winquist and Carlson, 2014; Chen et al., 2015, Phillips and Phillips, 2016; Vidic and Clark, 2016). In our case, we show how the flipped classroom can be applied to a statistics course, which is taught completely online. This means that we still flip the classroom concerning the nature of the learning process, i.e. the activities traditionally taught during the class take place out of class and vice versa. But, the whole teaching framework of the course is completely web-based, without face-to-face phases, in order to ensure the students better possibilities for time and location independent studies. In this paper, we present how the flipped classroom approach was applied to an advanced course in multivariate data analysis, in order to reduce the problems occurring in statistics courses by introducing active learning components in the course. We describe the settings of the flipped online course as well as lessons learned in the first in-class usage of the approach.

2. The Online Course Multivariate Data Analysis

In the era of data science, the statistics courses play an essential role in the computer science curricula at both B.Sc. and M.Sc. level at Frankfurt University of Applied Sciences (FRA-UAS). This is e.g. the case in the curriculum of the program High Integrity Systems (HIS), which is a two-year international M.Sc. program in computer science with an all-English curriculum (High Integrity Systems, 2018). This degree gives the students many challenging professional opportunities in different areas of software application. The concept *high integrity systems* refers to complex systems, controlled by software, having a large impact on humans and society in case of a failure, such as safety critical systems, which are directly influencing the life of humans, as e.g. aerospace, automotive, railway and marine systems or medical technology. We also talk

about high integrity systems in the context of mission critical systems, which, in case of a failure, can cause immediate, essential problems for the operation of an organization, such as ERP, CRM, etc. This means that the computer scientists with a Master's degree from HIS often hold a position, which involves both responsibility and a lot of data analysis.

Within the HIS program, two courses concerning statistics and data analysis are mandatory for all students: The students have to take the basic course Introductory Data Analysis and the semi-advanced course Data Mining Methods. In the first course, Introductory Data Analysis, it is assumed that the students already have passed a statistics course at B.Sc. level, i.e. the course starts with a review of basic concepts and then moves on to topics within inferential statistics and linear and logistic regression. The course Data Mining Methods introduces further methods for analyzing data, such as decision trees, neural networks and support vector machines, stressing the importance of data preparatory issues.

After these two courses, the students can choose to study the elective course Multivariate Data Analysis, which is a high-level course, directly based on the knowledge the students should have gained in the first two courses. The Multivariate Data Analysis course prepares for application of advanced multivariate statistical methods to practical problems and serves as a platform for those students who consider their master's thesis to be within data analysis. Among the contents included in the course, we find topics as principal component analysis (PCA), partial least squares regression (PLS), discriminant analysis, canonical correlation and cluster analysis. The objective of this course is to provide the students with both a sufficient theoretical foundation and in addition to this to give them a thorough knowledge of how to apply theory to some real-world situations by using statistical software, such as R, for analyzing large and complex data sets. The software R has been chosen, since it is used in many different industrial applications in companies in Germany. Furthermore, the software is easy to use, but still accommodates advanced statistical methods. Finally, another reason to use R is that it is free software, i.e. we do not have to deal with license costs and conditions. The examination of the course is a computer-based, written exam, i.e. the exam takes place in a computer room. This enables us to test the students' ability to apply the multivariate methods to real-world problems.

The program HIS is taught as a traditional M.Sc. program and in general, the courses take place face-to-face at the campus. However, in order to facilitate for the students to organize their busy schedule, the university seeks to offer online courses as an option for time and location independent studies. As a part of these efforts, the course Multivariate Data Analysis has been virtualized and is now offered as an online course.

3. Application of the Flipped Classroom Approach

For many years, the course Multivariate Data Analysis was taught as a traditional classroom course, consisting of a mixture of theory lectures and computer-based exercise sessions. In order to facilitate for the students to combine studies with work and family life, i.e. a usually busy schedule, the course has in later years been taught as an online course. The first time the course was taught online, this was done as a blended learning approach, i.e. the course contained mostly online components, but also on-site meetings at the campus. However, the students didn't show much interest in attending the voluntary on-site meetings and therefore the course now takes place completely online.

The first attempt to strengthen the student-centered activities was done in this online-taught course by the use of the cooperative learning technique jigsaw (Andersson and Logofatu, 2016). The application of the jigsaw technique to the multivariate data analysis course was successful in the sense that the student activity increased in the course and the resulting feedback from the students was mostly positive to the approach. A drawback with the jigsaw technique, applied to this course, was that the students suffered severely when group members dropped out and decided not to complete the course.

Therefore, we now consider the inverted classroom as an alternative method to activate the students. The online implementation of the course is based on the learning management system Moodle (Moodle, 2015). There are mainly two reasons for this choice of e-learning platform: Firstly, this platform is the most frequently used e-learning platform at FRA-UAS. Secondly, mathematical formulas can easily be constructed in Latex in Moodle.

In accordance with the traditional flipped classroom approach (Wilson, 2013; Touchton , 2015; Cilli-Turner, 2015; Winquist and Carlson, 2014; Chen et al., 2015, Phillips and Phillips, 2016; Vidic and Clark, 2016), we divided the learning activities in the course Multivariate Data Analysis into out-of-class and in-class activities. In our case, both kind of activities take place completely online in order to facilitate time and location independent studies for the students. The tuition components of the course can be described as follows:

Description of the out-of-class activities

- The students were partitioned into groups with three or four students in each group.
- Each group received a specific multivariate method (PCA, PLS, cluster analysis, discriminant analysis etc.) to learn about, i.e. the members of each group should become experts concerning their topic during the semester. This learning process should take place as a more or less completely student-centered activity, where the lecturers only act as guides if explicitly needed.

- Additionally, at the beginning of the course, the students received a time table, containing rough guidelines about what they, in the groups, were expected to process during the week. The guidelines were deliberately kept in an imprecise way, without details, in order to allow own initiatives of the students and to let self-centered learning take place. For example, the first week the guidelines encouraged the students to gather basic knowledge about their topic, the second week to search for existing algorithms for the method etc. The students were told to use the guidelines to structure their information retrieval, but that the main research about the multivariate method itself should be a student-centered inquiry-based learning process.

The guidelines suggested following steps to be processed during the semester:

1. Theoretical background of the method
2. A summary of required data preparation
3. Implementation of the method in R
4. Application to a real-world data set
5. Interpretation of results

- The students were offered to use the web conference tool Adobe Connect to perform their out-of-class group meetings online.
- As another out-of-class activity, the students were asked to construct a wiki-based website about their multivariate method.

Description of the in-class activities:

- Weekly web conferences, where both the students and the lecturer participated, were offered.
- In these web conferences, mainly the students were the actors, presenting the weekly progress of their research about their group's multivariate method. These presentations contained both the theoretical explanation of the studied method as well as practical implementation of the method in R and the application to a real-world problem, including a thorough interpretation of the R output and the numerical results.
- Weekly, a Questions-&-Answers session was offered in either the chat or the discussion room of Adobe Connect. In this session, the students had the opportunity to ask the lecturer about non-understandable issues, concerning their multivariate method.
- During one extra web conference, the students presented the wiki-based websites they had prepared.

As a prerequisite to take the final exam, the students had to attend at least 80% of the web conferences and, additionally, regularly report about their topic.

4. Discussion and Conclusions

Since this is an elective course, every year approximately twelve students participate in the course. This means that the sample is much too small to evaluate the results with standard statistical methods. In the written evaluation of the course, some of the students mentioned that it is more difficult to learn independently about their topic during the out-of-class activities, than it would have been to listen to a conventional lecture about the same topic. However, the students also mentioned that even if they experienced their working load to be higher than in a traditionally face-to-face taught course, they in the end felt more comfortable and confident in the use of the statistical methods than after completing similar traditionally taught courses. A few students commented in the evaluation that, according to them, it was appropriate to flip the classroom in an advanced course, since they already had basic knowledge about statistics. The students meant that it would have been more difficult in a beginner's course due to lack of basic knowledge concerning the subject.

The course lecturers, who previous years also have taught the same course in a traditional way, noticed that the students, in general, appeared well prepared to the in-class activities. During the web conferences, the students discussed with each other and posed several questions to the presentations of the different multivariate methods and their applications, i.e. an increased level of student interaction was observed, compared to the traditionally taught courses previous years.

We conclude that the sample of students is too small to perform a strict quantitative evaluation of this teaching approach, but that the qualitative results from the course evaluation and the positive observations of student interactions in the web conferences, encourage us to continue the implementation of the flipped classroom in advanced statistics courses, taught in an online setting.

References

1. Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. Journal of Statistics Education, 2(2).
2. Dillon, K. M. (1982). Statisticophobia. Teaching of Psychology, 9(2), 117.
3. Forte, J. A. (1995). Teaching statistics without sadistics. Journal of Social Work Education, 31(2), 204-218.
4. Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1998). Prior knowledge, attitude, and strategy use in an introduction to statistics course. Learning and Individual Differences, 10(4), 291-308.
5. Townsend, M. A., Moore, D. W., Tuck, B. F., & Wilton, K. M. (1998). Self-concept and anxiety in university students studying social science statistics within a co-operative learning structure. Educational Psychology, 18(1), 41-54.
6. Yilmaz, M. R. (1996). The challenge of teaching statistics to non-specialists. Journal of Statistics Education, 4(1), 1-9.
7. Väisanen, P., Ylonen, S., & Rautopuro, J. (2004). Modelling the impact of teacher education students' real and perceived maths ability and motivational-affective factors on their success in elementary statistics.
8. Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. Teaching in Higher Education, 8(2), 195-209.
9. Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in learning and teaching statistics. Journal of Statistics Education, 16(1).
10. Prins, S. C. B. (2009). Student-centered instruction in a theoretical statistics course. Journal of Statistics Education, 17(3).
11. Sciutto, M. J. (1995). Student-centered methods for decreasing anxiety and increasing interest level in undergraduate statistics courses. Journal of Instructional Psychology.
12. Prince, M. (2004). Does active learning work? A review of the research. Journal of Engineering Education, 93(3), 223-231.
13. Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410-8415.
14. Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183.
15. Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. AAHE bulletin, 3, 7.

16. Carlson, K. A., & Winquist, J. R. (2011). Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach. Journal of Statistics Education, 19(1).
17. Dolinsky, B. (2001). An active learning approach to teaching statistics. Teaching of Psychology.
18. Gnanadesikan, M., Scheaffer, R. L., Watkins, A. E., & Witmer, J. A. (1997). An activity-based statistics course. Journal of Statistics Education, 5(2).
19. Knypstra, S. (2009). Teaching statistics in an activity encouraging format. Journal of Statistics Education, 17(2).
20. Kvam, P. H. (2000). The effect of active learning methods on student retention in engineering statistics. The American Statistician, 54(2), 136-140.
21. Dierker, L., Flaming, K., Cooper, J. L., Singer-Freeman, K., Germano, K., & Rose, J. (2018). Evaluating Impact: A Comparison of Learning Experiences and Outcomes of Students Completing A Traditional Versus Multidisciplinary, Project-Based Introductory Statistics Course. International Journal of Education, Training and Learning, 2(1), 16-28.
22. Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. The Journal of Economic Education, 31(1), 30-43.
23. Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. In ASEE national conference proceedings, Atlanta, GA (Vol. 30, No. 9, pp. 1-18).
24. Lockwood, K., & Esselstein, R. (2013). The inverted classroom and the CS curriculum. In Proceeding of the 44th ACM technical symposium on Computer science education (pp. 113-118). ACM.
25. Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. Learning Environments Research, 15(2), 171-193.
26. Bergmann, J., & Sams, A. (2012). Flip your classroom: Reach every student in every class every day. International Society for Technology in Education.
27. Roehl, A., Reddy, S. L., & Shannon, G. J. (2013). The flipped classroom: An opportunity to engage millennial students through active learning strategies. Journal of Family & Consumer Sciences, 105(2), 44-49.
28. Wilson, S. G. (2013). The flipped class: A method to address the challenges of an undergraduate statistics course. Teaching of Psychology, 40(3), 193-199.
29. Touchton, M. (2015). Flipping the classroom and student performance in advanced statistics: Evidence from a quasi-experiment. Journal of Political Science Education, 11(1), 28-44.

30. Cilli-Turner, E. (2015). Measuring learning outcomes and attitudes in a flipped introductory statistics course. Primus, 25(9-10), 833-846.
31. Winquist, J. R., & Carlson, K. A. (2014). Flipped statistics class results: Better performance than lecture over one year later. Journal of Statistics Education, 22(3).
32. Chen, L., Chen, T. L., & Chen, N. S. (2015). Students' perspectives of using cooperative learning in a flipped statistics classroom. Australasian Journal of Educational Technology, 31(6).
33. Phillips, L., & Phillips, M. (2016). Improved student outcomes in a flipped statistics course. Administrative Issues Journal, 6(1), 10.
34. Vidic, N. S., & Clark, R. M. (2016). Comparison of a Partially Flipped vs. Fully Flipped Introductory Probability and Statistics Course for Engineers: Lessons Learned. In American Society for Engineering Education. 2016 ASEE Annual Conference Proceedings, New Orleans, LA.
35. High Integrity Systems, http://frankfurtuniversity.de/fileadmin/de/Fachbereiche/FB2/RZ_FUAS_HIS_2016.pdf, accessed 10 December 2018.
36. Moodle. Moodle – Learning Management System. http://moodle.de, 2015. Accessed: 2019-01-21.
37. Andersson, C. & Logofatu, D. (2016), Application of the Jigsaw Technique in a Blended Learning Course in Multivariate Statistics, ISI World Statistics Congress Proceedings (to be published), Marrakech.

The survey and research of "non-observed finance" -- Take Shenzhen as an example

Yang Xinhong

Statistics Bureau of Guangdong Province, Guangzhou, The People's Republic of China

Abstract

"Non-observed Finance" includes folk informal finance, underground finance and illegal finance. Taking self-employed households in Shenzhen as samples, this study is the first time in China to explore the use of direct-survey method to measure the "Non-observed Finance" scale and its added value. The result shows that the total amount of "Non-observed Finance" of self-employed households in Shenzhen municipality is tens of billions of yuan, which is basically consistent with the data collated by industry associations and the data of formal loan provided by banking regulatory department. FISIM method is used to calculate the added value of self-employed households in 2015.

Keywords

informal finance; underground finance; illegal finance; direct-survey-method; FISIM method

1. Introduction

In recent years, "Non-observed Finance" ("NOF") , a terminology referring to the "Non-observed Economy" ("NOE"), has been used by domestic theoretical circles (Tian Guangning, 2008; Li Jianjun, 2008, 2010; Ren Biyun, 2013). It refers to the financial organizations which exist outside the System of National Accounts, financial statistics monitoring system and economical and financial supervision system, their activities, as well as the value forms of monetary and financial assets resulting from their activities. In general, "NOF" is referred to as folk informal finance, underground finance and illegal finance. The existence of "NOF" affects the government's judgment about the financial and economic situation, and impacts the effect of macro monetary policy. To make a judgment about the impact and effect mentioned above, we must master the scale of "NOF". That how to monitor "NOF" has also became the focus of theoretical research and practical work.

At present, only a limited number of domestic scholars, such as Tian Guangning (2008) Jianjun (2006, 2008, 2010) [4-5, 7] ;Ren Biyun (2013) have studied "NOF", focusing on the definition of "NOF", the estimation of its scale and its impact on the money supply, etc. However, the scholars apply the

"macro-index-estimation-method" or "indirect-estimation-method" to estimate the scale of "NOF", which is apt to subject to the theoretical basis of the relationship between "NOF" and other macroeconomic indicators. Without being able to clarify the theoretical relationship between "NOF" and the official economy, the estimation outcome will inevitably be biased. Therefore, this paper intends to explore the application of micro-method, that is, "direct-estimation-method", with self-employed households in Shenzhen as samples. This method calculates the scale of "NOF" and the added value of "NOF" activities, basing on the data obtained through the statistical survey of the objects involved in the "NOF". The theoretical value of this research lies in the first attempt in China to determine the scale of "NOF" activities by direct-survey-method, while its practical significance lies in providing reference for further improving the system of statistical methods, improving statistical work, understanding the situation of off-balance-sheet finance and cash flow from abnormal channel, and facilitating financial activities better serve the real economy.

2. The Design of "NOF" Survey Scheme

As both a leading frontier of reform and opening up and a model city of economic transformation, Shenzhen's pillar industry is the financial industry. With both the enormous scale of traditional financial activities over the years and the vigorous development of various emerging financial industries in recent years, the financial indicators have maintained sustainably stable growth. By researching the scale of "NOF" in the real economy of Shenzhen and further studying its disturbance to the economic operation, the government is able to correctly judge the financial and economic situation, thus to facilitate the decision-making to accurately implement various financial policies.

2.1 The Scope, Object and Content of the Survey

With various economic entities in the System of National Accounts, capital flow accounting is classified into the sectors of non-financial enterprises, financial institutions, government and household. According to the characteristics of China, economic units are divided into enterprises, self-employed households, households and administrative institutions. Considering the principle of feasibility, the scope and object of this survey are "NOF" loan activities of self-employed households within Shenzhen. "NOF" loan activities refer to the activities to finance the operating capital by economic units through commercial credit (inter-enterprise loan, credit), folk lending usury, social fund-raising and interpersonal loan. The specific classification of economic units and the subjects of this survey are shown in Table 1.

Table 1 Classification of economic units and "NOF" survey subject

Classification of economic units		Judgment about "NOF" amount	Is it the subject of this survey?
Non-financial enterprises	Large and medium-sized	A large amount	No
	Small and micro-sized	A large amount	No
Self-employed households	Big self-employed household	A reasonable amount	Yes
	General self-employed households	A small amount	Yes
Households		Very few	No
Financial enterprises		Very few	No
Administrative institutions		Non	No

Note: "NOF" amount refers to the economic amount of operating capital that economic units raise through commercial credit (that are credit and lending between enterprises), folk usury, social fund-raising and private borrowing Judgment about "NOF" amount is a preliminary judgment.

The main contents of the survey are the operation status, financing channels and financing scale of self-employed households, mainly inquiring into the turn of capital and the proportion of "NOF" in financing operating capital. In order to enhance the reliability of self-employed households' answers, Shenzhen Gold and Jewelry Industry Association, Shenzhen Service and Trade Association and Shenzhen Mobile Phone Association were selected to survey their membership scale, overall operation status and relevant empirical data. Self-employed households were surveyed using the "*Self-employed Household Operating and Financing Questionnaire*" and industry associations were surveyed using the "*Shenzhen Industry Development and Financial Activities Questionnaire*". The survey-period index was 2015, and the time-point index was the end of 2015.

2.2 Survey Methods and Parameter Estimation

2.2.1 Survey Methods

The three-stage stratified systematic sampling is adopted in the survey, which requires the samples to be representative of strata. Large sample sampling (n 2 30) is applied within strata, and sample statistics can be used to calculate parameters within strata.

The "big self-employed households" and "the third national economic census" self-employed household data base, which was held by Shenzhen Statistical Bureau on September 30, 2016, was taken as sampling box. All self-employed households within Shenzhen are divided into two strata, namely "big self-employed households" and "general self-employed households".

For each stratum of "big self-employed households", including the "gold and jewelry wholesale", "flower retail", "mobile phone wholesale" and "other big self-employed households ", all units are ranked according to the size of operating income, and then systematic (equidistant) sampling is adopted to extract samples. "General self-employed households" are ranked by the sequence numbers in the "the third national economic census" data base, and then systematic sampling is adopted to extract samples. See Table 2 for the unit number in relevant strata and the sample size.

Table 2 "NOF" unit number in relevant strata and the sample size

Strata		Unit number	Sample size	Sampling method within strata
Total		497474	190	
Self-employed households	1.Big self-employed households	3099	140	Systematic (equidistant) sampling after units are ranked according to the size of operating income
	1.1 Jewelry wholesale	110	30	
	1.2Flower retail	285	30	
	1.3Mobile phone wholesale	91	30	
	1.4Other big self-employed households	2613	50	
	2.General self-employed households	494375	50	Systematic (equidistant) sampling after units are ranked by the sequence numbers

In the practical survey, sample replacement happens when sample unit refuses to obey, the data provided is undoubtedly false, or contacts cannot be reached. The principle of sample replacement is to extract the next one of the isometric sampling position of the original sample, excluding the following situations:(1) samples without "NOF" activities during the reporting period (the "NOF" amount is 0); (2) samples of extinction (closing down) (the "NOF" amount is treated as 0); (3) samples having moved out of Shenzhen (the "NOF" amount is treated as 0).

2.2.2 Parameter Estimation

Point estimation-The total "NOF" amount of self-employed households in the whole city is obtained by point estimation, and the data are obtained by adding the parameters of the "big self-employed households" stratum and the "general self-employed households" stratum. The detailed calculation formula is shown in Table 3.

Table 3 "NOF" population amount and calculation formula for parameter in each stratum

Strata		Calculation formula	Explanation for formula symbols
Total		$\hat{Y}_{st}=\sum_{h=1}^{L}\hat{Y}_h$	$\hat{Y}_{st}$: "NOF" population amount $\hat{Y}_h$: Stratified parameter $\sum_{h=1}^{L}$: Accumulated value from stratified parameter 1 to the stratified parameter L
Self-employed households	1.Big self-employed households 1.1 Jewelry wholesale 1.2Flower retail 1.3Mobile phone wholesale 1.4Other big self-employed households	$\hat{Y}_h=(\sum_{i=1}^{N_h}A_i)\cdot(\frac{\sum_{i=1}^{n_h}y_i}{\sum_{i=1}^{n_h}\alpha_i})$	A_i : Operating income of population units α_i : Operating income of sample units y_i : "NOF" amount of sample units
	2.General self-employed households	$\hat{Y}_h=N_h\cdot\bar{Y}_h$	N_h : Unit number in stratum h $\bar{Y}_h$: Sample mean in stratum h

Interval estimation-The interval estimation (95% confidence level Z distribution) of the total "NOF" amount of the self-employed households in Shenzhen can be calculated by the following formula:

$$\hat{Y}_{st}+\mu_{0.025}\cdot\sqrt{v(\hat{Y}_{st})} \tag{1}$$

$$\hat{Y}_{st}-\mu_{0.025}\cdot\sqrt{v(\hat{Y}_{st})} \tag{2}$$

The population variance of "NOF" amount of self-employed households and "big self-employed households" in Shenzhen can be calculated by the following fonnula.

$$v(\hat{Y}_{st})=\sum_{h=1}^{L}v(\hat{Y}_h) \tag{3}$$

The population variance of "NOF" amount within strata of "gold and jewelry wholesale", "flower retail", "mobile phone wholesale" and "other big self-employed households" can be calculated by the following formula:

$$v(\hat{Y}_h) = \sum_{h=1}^{L} \left(\sum_{i=1}^{N_h} A_i \right)^2 . S_h^2 / n_h \qquad (4)$$

N_h represents the total number of units within stratum h, n_h represents the number of samples within stratum h_i S^2_h represents the population variance of stratum h_i and A_i represents the total operating income of population units.

The population variance of "NOF" amount within strata of "general self-employed households" can be calculated by the following formula:

$$v(\hat{Y}_h) = \sum_{h=1}^{L} N_h (N_h - n_h) . S_h^2 / n_h \qquad (5)$$

N_h represents the total number of units within stratum h_i represents the number of samples within stratum h_i and S^2h represents the population variance of stratum h.

The variance of "NOF" amount of sample units within strata of "gold and jewelry wholesale", "flower retail", "mobile phone wholesale" and "other big self-employed households" can be calculated by the following formula:

$$S^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} \cdot (1 - \frac{n}{N}) \qquad (6)$$

The variance of the proportion of "NOF" in financing operating capital within strata of "gold and jewelry wholesale", "flower retail", "mobile phone wholesale" and "other big self-employed households" can be calculated by the following formula:

$$S^2 = \frac{\sum_{i=1}^{n} (\frac{y_i}{\alpha_i} - \frac{\bar{y}_i}{\bar{\alpha}_i})^2}{n-1} \cdot (1 - \frac{n}{N}) \qquad (7)$$

In the interval estimation formula, $\sum_{i=1}^{N_h} A_i$ is the total financing operating capital within the strata (operating income in the statistical statement by Foreign Trade and Economic Cooperation Office of the Shenzhen Statistics Bureau + turn of capital); y_i/α_i is the proportion of "NOF" in the operating capital raised of the sample units (data calculated in this survey).

2.3 The Organization and Implementation of the Survey

2.3.1 Set up a survey team to carry out the survey of sample units.

A "NOF" survey team is set up, which is composed of survey interviewers, managers of the markets where self-employed households operate and

industry associations' staffs that are familiar with the situation. The sample size of "NOF" is 190, including 140 "big self-employed households" and 50 "general self-employed households".

For the sample of "big self-employed households", a group of three people will carry out the interview and survey. First, the market manager contacts the interviewees and makes an introduction in order to dispel their misgiving; then the survey interviewer explains the purpose of the survey, the method of sampling, the significance of the authenticity of the extracted samples to the representativeness, the scope of use of the survey data and the principle of confidentiality. After the interviewees' approval, the interviewer will inquire in detail about the turnover times, capital amount, the proportion of "NOF" and other indicators. Industry associations' staff makes a judgment about their credibility.

For the sample of "general self-employed households", in order to dispel the interviewees' misgiving and raise the credibility, only market managers are entrusted to conduct the survey by applying unstructured interview (that is, market managers who are familiar with the main content of the survey chat to get the survey results without questionnaire), with focus on "NOF" amount.

2.3.2 Hold a forum to survey the industry development and financial activities.

Invite the director and professional personnel from Shenzhen Trade and Services Association, Shenzhen Gold and Jewelry Industry Association, Shenzhen Cell Phone Association and Unifortune Supply Chain Management Co., Ltd, a total of 12 people to carry on a forum for the purpose of understanding the operation status of the whole industry, the financing channel and scale of empirical data. "Shenzhen Industry Development and Financial Activities Questionnaire" will be distributed and collected in the forum.

3. "NOF" Scale Measurement and Evaluation

3.1 Distribution of Sample Values and Statistics

3.1.1 Distribution of Character Values of the Sample Unit

Among the 190 samples, 56 self-employed households had no "NOF" in 2015. Besides their owned capital, the operating capital mainly comes from formal financial institutions (banks) through mortgage (houses and goods). 134 self-employed households had "NOF", among which gold and jewelry account for the largest amount. They mainly make loans through "NOF" activities, with the amount over RMB 10 million. This is mainly due to the large amount of the gold and jewelry business: in 2015, the operating income of the 30 sample units was RMB 10.268 billion, with an average of RMB 340 million for each sample. The volume of transactions is so large that it requires

frequent "NOF" loan to raise funds. The frequency distribution of character value of "NOF" sample units is shown in Table 4.

Table 4 Frequency distribution of character value of "NOF" sampling unit

Strata		Sample number	Sample number without "NOF"	Sample number classified according to "NOF" amount(' 000 RMB)					
				0~100	100~500	500~1,000	1,000~5,000	5,000~10,000	> 10,000
Total		190	56	13	16		12	24	49
Self-employed households	1.Big self-employed households	140	35				12	24	49
	1.1 Jewelry wholesale	30	6						24
	1.2Flower retail	30	10					18	2
	1.3Mobile phone wholesale	30	7				7		16
	1.4Other big self-employed households	50	16			1	9	7	18
	2.General self-employed households	50	21	13	16				

Note: Some of the "NOF" amount is the adjusted survey data calculated by operating income of the sample unit *turn of capital x the proportion of "NOF" in operating capital.

3.1.2 Statistics of Each Stratum

In order to dispel the misgiving of sample units of "big self-employed households", this survey mainly focuses on turn of capital and the proportion of "NOF" in operating capital, without direct access to the sensitivity problem such as the operating income, operating capital and the "NOF" amount. The "NOF" amount is calculated by the following formula: total operating income in the normal statistical statement by Shenzhen Statistics Bureau÷turn of capital ^{x} the proportion of "NOF" in operating capital. Since there is no revenue data for the "general self-employed households the "NOF" amount is directly asked in the interview with "general self-employed households ^{t}'. Table 5 and Table 6 show the sampling survey statistics of "NOF" after weighted calculation.

Table 5 "NOF" sampling

Strata		Sample number	Turn of capital	Sample number with "NOF"		The proportion of "NOF" in operating capital (%)	"NOF" amount (' 000 RMB)
				Sample number	Proportion (%)		
Total		190	—	114	67.1	—	
Self-employed households	1.Big self-employed households	140	4.0	85	70.8	53.1	Indirect survey
	1.1 Jewelry wholesale	30	2.6	24	80.0	73.2	
	1.2Flower retail	30	4.2	20	66.7	35.6	
	1.3Mobile phone wholesale	30	5.7	23	76.7	41.7	
	1.4Other big self-employed households	50	4.3	34	68.0	31.5	
	2.General self-employed households	50	—	29	58.0	—	2890.00

Note: In "Total" and "Big self-employed households", the proportion of "NOF" in operating capital is obtained by weighted average

Table 6 Statistic of "NOF" sampling

Strata		Sample number	Operating capital ('000 RMB)	"NOF" amount ('000 RMB)		
				Statistic	Mean	Standard deviation
Total		190	—	—	—	—
Self-employed households	1.Big self-employed households	140	—	—	—	—
	1.1 Jewelry wholesale	30	3949203.1	2890816.7	96360.6	50022.3
	1.2Flower retail	30	478288.1	170270.6	5675.7	4213.6
	1.3Mobile phone wholesale	30	2745253.6	1144770.8	38159.0	61374.4
	1.4Other big self-employed households	50	3045666.7	959385.0	19187.7	17606.2
	2.General self-employed households	50	—	2890.0	57.8	74.7

Note: Operating capital=operating income ÷ turn of capital;
"NOF" amount= operating income ÷ turn of capital x the proportion of "NOF" in operating capital

Although the turnover times of "gold and jewelry wholesale" stratum is the least (2.6 times), the proportion of "NOF" in the operating capital is the largest (73.2%). The "NOF" calculated through the operating income of 30 sample units in 2015 (RMB 10.268 billion) is RMB 2.891 billion, accounting for 28.2% of the operating income. Table 7 shows the sampling survey statistics of the proportion of "NOF" in the operating income after weighted calculation.

Table 7 Statistic of proportion of "NOF" in operating capital

Strata	Sample number	Operating income of 2015 ('000 RMB)	Turn of capital	"NOF" amount		
				Statistic ('000 RMB)	Mean ('000 RMB)	Proportion (%)
1. Jewelry wholesale	30	10267928.0	2.6	2890817	96360.6	28.2
2.Flower retail	30	2008810.0	4.2	170270.6	5675.7	8.5
3.Mobile phone wholesale	30	15647946.0	5.7	1144771	38159	7.3
4.Other big self-employed households	50	13096366.7	4.3	959385	19187.7	7.3

3.2 Population Parameters

3.2.1 Points Estimation of "NOF"

According to the calculation formula of population parameters (point estimation is adopted for parameters of each stratum), the "NOF" of self-employed households in the jurisdiction of Shenzhen is RMB 72.417 billion. Among them, the "NOF" of the "big self-employed households" is RMB 43.842 billion, accounting for 60.5% of the total, and the "NOF" of "general self-employed individuals" is RMB 28.575 billion, accounting for 39.5% of the total. The results of the "NOF" parameters of each stratum are shown in Table 8.

Table 8 Points Estimation of "NOF" Amount

Strata		"NOF" (billion RMB)	Indicators				
			Operating income (billion RMB)	Turn of capital	Operating Capital (billion RMB)	Proportion of "NOF" in operating capital (%)	"NOF" mean of sample ('000 RMB)
Total		72.417	—	—	—	—	—
Self-employed households	1.Big self-employed households	43.842	500.289	4.2	118.590	36.97	—
	1.1 Jewelry wholesale	9.443	33.539	2.6	12.900	73.2	—
	1.2Flower retail	1.564	18.447	4.2	4.392	35.6	—
	1.3Mobile phone wholesale	3.789	51.791	5.7	9.086	41.7	—
	1.4Other big self-employed households	29.047	396.512	4.3	92.212	31.5	—
	2.General self-employed households	28.575	—	—	—	—	57.8

Note: For "big self-employed households" , points estimation of "NOF"=operating income÷turn of capital x the proportion of "NOF" in operating capital : For "general self-employee households", points estimation of "NOF"="NOF" mean of sample x population unit number.

3.2.2 Interval Estimation of "NOF"

Basing on the 95% probability to calculate the parameter interval of the "NOF" of both the population and each stratum, it is estimated that the total "NOF" amount of the self-employed households is at least RMB 56.564 billion and at most RMB 88.27 billion. From the perspective of variance, "gold and jewelry wholesale", "flower retail" and "mobile phone wholesale" have relatively small variance due to operational characteristics. Without further stratification, the variances of "other large self-employed households" and "general self-employed households" are very large. So it can be seen that the diversity of commodities is related to the difference of "NOF" to some extent. The results of the parameter interval estimation of "NOF" are shown in Table 9

Table 9 Interval Estimation of "NOF" Amount

Strata		Points Estimation	Interval Estimation		Indicators		
		(billion RMB)	Lower limit (billion RMB)	Upper limit (billion RMB)	"NOF"population variance	"NOF"standard error (billion RMB)	"NOF"limit error (billion RMB)
Total		72.417	56.564	88.270	6542.32	8.088	15.853
Self-employed households	1.Big self-employed households	43.842	31.736	55.948	3814.64	6.176	12.106
	1.1 Jewelry wholesale	9.443	7.670	11.216	81.87	0.905	1.773
	1.2Flower retail	1.564	0.840	2.288	13.65	0.369	0.724
	1.3Mobile phone wholesale	3.789	2.454	5.124	46.39	0.681	1.335
	1.4Other big self-employed households	29.047	17.169	40.925	3672.74	6.060	11.878
	2.General self-employed households	28.575	18.338	38.811	2727.68	5.223	10.237

3.3 Evaluation

3.3.1 Comparison with the Data from Industry Associations

The survey choose Shenzhen Gold and Jewelry Industry Association (with more than 700 members and over RMB 180 billion operating income in 2015), Shenzhen Service and Trade Association (with 735 members, and over RMB 100 billion operating income) and Shenzhen Mobile Phone Industry Association (with more than 1200 members, and over RMB 1 10 billion operating income) at the same time in order to understand the average turnover, the proportion of "NOF" in operating capital, and so forth of their members in 2015. Three industry associations were entrusted to use Delphi Method (expert-opinion-method) to collect data and fill out the "Shenzhen Industry Development and Financial Activities Questionnaire". The data collected by the industry associations is basically consistent with the survey data (see Table 10)

Table 10 Comparison between data from "NOF" survey and data from industry association survey with Delphi method

Strata		Interviewer-survey-method		Delphi method	
		Turn of capital	The proportion of "NOF" in operating capital (%)	Turn of capital	The proportion of "NOF" in operating capital (%)
Big self-employed households	1. Jewelry wholesale	2.6	73.2	2~3	60~80
	2.Flower retail	4.2	35.6	4~5	30~40
	3.Mobile phone wholesale	5.7	41.7	5~7	35~45

Note: "NOF" survey data is collected with Interviewer survey method.

3.3.2 Comparison with the Data of Loan Obtained from the Banking Regulatory Department

According to the statistics of Shenzhen Banking Regulatory Department, the lending balance of financial institutions to the self-employed households (formal financial institutions lending) was RMB 63.594 billion at the end of 2015 and RMB 78.764 billion on June 30, 2016 respectively. The estimated "NOF" amount in this survey is RMB 72.417 billion. It can be seen that the "NOF" amount of the self-employed households is no less than the amount lend from formal financial institutions.

Table 11 Loans from financial institutions in Shenzhen

Unit: billion RMB

Indicators	Code	2015	First half of 2016
Loan balance in Shenzhen	1	3242.922	3698.399
Among: Loan balance to small and micro enterprises	2	328.892	364.549
Loan balance to small and micro enterprises owners	3	101.001	120.007
Loan balance to self-employed households	4	63.594	78.764

4. The Value-added Accounting of Self-employed Households' "NOF"

In this survey, the data sources of "NOF" loan activities of self-employed households are widespread, and the same survey subject has a variety of channels of loan in different periods or in the same period, so it is difficult to

collect the proportion of "NOF" loan capital according to the source channels in the survey.

4.1 Financial Intermediary Services by Indirect Measure

From the perspective of national accounting, this kind of capital borrowing and lending is the output of financial intermediary services, which does not directly charge service fees, but actually produces financial services related to deposit and lending interest fees. SNA provides a method to indirectly measure the financial intermediary services (FISIM).

SNA2008 improves the calculation method of FISIM. It is suggested to use the reference-rate-method to calculate FISIM for all deposit and lending (including self-owned capital), and to presume that all deposit and lending services provided by financial institutions have been charged indirect service fees, regardless of the source of capital. At the same time, it is concluded from SNA that financial institutions do not necessarily provide both deposit and lending services. Therefore, based on SNA2008, it is suggested that an unincorporated lender with self-owned capital should also be regarded as a financial institution that provides lending services, while the output of its lending services should be regarded as the virtual lending service fee according to the calculation method of FISIM.

4.2 The Value-added Accounting of Self-employed Households' "NOF" of Shenzhen in 2015

For lenders with their own capital, lending service output FISIM = lending service fee = lending amount x (lending interest rate-reference rate). The added value of lending activities of the self-employed households' "NOF" is calculated as follows. First, it is assumed that:

1. The " annual average NOF balance" of self-employed households in 2015 is RMB 72.417 billion in this survey (this data should be time points numbers, and the lending balance at different time points throughout the year changes; thus it makes it simply by assuming that the self-employed households' "NOF" scale throughout the year is basically stable).
2. The lending interest rate of the unincorporated lender refers to the annual interest rate of the bank for the personal unsecured credit lending, which is about 9%.
3. The reference interest rate is weighted by the average deposit and lending balance of Shenzhen in 2015, and the one-year deposit and lending benchmark interest rate of the People's Bank of China at the end of the year.

$$\text{Reference interest rate} = \frac{56725.61}{30248.86 + 56725.61} \text{X} 1.5\% + \frac{30248.86}{30248.86 + 56725.61} \text{X} 4.35\% = 2.49\%$$

Based on the FISIM calculation method and the assumptions above, the output and added value of "NOF" activities of self-employed households in 2015 are calculated as follows:

Lending service output = 72.417 X (9%-2.49%) =RMB4.714 billion
Interval lower limit: 56.564 X (9%-2.49%) = RMB 3.682billion
Interval upper limit: 88.270.X (9%-2.49%) = RMB 5.746billion

For unincorporated lenders, the intermediate consumption of such capital lending activities is small. Assuming the added value rate is 95%, then:
Added value = Lending service output x 95% = RMB 4.479 billion
Interval lower limit: 3,682 x 95% = RMB 3.498 billion
Interval upper limit: 5.746 x 95% = RMB 5.459billion

From the financial industry data of Shenzhen, the lending balance at the end of the 2015 was RMB 3244.904 billion, and the "NOF" amount of the self-employed households was RMB 72.417 billion, with 2.2% as a ratio of the two figures. The added value of monetary and financial services in Shenzhen was RMB 141.192 billion, and the added value of self-employed households' "NOF" was calculated to be RMB 4.479 billion, with 3.2% as a ratio of the two figures.

5. Conclusion and Prospect

In this paper, the direct-survey-method is adopted to study the "NOF" amount of self-employed households in Shenzhen, and the added value of "NOF" is calculated. The result shows that the total "NOF" amount of the self-employed households in Shenzhen is RMB 72.417 billion, the interval lower limit is RMB 56.564 billion, and interval upper limit is RMB 88.27 billion, which is basically consistent with the data collected by the industry association and the banking regulatory department The added value of "NOF" of the Shenzhen's self-employed households was estimated at RMB 4.479 billion by the FISIM method.

In recent years, with the raising of resident income and the advance of internet technology, the innovation and production mode of the financial activities develop with diversification and complication. Since folk loan in China is expanding, the capital amount of the personal loan activities, as well as the effect on social financing scale and financial system by personal loan activities have grown constantly. It is suggested that "NOF" should be included in the financial activities accounting. Survey of "NOF" should be carried on, where survey of "NOF" should be further carried out in small, micro-sized enterprises and households. Surveys of "NOF" activities in small, micro-sized enterprises and households are supposed to be more complex, because such economic units are more sensitive to talking about "NOF". As for the households' "NOF", it is advised to design relevant indicators in the Urban and Rural Household Survey to conduct a pilot survey to obtain the proportion and composition of households "NOF". For the purpose of improving the accuracy of "NOF" interval estimation in Shenzhen, we must study the stratification characteristics of various economic units further,

develop more scientific sampling methods, design survey tools with high validity, and utilize more flexible survey methods and interview techniques to improve the reliability.

Reference

1. OECD/IMF/LO/CIS STAT. Measuring the non-observed economy: a handbook [M] France, 2002: 10-12
2. United Nations. Economic Commission for Europe, Non-observed Economy in National Accounts: Survey of National Practices, 2003
3. Tian Guangning. Research on non-observed financial and monetary equilibrium [M]. China finance press, 2008.
4. Li Jianjun Non-observed monetary and financial condition index and economic prosperity index in China -- empirical study on theoretical design and internal relationship [J] Finance and trade economics, 2008(7).
5. Li Jianjun. Design and measurement of China's "non-observed finance" indicator system [J]. Research on quantitative economy and technical economy, 2010(5).
6. Ren Biyun, Gao Zhiyan, Zhang Tongjin, Analysis of the influence of non-observed finance on money supply [J]. Joumal of Beijing technology and business social science edition, 2013(2).
7. Li Jianjun, Non-observed currency size estimation based on currency absorption analysis and GDP revision data [J]. Finance and trade economics, 2006(6): 1978-2005.
8. Jiang Xuchao, Ding Changfeng. Theoretical analysis of folk finance: category, comparison and institutional change [J] Financial research, 2004(8).
9. Li Jinchang Xu Aiting. A new approach to non-observed economic estimation Statistical research, 2005(11).
10. Xu Aiting, Li Jinchang China's non -observed economic size -- a new discovery based on MIMIC model and economic census data [J], Statistical research, 2007(9).
11. Xu Aiting. Estimation of non-observed economic size: applicability and innovation of income and expenditure difference method [J]. Statistical research, 2008(12).

Plutus – A new tool to standardise the metadata of seasonal adjustment

Tímea Baczakó, Mária Pécs
Hungarian Central Statistical Office

Abstract

Standardisation and process-driven developments of methodological and information technological services have always existed at the Hungarian Central Statistical Office, even if they have not been common and uniform for the whole institution or the whole statistical business process chain. The philosophy of this Business Architecture model is cross-cutting, thus it covers all subject-matter domains, their data and metadata flows in an integrated way. The driving force behind using standards in the Hungarian Central Statistical Office is to support the efficient operation of the organisation and to foster harmonisation of its statistical activities and also the statistical and the non-statistical (supporting) activities. The Hungarian Central Statistical Office currently concentrates on the modernisation and operationalisation of its BA and considers it as a main driver for the modernisation activities. As part of this modernisation, the Hungarian Central Statistical Office revised its previous business process model according to the Generic Statistical Business Process Model ver. 5.0; the Hungarian adaptation of the Generic Statistical Business Process Model, called 'Egységes Statisztikai Folyamat Modell' in Hungarian was created. After this, in the last year the seasonal adjustment, one of the sub-processes of 'Egységes Statisztikai Folyamat Modell', was part of the modernization and standardisation. The focus was on the handling metadata of seasonal adjustment. Seasonal adjustment is an integrated step of the statistical business process in the official statistics, therefore a proper system that guarantees the high quality results is important for every National Statistical Institute. The aim of Plutus, the system presented in this paper could support seasonal adjustment in more integrated way in a National Statistical Institute with a standard metainformation system in place. Quality of metadata can be assured among others when they are up-to-date, comparable, consistent, available, standardised, related, complete, and understandable.

The Hungarian Central Statistical Office has more than 40 years of history on the field of metadata integration where development is never finished, but we can stop at a certain point at take a snapshot of the current state-of-the-art. Creative thinking for developing and implementing standards and general overview of the integrated system is necessary to conduct the methodological improvements.

Keywords
seasonal adjustment; modernization; business architecture; metadata

1. Introduction

The Hungarian Central Statistical Office constantly develops process-oriented operations and standardises its Business Architecture.[1] (UNECE, 2015) Integrated methods, integrated methodology and integrated tools are aimed to provide the standardisation for various statistical subject-matter domains and for the whole system of official statistics.

The Hungarian Central Statistical Office has been involved in important standardisation projects, e.g. ESSnet on Standardisation (ESSnet on Standardisation, 2015) and one of the UNECE Common Statistical Production Architecture Catalogue team (Catalogue of Common Statistical Production Architecture services, 2014). In 2015 the management of the Hungarian Central Statistical Office has accepted the enterprise architecture (based on Common Statistical Production Architecture), and since then we are working on its implementation.

The aim of this paper is to describe a new tool, called Plutus that handles the metadata of seasonal adjustment. Plutus is a good example for the standardisation a sub-process of the statistical data production business process.

2. Business Architecture and Seasonal adjustment

The Hungarian Central Statistical Office is constantly moving towards a process-oriented operation and standardising its Business Architecture. The Hungarian Central Statistical Office has already begun to move towards a more standardised BA. The adaptation of the UNECE standard Generic Statistical Business Process Model ver. 5.0 to the Hungarian needs (called 'Egységes Statisztikai Folyamat Modell') a few years ago was the major first step to realise this strategic goal (HCSO, 2018). Since the first version of the 'Egységes Statisztikai Folyamat Modell' many developments have started to move the statistical business processes towards process-orientation. This Hungarian Generic Statistical Business Process Model is the model for statistical business processes, by which we, at the Hungarian Central Statistical Office mean the statistical data production process as well as processes related to statistical registers' maintenance and survey frame production. This model is also suited for describing other statistical processes like classifications'

[1] 'Business Architecture covers all the activities undertaken by a statistical organization, including those undertaken to conceptualize, design, build and maintain information and application assets used in the production of statistical outputs. Business Architecture drives the Information, Application and Technology architectures for a statistical organization.'

maintenance, producing questionnaires, data files, publications, etc, but only to those processes, which are directly related to statistical data.

Practical areas of use of the Hungarian Generic Statistical Business Process Model are methodological standardization, process documentation, process monitoring, process management, process optimization, quality measuring. Standardised methodologies can be developed based on existing practices. Furthermore existing solutions and creating methodological resolutions for potentially missing elements can be standardised. With the process optimisation the permanent process can be improved (even reorganised) in favour of efficiency and success.

The Hungarian Central Statistical Office considers the common statistical business process model the core of its Business Architecture and is currently investing a lot of resources on ongoing organizational, process-management, methodological and IT activities.

This paper presents a methodological development from this modernisation activity for the field of seasonal adjustment. The Hungarian Central Statistical Office implemented modernization in seasonal adjustment which contributed to the use of integrated, standard applications.

The comprehensive goal of the Hungarian Central Statistical Office is to create generic and independent applications (keeping in mind the principles of service-oriented architecture) for the support of the data process phases (horizontally) and this was also the case with Plutus.

The Hungarian Central Statistical Office's integrated, standard information technology systems are mostly metadata-driven systems and cooperate with each other as shown on Figure 1. The Plutus is not yet a metadata-driven tool, but in the future it is planned to be developed according to this principle. This development helps the standardisation of the sub-process and improves the quality of seasonal adjustment.

During the modernization this standard information technology tool (Plutus) has been developed. Plutus allows the subject-matter domains to manage the information of time series through this standard application instead of a questionnaire. Plutus is a system of databases with a user friendly interface, which fully uses metadata.

The purpose is to ensure the controllability of the maintenance through proper coordination. The main aspects of the overview is the supervision of the technical requirements, the content and the main observation factors. These are the completeness and the correctness of the metadata.

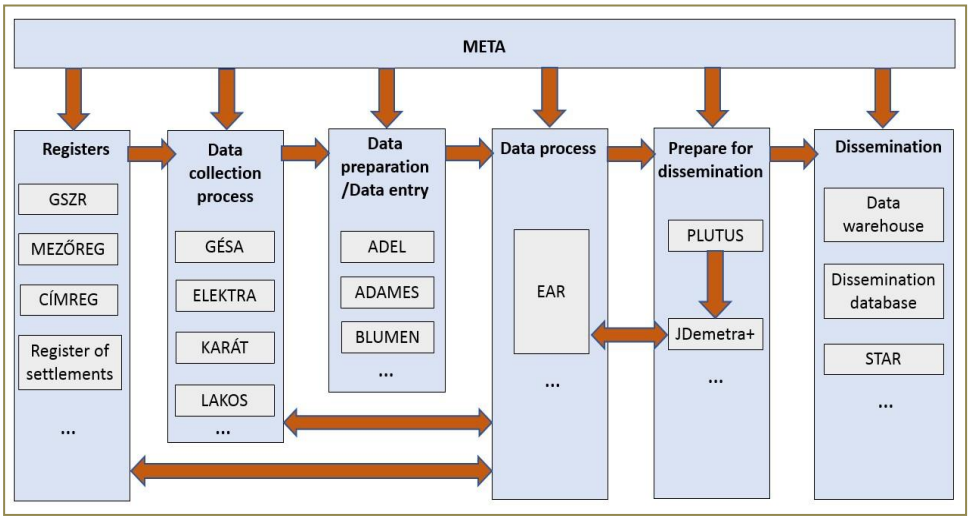


Figure 1 – Connection of integrated information technology systems of the Hungarian Central Statistical Office

The integrated information technology systems, aimed to support one or more sub-processes or process phases of the statistical business process can be positioned according the statistical business process model, more or less the same way as methods. This makes the connections between the methods and the information technology tools more self-evident as information technology systems are expected to implement the methods described by the methodological framework.

This is also a great opportunity for statisticians to use information technology tools that are fully in line with the guiding methodology. From methodological perspective, the integration of methods into the information technology systems and their use in practice increases the practical relevance of the methods proposed.

Plutus indirectly connects with JDemetra+ (Grudkowska, S., 2016), the experts of the Methodology Department are needed in the management of the process. In the Plutus application the subject-matter domains update the key information of the specific time series. Based on these given metadata the methodological experts make the annual revision of the models with software JDemetra+. Furthermore during the year the data and models are refreshed by the subject-matter domains, supported by the methodological experts.
In this process all participants are key players, the information technology tool Plutus cannot replace the professional knowledge in itself, but it is necessary for the effective, punctual, standardised, good quality work.

The Hungarian Central Statistical Office also considers this is a good example that the methods and information technology solutions need to go

hand-in-hand by providing in-house architectural services for the subject-matter statisticians.

3. Details on the seasonal adjustment process

Seasonal adjustment (Mazzi, G. L. (2018), HCSO (2017), EUROSTAT (2015a)) is an integrated step of the (Hungarian) Generic Statistical Business Process Model in official statistics, therefore a proper system that guarantees the uniform and high quality results is important for every National Statistical Institute.

In general we can say that seasonal adjustment needs two types of knowledge: the mathematicalstatistical knowledge which is extraordinarily important for time series analysis, and the sound knowledge of the specific time series. In an optimal world one person owns these knowledge but commonly this is not the case.

In Europe there are two main scenarios to handle seasonal adjustment. One is when one person actually owns the required knowledge (mathematical-statistical and domain-specific) and can conduct the statistical business process by himself/herself from the beginning until the publication. This solution does not guarantee standard, accepted solutions within the entire institute. The other one is the scheme of having experts of the subject-matter domain fields who know without a doubt their time series and experts who know everything about time series analysis (they are generally methodologist).

At the Hungarian Central Statistical Office the experts of the Methodology Department work firmly together with the experts from the subject-matter domain departments during the whole seasonal adjustment process (second scenario above). Methodological experts own the mathematical, statistical knowledge on time series analysis and the subject-matter experts know every deep detail on the subject matter and their time series. This means that the Hungarian Central Statistical Office performs a centralized internal system in case of seasonal adjustment process.

The division of labour has two big runs in every year. At the beginning of a given year, before the first publication of seasonally adjusted data of the year, the whole annual revision of the models is made in case of every single time series, with the help of software JDemetra+. The Methodology Department is responsible for this task but key information is needed from the subject-matter domains on the time series. During the year the subject-matter domains update the data and models, however, this part of the work is also supported (checked, and if needed, modified) by the Methodology Department.

Figure 2 is an understandable flow chart representing our system during a year in case of one time period where blue means tasks for the subject-matter domain departments, green means tasks for the Methodology

Department and yellow means tasks for the Dissemination and Publishing Department:

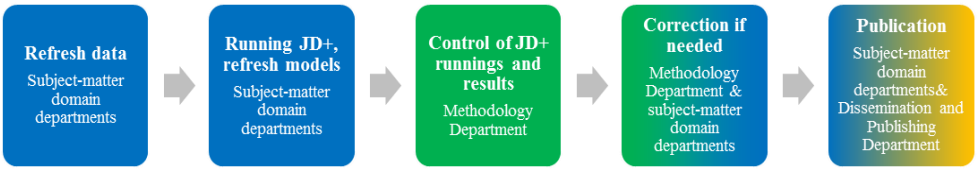


Figure 2 - Understandable flow chart about the centralised system during a year in case of one time period

How can we ensure quality in this system in case of seasonal adjustment? Until now, at the beginning of every year the Methodology Department sent a questionnaire to the subject-matter domains to collect as much information as possible on their time series to prepare for the annual model revisions. Year after year the Hungarian Central Statistical Office have created more and more time series and at the same time more and more information on them. This old system cannot or hardly be managed anymore. The main aim of Plutus is to help the collaboration of colleagues from different departments, to solve the problem of increased data set and to further standardise the process. Exceptionally during the annual model revisions.

Plutus is a system of database tables with a user friendly interface, which fully uses metadata. In this way everything is fully trackable. We have information on who modified the data and when. The biggest advantage of this whole system is that the whole process is more transparent than before. Now every user has reading rights to all information which makes their work easier than before. This also mean that in every year at the beginning the users do not have to fill a new questionnaire, just update the last year's information which results in less work. The developed system has a lot of automated options such as sending notifications which also make easier the workflow.

These steps lead us to guarantee the more complete service-oriented architecture. In addition this also increases the amount of metainformation on seasonally adjusted time series. For example there is a possibility for the users to upload the domestic and international standards linked to their time series. The collected metainformation can be linked to any other metadata object like descriptive information on subject-matter domains or any publication system to inform our users more completely. We can publish more detailed information on our time series and their seasonal adjustment methods.

The whole system of Plutus will be connected to the Single Integrated Metadata Structure (EUROSTAT, 2015b) too which connection is really important in case of linking metainformation and thus quality management.

The Hungarian Central Statistical Office have future plans too to use these in-house data to inform their users more effectively. Unfortunately now

we do not publish any deep details about our seasonally adjusted time series but with the Plutus we have the chance to perform these data. For example any model information or reasons of the outliers can appear on the website. And the best part is that it does not indicate more time investment from our colleagues. Simply we need to develop a system which is read and implement these metainformation from Plutus's database tables to our publications. The main advantage of Plutus is that until now the main methodological parts of prepare for dissemination process of Hungarian Generic Statistical Business Process Model were not adapted in a metadata system such as apply disclosure control and produce seasonally adjusted outputs. Now the last one uses metadata and that is a big step ahead for our institute.

References

1. UNECE (2015). Business Architecture. *Common Statistical Production Architecture v1.5*. UNECE. https://statswiki.unece.org/display/CSPA/Business+Architecture Accessed on 15. 01. 2019.
2. *ESSnet on Standardisation (2015)*. https://ec.europa.eu/eurostat/cros/content/standardisation_en Accessed on 15. 01. 2019.
3. *Catalogue of CSPA services (2014).* https://webgate.ec.europa.eu/fpfis/mwikis/cspacatalogue/index.php/CSPA_catalogue Accessed on 15. 01. 2019.
4. HCSO (2018). *Hungarian Generic Statistical Business Process Modell (HGSBPM) v2.3.,* Budapest: HCSO. http://www.ksh.hu/docs/bemutatkozas/eng/estfm_eng.pdf Accessed on 15. 01. 2019.
5. Grudkowska, S. (2016). *JDemetra+ Reference Manual Version 2.1.,* Narodowy Bank Polski. https://ec.europa.eu/eurostat/cros/system/files/jdemetra_reference_manual_version_2.1_0.pdf Accessed on 15. 01. 2019.
6. Mazzi, G. L. (ed.) (2018). *Handbook on Seasonal Adjustment*, Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/documents/3859598/8939616/KS-GQ-18-001EN-N.pdf Accessed on 10. 01. 2019.
7. HCSO (2017). *About Seasonal Adjustment,* Budapest: HCSO. http://www.ksh.hu/docs/eng/xftp/modsz/eszezonalis_kiig.pdf Accessed on 15. 01. 2019.
8. EUROSTAT (2015a). *ESS Guidelines on Seasonal Adjustment*. EUROSTAT Manuals and Guidelines. https://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf Accessed on 15. 01. 2019.

9. EUROSTAT (2015b). *Single Integrated Metadata Structure v 2.0 (SIMS v2.0) and Its Underlying Reporting Structures. The ESS Quality and Reference Metadata Reporting Standards ESMS 2.0 and ESQRS 2.0*, Luxembourg: EUROSTAT. https://ec.europa.eu/eurostat/documents/64157/4373903/SIMS-2-0-Revised-standards-November2015-ESSC-final.pdf Accessed on 15. 01. 2019.

Handling technological changes by time varying coefficient model analysis in flash estimate of gross value added in information and communication industry

Klaudia Máténé Bella, Ildikó Ritzlné Kazimir
National Accounts Department, Hungarian Central Statistical Office, Budapest, Hungary

Abstract

This paper investigates the linkage of physical indicators as expression of technological changes and gross value added of the industry telecommunication by time varying coefficient model. Empirical results for quarterly data from 2000q1 to 2018q3 in Hungary indicates that the physical indicators follow a logistic curve and they change each other overlapping with technological progress. We find the relationship that the most physical indicators effect significantly the gross value added of industry telecommunication until their growth reaches the inflection point. In each sub period one or two variables could be detected to be driving force behind the growth. The goal was the construction of a model that enables the forecasting of gross value added of telecommunication using the selected physical indicators. We argue that time varying coefficient model is able to handle technological changes and the quick changes of the explanatory power of exogenous variables. With the handle the full time series time varying coefficient model is an effective method to flash estimate of gross value added of the information and communication industry.

Keywords

Logistic curve; state space model; GDP estimation

1. Introduction

In the flash gross domestic product (GDP) estimation the Hungarian Central Statistical Office (HCSO) applies a bottom up approach in the production side (Cserháti et al. (2009)). The HCSO utilizes the most available physical indicators, and fits autoregressive integrated moving average (ARIMA) models completed with explanatory variables usually. In the case of information and communication industry the physical indicators have low explanatory power separately, and the significant multicollinearity hinder the accurate estimation of gross value added. The strange and unusual relationship between physical indicators is due to the rapid technological changes in whole economy.

A radical change in circumstances of production results changes in different areas of economic environment, the paradigms of production can transform. For example, a significant new technology requires new or

developed infrastructure, standards, changes in market structure, and so one. From the eighties to the middle of second decade of 21th century had developed the automatization, this period is called the third industrial revolution. The automatized technologies in industry resulted a cross-fertilized development in computer networks, telecommunication, mobile internet and industrial communication. The process influenced the business models, the globalization, as well, and it was a technological driving force in development of consumer electronics.

Since the second decade of 21th century the new technologies have spread in fields of cyber physical systems, the internet of things and developed networks. This radical new technology in industry is called industry 4.0, and it requires real time solutions in telecommunication systems. (Wollshlaeger et al. (2017))

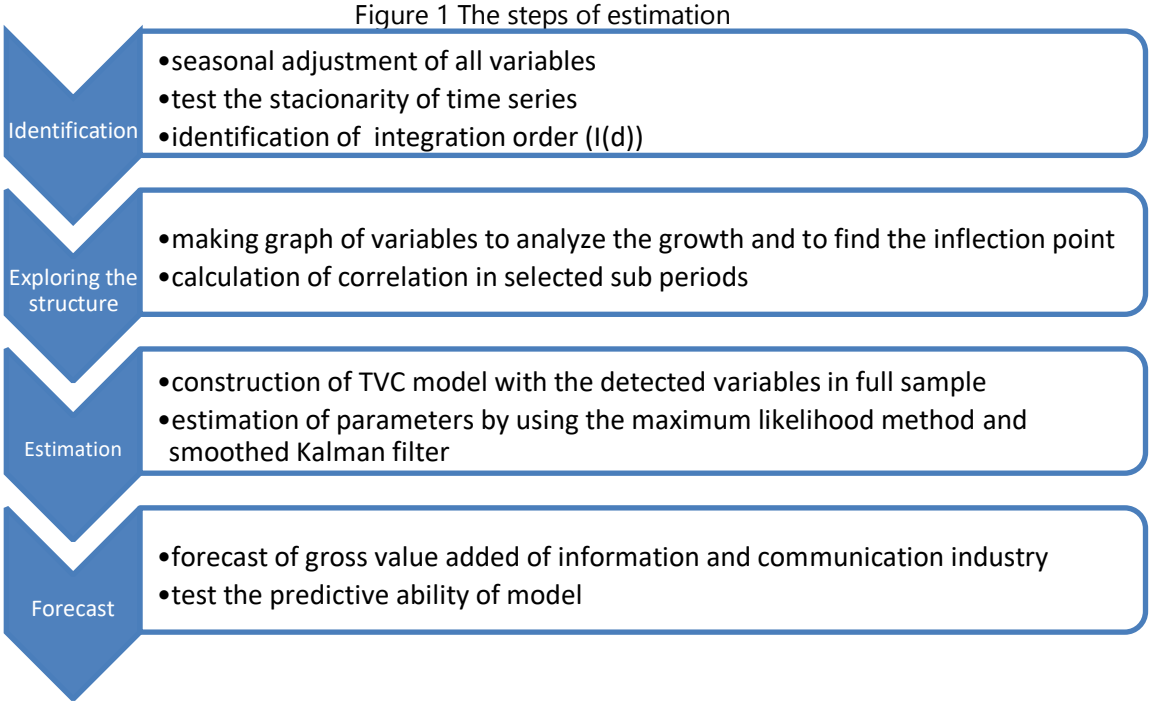
The new technologies penetrate usually along an s-curve (logistic curve). The innovation spreads out at a relatively low speed, because it needs time to develop the new technology, and the enterprises, markets have to adopt to the new circumstances. However, the adjustment results increasing penetration rate of the new technology, until the all opportunities will be utilized. In this case the limit of growth is reached, and the increase of profit can be realized by adaptation of other new technology. The time between technological changes is getting shorter. (Brynjolfsson & McAfee, (2014))

Due to the above mentioned processes the most physical indicators of telecommunication networks have explanatory power only in a short period until the growth of their spread reaches the inflection point. To all sub periods can be specified one or two variables that had a significant impact on the growth of gross value added. Because of the change of the explanatory power, time varying coefficient model is selected to handle the time series and to forecast. The paper shows that this approach for estimation has a more accurate predictive ability compared to the previous applied model.

2. Methodology

If the explanatory variables change from period to period, the correlation between physical indicators and gross value added of information and communication should be calculated to detect the most suitable variables for all sub periods. If the variables belonging to sub periods are confirmed, the construction of time varying coefficient model (TVC) model is appropriate.
The fit of TVC model should follow the steps included in the Figure 1:

Figure 1 The steps of estimation



This paper follows the above steps to determine the TVC model. First, the series have to be seasonally adjusted by X11 method and tested whether they are I(0). A series X_t is said to be integrated with order d, written I(d) means that it needs to be differenced d times to make it stationary. The inflection points of physical indicators help to determine the intervals where this indicators have a significant effect on the gross value added. To the accurate determination of intervals was applied the rolling window method.

On the basis of second step analysis and Figure 1, six variables would be selected for the TVC model. The TVC model is a kind of state space models. Our fitted model follows the model of Hall, Swamy and Tavlas, and it is described by the two types of equations. (Hall et al. (2014)) First the basic equation is determined, in this formula the parameters depend on time. In the state space model this equation is the signal:

$$y_t = \beta_{0t} + \beta_{1t} x_{1t} \quad (1)$$

The state equations are described for the parameters of the signal equation. In the model of Hall et al. it is the driver equations.

$$\beta_{0t} = \pi_{00} + \sum_{i=1}^{p-1} \pi_{0i} Z_{it} + \varepsilon_{0t} \quad (2)$$

$$\beta 1t = \pi_{10} + \pi_{1p+1} x_{1t} + \sum_{i=1}^{p-1} \pi 1 i Zit + \varepsilon 1t \quad (3)$$

The advantage of state space model is that describes linear connections between variables and can handle non-linearity of variables, therefore it is suitable for the estimation of gross value added in information and communication with taken into consideration the changes in physical variables.

3. Data

The GDP flash estimation in Hungary built on bottom up approach by production side, and fit autoregressive models with explanatory variables on ten sections (A10) breakdown of Statistical

Classification of Economic Activities in the European Community (NACE). (Cserháti et al. (2009)) The information and communication industry had 4.3% share in the GDP in 2017, and it included the branches which are listed in the Table 1.

Table 1 The sections of information and communication industry

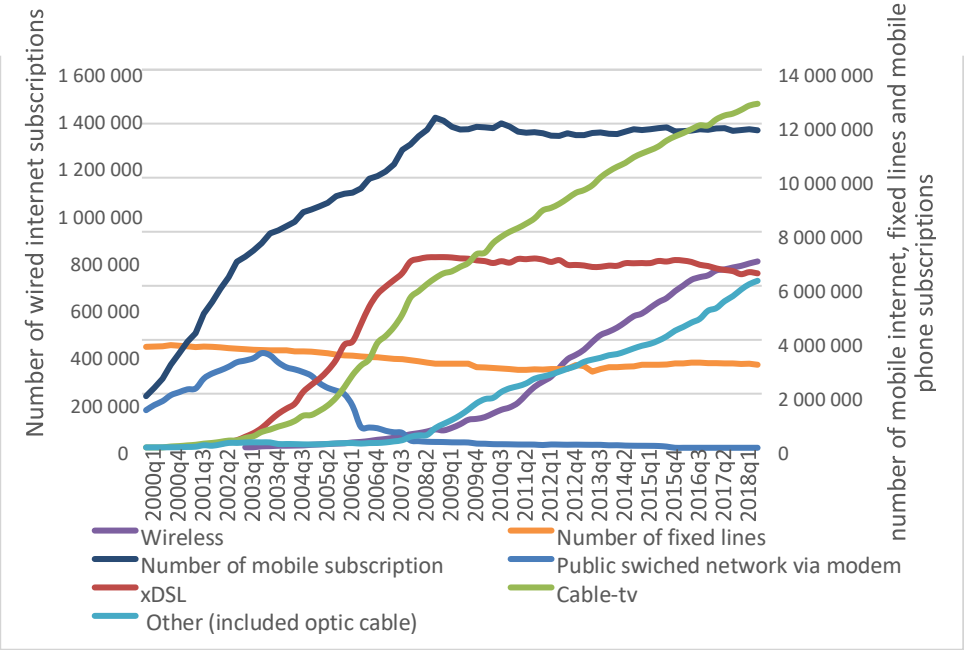
J58.. - Publishing activities
J59.. - Motion picture, video and television programme production, sound recording and music publishing activities
J60.. - Programming and broadcasting activities
J61.. - Telecommunications
J62.. - Computer programming, consultancy and related activities
J63.. - Information service activities

The first three activities include activities which are connected to the media activities, and the last three include the activities relating telecommunication and information technology. Therefore the sections were grouped. The 61-63 sections are estimated utilizing the physical indicators of information technology and telecommunication by TVC model. The estimated result will be built in the autoregressive model of information and communication. The 61-63 sections had a share of 78% of gross value added in information and communication industry in 2017.

The available physical indicators such as mobile phone minutes, landline phone minutes, number of pieces of internet in connected lines, other internet access, wireless, x digital subscriber line (xdsl), cable TV and the number of employees in information and communication industry seem to be good proxies to forecast of gross value added of the Information and communication industry. Based on preliminary examination two main problems are faced. The first is that in some periods are the physical indicators significant and in other periods appear they to be insignificant because of structural breaks. The development of physical indicators cause two problems. The first problem occurs due to the substitution of telecommunication channels. The second problem is related to the diffusion of new technologies with the further existence of previous technologies.

The spread of telecommunication networks in Hungary is a good example to show the penetration of new infrastructure related to cross-fertilization of technological changes. The Figure 2 draws up the number of wired, mobile internet subscription, as well the number of fixed and mobile phone lines.

Figure 2 Number of internet and phone subscriptions in Hungary



Source: HCSO

Some physical indicators follows a logistic curve. The above figure shows that the telecommunication channels – except of wireless, other (included optic cable) and cable-tv internet subscriptions– reached their saturation. The saturation in case of wireless internet may be expected in the near future, but the 5th generation wireless systems (5G) technology can give new impetus to growth. The 5G is expected to become widely available in Hungary in 2020-2022.

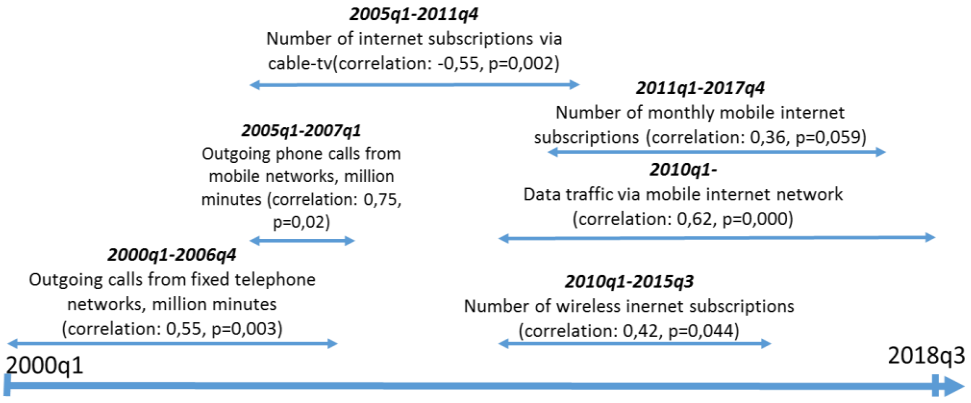
The 5G technology is appropriate for real time, mobile data traffic. However, the possibilities of 5G line are not only in its speed, but in its integrated network intelligence and its collaboration with other technologies. The new technology can bring significant improvements, for example in health, transport, education and industry. (Portfólió (2018b))

The high quality optical cable network is a precondition for 5G. Therefore the significant growth in optical cable networks is currently characteristic in Hungary, over the next ten years, Hungary will be connecting more than 100,000 kilometres with new optical cables. This improvement can cause an additional increase in cable tv subscriptions, which is one of the typical form of households' internet subscription in Hungary. (Portfólió (2018a))

The internet subscription forms can substitute each other in a certain extent, but it is complementary relation between wired and wireless network as well. Moreover, the changing in user needs causes different penetration rates and structure of telecommunication channels.

The analysis of correlations of gross value added in 63-61 industry and each physical indicator, we detected the following overlapped structure of dependency. The Figure 3 shows the determining periods of telecommunication services.

Figure 3 Penetration of telecommunication channels (p means the level of significance)



The graphs of the variables can help to understand the processes and includes the variables for TVC model.

4. Result

The analysis and estimation were prepared in Eviews software. As a first step we have investigated the seasonally adjusted time series whether variables have a unit root. According to the Augmented-DickeyFuller test, all first difference of variables have a unit root, but the second difference of the variables are stationary. Because of overlapping of technologies have some sub period defined which are proved by the correlations between the physical indicators and the gross value added.

As a second step we have constituted a TVC model among the variables to handle the structural breaks and so the change of explanatory power of exogenous variables. The equations in the TVC model are described as follow:

$$gva_t^{61-63} = \beta_{0t} + \beta_{1t} subscriptions_{1t} \quad (4)$$

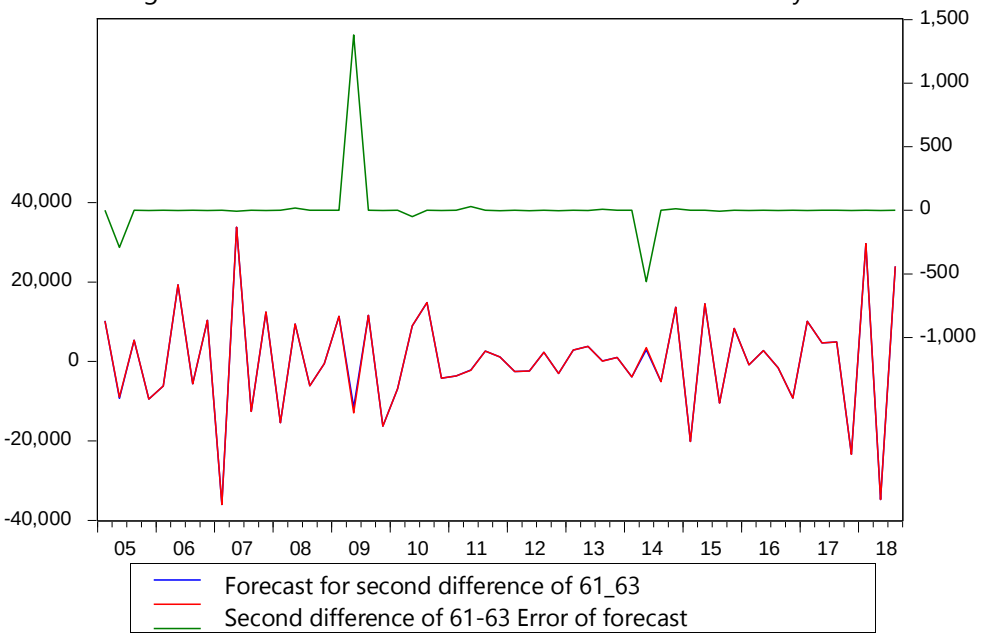
$$\beta_{0t} = \pi_{00} + \sum_{i=1}^{p-1} \pi_{0i} Z_{it} + \varepsilon_{0t} \quad (5)$$

$$\beta 1t = \pi 10 + \pi 1p + 1 subscriptions 1t + \sum_{i=1}^{p-1} \pi 1i Zit + \varepsilon 1t \quad (6)$$

where $gva^{61\text{-}63}$ is the gross value added in the industries 61-63, the subscriptions include the sum of all kind of telecommunication subscriptions. Z_i 's contain the explanatory variables from Figure 3.

As a last step we have made a forecast for the period 2005q1-2018q3. The Figure 4 illustrates the actual and the forecasted value of the gross value added of information and communication industry at constant price HUF 20005 in Hungary.

Figure 4 The actual and estimated value added of industry 61-63



We argue based on this result, the gross value added of information and telecommunication (q61_63) presents a good forecast efficiency. The results show that the crisis in 2008-2009 and the recovery at the beginning of 2014 means a relative significant challenge for the forecast algorithm.

5. Discussion and Conclusion

The current applied flash estimation based on bottom-up concept with ARIMA models faces significant problems in the case of rapid economic and technological challenges and considerable multicollinearity of time series. Both problem appear in the information and communication industry in Hungary. The rapid changes in telecommunication technologies are followed in the development of gross value added as well, therefore it is a great motivation to improve a new model for estimation.

The applied TVC model is a kind of state space models and can handle the nonlinearity in variables, although it uses linear equations. The fitted TVC model utilizes the physical indicators for estimation process and follows the structural changes of telecommunication. The predictive ability is appropriate, it allows more accurate estimation for the industry. The disadvantage of the method is that it cannot handle the changes in trend, although changes in trend causes problems in ARIMA models as well.

References

1. Brynjolfsson, E. & McAfee, A. (2014). The Second Machine Age, Work, Progress and Prosperity in Time of Brilliant Technologies. New York: W. W. Norton.
2. Cserháti, I. & Keresztély, T. & Takács, T. (2009): A GDP gyorsbecslés, Statisztikai Szemle http://www.ksh.hu/statszemle_archive/2009/2009_04/2009_04_345.pdf
3. Hall, S. G. & Swamy, P. A. V. B. & Tavlas, G. S.(2014): Time Varying Coefficient Models; A Proposal for selecting the Coefficient Driver Sets https://www.le.ac.uk/economics/research/RePEc/lec/leecon/dp14-18.pdf?uol_r=d307e306
4. Portfólió (2018a): Kábeleken múlik, lesz-e digitális robbanás Magyarországon https://www.portfolio.hu/vallalatok/telekom/kabeleken-mulik-lesz-e-digitalis-robbanasmagyarorszagon.279591.html
5. Portfólió (2018b): Létrehozták Magyarország első 5G-s kapcsolatát https://www.portfolio.hu/vallalatok/it/letrehoztak-magyarorszag-elso-5g-skapcsolatat.290630.html
6. Wollshlaeger, M. & Sauter, T. & Jasperneite, J. (2017): The Future of Industrial Communication, Automation Networks in the Era of the Internet of Things and Industry 4.0. IEEE Industrial Electronics Magazine, https://www.hs-owl.de/init/uploads/tx_initdb/IEEEMagazine.pdf, 2019. március, pp. 17-27

Employment of domestic concept in the framework of process table

Klára Anwar , Ildikó Ritzlné Kazimir
National Accounts Department, Hungarian Central Statistical Office, Budapest, Hungary

Abstract

The European Commission emphasised that labour input (Commission Decision, 94/168) could be one of the useful mechanism for assessing the exhaustiveness of gross domestic product (GDP) estimates. (OECD (2002)) From that time employment estimation in domestic concept (i.e. in line with the National Accounts) was developed in Hungary. The purpose of the development is to ensure the harmonization of production account and employment data. In this study we develop the methodology of non-observed employment estimation by categories of non-observed economy (NOE), prepared by Eurostat, (Eurostat (2005a)) and prepare the process table (PT) for the employment in domestic concept. This new approach, includes – similarly to the production of gross value added – data in institutional sector, Statistical Classification of Economic Activities in the European Community (NACE) and process table categories breakdown. Therefore it will provide a complete picture of the employment structure of the Hungarian economy. Furthermore it can facilitate the validation of gross value added from production side, and it allows detailed analysis.

Keywords

Non-observed economy; National Accounts; employment method

1. Introduction

The non-observed economy in Hungary spread considerably after the regime change, at the beginning of nineties. The transformation from socialism to market economy caused uncertain legal, institutional and economic environment, therefore the tax fraud and the illegal activities became more and more significant in the economic performance. The ratio of the non-observed activities reached 30% of GDP in 1992. (Árvay &Vértes (1994)) After the transformation period, the ratio of non-observed economy had decreased, and its share became 14.9% of GDP in 2005. (Murai &Ritzlné (2011)) However Medina and Schneider estimated 22.5% share of shadow economy from Hungarian GDP for the year 2005. (Medina &Schneider (2017))

The non-observed economic activities have to be estimated and included into the national accounts according to the European System of National Accounts 2010 (ESA 2010). Their estimation raises several problems. These activities include illegal or tax fraudulent activities, as well, while such activities of households result problems in the data collection process. Due to lack of appropriate available information, econometric models, statistical estimation processes are widely used in these estimations. However, precise delimitation of activities is also essential. For example the differences between the two above mentioned results for the year 2005 are due to the fact that not the same nonobserved economic categories were covered. In the official statistics the Tabular Approach to Exhaustiveness determines the activities included in the non-observed economy. (Eurostat (2005a)) Cross-validation of the results is very crucial in the estimation process of non-observed economy, because of their sensitivity and the economic actors' interest to hide the information. Obvious method for cross validation is the comparison of output and input side of non-observed economy. The labour input can validate the gross value added and give information about the productivity as well.

The basis of the labour input calculations is the labour force survey (LFS). In order to validate the gross value added data, first the LFS data have to be adjusted, because it represents the national concept. To be able to make comparison of different concepts possible, the labour input data have to be compiled in such a detailed breakdown as the gross value added.

This paper describes a framework that allows the comparison and validation of gross value added and employment data. To this purpose we apply the process table approach. In the first step, we identify the PT categories, i.e. adding labour input into production. In the second step the potential data sources are determined. Finally, we provide estimation of the non-observed employment for the average of years 2014-2016.

2. Methodology

In order to reach employment data in domestic concept, from that of national concept, like LFS several bridges should be taken into account according to ESA 2010 (Eurostat (2013)). This is illustrated in the right side of Table 1. Employed persons living abroad but working in Hungary, in a resident company should be part of the employment data of the country whereas employed persons living in Hungary but working abroad, i.e. for a non-resident company are not part of it, instead they are part of the employment of the country where the company they are working for is resident. The same is the case with the employees of the embassies. If Hungarian embassies, wherever they are, employ a foreign resident, it

should be part of the employment according to the domestic concept.

Moreover, employed persons living in institutions and those above the age 74 are not accounted in LFS therefore should be added to the number of employment. Due to the fact that the agricultural producers for own final use are not covered correctly in LFS due to statistical deficiencies, an additional estimation is provided as an adjustment. The same is the case with some other adjustments like drug trafficking, which is a sensitive field that persons do not like to confess publicly and probably do not consider as a job, to mention in such a survey.

The employment data in domestic concept can facilitate the detailed analysis of gross value added from productivity point of view and allows the validation. (Commission Decision, 94/168) The gross value added available by different dimensions, should be calculated for institutional sectors, in NACE breakdown, and it is the sum of data extracted from different data sources and result of several estimation. The breakdown of gross value added by data sources and estimation methods is classified into process table (PT) categories. The Eurostat has developed a unified table system to standardize and facilitate verifications. For the Member States, besides the Gross National Income (GNI) Inventory, a PT must be prepared in a defined structure for the specified reference year. (Eurostat (2004), (2005b)) The purpose of PT is to provide an overview of the compilation of the national accounts, the data sources and the estimation procedures used for the GDP-GNI calculation. The process table facilitates the verification of calculations and also improves the quality of national accounts. (Murai (2011)) The Hungarian GDP has been generated in PT structure for the whole annual time series since the calculation year 2013. This new system makes possible to follow up revisions among different calculation years.

The process table approach in the compilation of employment data according to domestic concept (see Table 1. left side) help to understand input needs of the production. It is necessary to cross-validate the gross value added and the items of income generation account. The income generation account, the gross value added and employment in domestic concept should be coherent and their items should be comparable to each other. Therefore, in the PT concept of production approach not all categories induce additional labour input.

The survey and census data, the administrative records as well as the combined data sources are the basis of national accounts data in the PT framework. These data sources cover all observed data for the reference period. In the case of employment in PT concept these lines include the social contribution return data, the monthly institutional labour statistics (regarding to short-term business statistics (STS)), the employment data

from personal income tax returns and corporation tax returns. The lines under the heading 'extrapolation and models' encompass all estimations, which based on computations or fixed percentage of survey information or administrative records. In Hungary for example the data of special purpose entities (SPE) is included in this item. The labour input of extrapolation and models should be validated by the basis of national accounts' figures, and should be re-estimated, if necessary.

Table 1.: Possible approaches to deduce employment in domestic concept

Process table approach			**Compilation from national concept**	
Basis for GVA Figures	Surveys & Censuses		**Basis for Employment in domestic concept**	Labour Force Survey (employment in national concept)
	Administrative Records			
	Combined Data			
	Extrapolation and Models	Benchmark extrapolations		Employed / Self-employed
		Commodity Flow Model		
		Consumption of fixed capital (Perpetual Inventory Method (PIM))		
		Dwellings - stratification method		
		Financial Intermediation Services Indirectly Measured (FISIM)	**Adjustments: Bridges from national to domestic concept**	Residents working outside the economic territory (-)
		Other extrapolations and models		
Adjustments	Data validation			Non residents working inside the economic territory (+)
	Conceptual	Allocation of FISIM		
		Other conceptual		Non residents employed at Hungarian embassies (+)
	Exhaustiveness	Producer should have registered (underground producer) N1		
		Illegal producer N2		Employed persons above the age of 74 (+)
		Producer not obliged to register		

		N3		
		Registered legal person not included in statistics N4		Employed persons living in institutional households (+)
		Registered entrepreneur not included in statistics N5		
		Misreporting by producer N6		Agricultural producers for own final use and other adjustment (+)
		Statistical deficiencies in the data N7		
	Balancing			
Employment in domestic concept			**Employment in domestic concept**	

Source: Own editing

The data validation adjustments remedy the uncovered source data problem. In this case, the difference between the initial and the adjusted data should be analysed, and the adjusted items should also be validated from all aspects. For example, if an enterprise reports incorrect sales data, other related data – including the employment should be also examined.

The conceptual items adjust the basic data to bring it in line with the ESA definitions. For example it includes net taxes on production to ensure the data at basic prices. (Eurostat (2005b)) These conceptual adjustments in gross value added do not induce additional employment in Hungary. The exhaustiveness adjustments (Murai, B., &Ritzlne Kazimir, I. (2011)) show the non-observed economy that may infer non-observed employment and even non-observed wages and salaries, while some non-observed subcategories, like value added tax (VAT) fraud without complicity (N6), wages and salaries in kind, gratuity, and tips (N7) do not induce additional non-observed employment. The mostly employment related non-observed category is the mentioned N7 (salaries in kind, gratuity, and tips) can be subtracted from survey, census, and administrative data, therefore it is included in the first category of the process table called 'survey, census and administrative records', as it is shown in Table 1. The non-observed economic categories are the following:

- N1 includes private accommodation of households, and rents for other real estates and unregistered educational activity of households.
- N2 category covers drugs, smuggling and prostitution.

- N3 category covers both agricultural own account production and costs of own account construction.
- N4 includes legal persons included in Register but not included in statistics.
- N5 covers agricultural production of small enterprises and sole proprietors included in register but not in statistics.
- Economic performance due to deliberate misreporting and VAT fraud without complicity belongs to N6 category.
- Sole proprietors not subject to VAT but submit Tax return, wages and salaries in kind, gratuity, tips and reimbursed costs included in N7 category.

The estimations for these items should be harmonized by the three mentioned approaches: production, income generation and employment method.

The balancing item in National Accounts process table reconciles the production, the expenditure and the income generation sides, while in the employment PT it reconciles the two approaches of compilation of employment in domestic concept: the PT approach and that derived from the national concept as Table 1. shows.

We have examined each exhaustiveness adjustment category whether covering additional nonobserved employment or not. It is illustrated in Table 2.

The investigation of these non-observed economic categories should be based on different data sources and from different concepts in order to facilitate the cross validation. For example, the N6 category of exhaustiveness adjustment from production side can be based on tax audit data, or theoretical VAT estimation (Eurostat (2005a)), while the N6 labour input should based on labour inspection data.

Table 2.: Possible non-observed employment by Exhaustiveness categories

Exhaustiveness adjustment category		**Type of data source for estimation**	**Improvement required in estimation method**	**Kind of activities, which may not result additional nonobserved employment**
N1	Producer should be registered (underground producer)	Surveys	yes	Private accommodation of households
N2	Illegal producer	Administrative data	no	
N3	Producer not obliged to be registered	Surveys, administrative data	yes	
N4	Registered legal person not included in statistics	Surveys, administrative data	no	

N5	Registered entrepreneur not included in statistics	Surveys, administrative data	no	
N6	Misreporting by producer	Employment and labour audit data	yes	VAT fraud without complicity
N7	Statistical deficiencies in the data	Administrative data	yes	Tip, gratuity, wages and salaries in kind, reimbursed costs

Source: Own editing

According to the results of our analysis of exhaustiveness categories, non-observed employment should appear in all categories of non-observed economy. In N1, N3, N6 and N7 categories new methodological improvement is required as a further step of our research.

3. Result

For the average of years 2014-2016, we compiled employment in domestic concept from data in national concept through the required bridges, as a first step. Results are shown in Table 3.

Table 3.: Compilation from employment in national concept, year 2014-2016 average

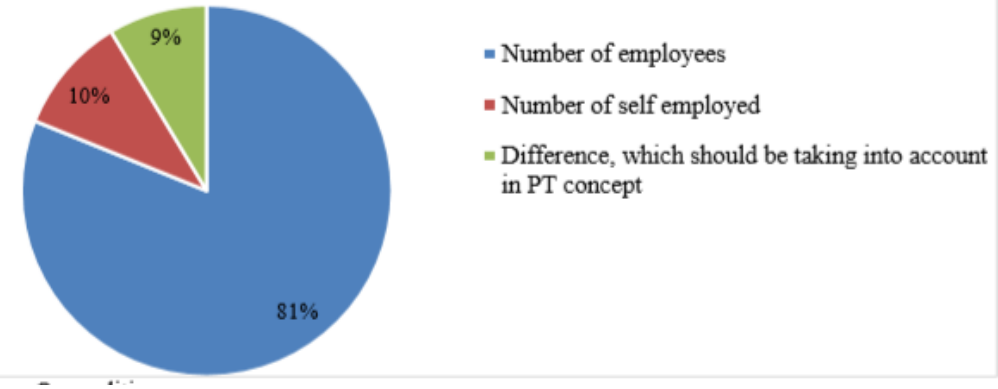
Labour Force Survey		**4221**
Adjustments, % of LFS	Residents working outside the economic territory	2.6
	Non residents working inside the economic territory	1.6
	Non residents employed at Hungarian embassies	0.01
	Employed persons above the age of 74	0.2
	Employed persons living in institutional households	0.6
	Agricultural producers for own final use and other adjustment	2.6
Employment of domestic concept		**4325**

Source: own calculation

According to these figures, 2.6% of the employment working abroad, therefore they are deducted from the total employment in domestic concept, while only 1.6% of non-residents are working inside the economic territory, increasing the number of employment. The same percentage that reduces the number of employment in one side due to residents working outside the economic territory, increases the number of employment on the other side due to the agricultural producers for own final use. The employment in domestic concept – the LFS data adjusted by ESA bridges – is 4325 thousand

persons in the average of years 2014-2016. Comparing this data to the result derived by process table approach of the employment is shown in Figure 1

Figure 1: The employment data in domestic concept according to PT approach, average of 2014-2016(thousand persons)



The number of employees – including the institutional labour statistics (employment of enterprises with more than five employees), the employment of small enterprises with less than five employees from combined data source[2], and the number of employment in household sector (employees of sole proprietors) from personal income tax returns – have 81% of total employment in domestic concept.

The number of self-employed persons, derived from information available in personal income tax returns and business register of Hungarian Central Statistical Office, gives 10% of total employment. As a result, 91% of total employment in domestic concept is covered by surveys and census as well as administrative records and combined data source.

The rest part of total employment according to process table approach comes from 'extrapolation and models', 'data validation' i.e. mostly non-observed employment as well as the 'balancing' item. These items provide 9% of the total employment.

4. Discussion and Conclusion

The paper argues that the process table concept of national accounts data compilation is a suitable tool to analyse employment data. Employment derived from national concept should be compared with the detailed employment data compiled in process table in order to explain differences, and cross validated national accounts data from gross value added, income generation and employment point of view.

The first results of our analysis show that 9% of the total employment data should be divided into 'extrapolation and models', 'exhaustiveness adjustment

[2] Combined data source is a harmonised and consistent database which includes statistical surveys data supplemented with corporate tax returns data at individual (enterprise) level.

data validation' and 'balancing items'. Considering the decreasing trend of the non-observed value added in Hungary since the middle of nineties, this result seems to be in line with the estimation of non-observed value added from output side. The further work in this field is to compile employment for the previously mentioned process table categories in the appropriate NACE and institutional sector breakdown.

References

1. Árvay, J. & Vértes, A. (1994): Share of the Private Sector and Hidden Economy in Hungary 19801992. Hungarian Statistical Review, 72 (7). (in Hungarian) pp. 517-529. http://www.ksh.hu/statszemle_archivum#year=1994/issue=07.
2. Eurostat (2004): GNI Process Tables Analysis PT/01. Luxembourg
3. Eurostat (2005a): Eurostat's Tabular Approach to Exhaustiveness Guidelines, Eurostat/C1/GNIC/050 EN, www.dst.dk/ext/739814884/0/...Tabular-Approach-part-1-2_ENG--pdf
4. Eurostat (2005b): Process Tables Compilation Guide GNIC/054. Luxembourg
5. Medina, L. & Schneider, F., 2017. Shadow Economies Around the World: What Did We Learn Over the Last 20 Years? https://www.imf.org/en/Publications/WP/Issues/2018/01/25/ShadowEconomies-Around-the-World-What-Did-We-Learn-Over-the-Last-20-Years-45583: IMF.
6. Murai, B. & Ritzlne Kazimir, I. (2011): Possibilities of measuring non-observed economy, Hungarian Statistical Review, 89 (5). (in Hungarian) pp 501-522
7. Murai, B. (2011): The framework for annual GDP calculations, Hungarian Statistical Review, 89
8. (6). (in Hungarian) pp 609-623 http://www.ksh.hu/statszemle_archive/2011/2011_06/2011_06_609.pdf
9. OECD (2002): Measuring the non-observed economy, A Handbook. Paris: OECD Publications Service
10. Eurostat (2013): European System of Accounts 2010, Luxembourg: Publications Office of the European Union, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-02-13269

Evaluating South Africa's market risk using APARCH model under heavy-tailed distributions

Retius Chifurira, Knowledge Chinhamu
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, South Africa

Abstract

Estimating Value-at-risk (VaR) of stock returns, especially from emerging economies, has recently attracted attention of both academics and risk managers. VaR and other risk management tools, such as expected shortfall (conditional VaR) are highly dependent on an appropriate set of underlying distributional assumptions. Thus, identifying a distribution that best captures all aspects of financial returns is of great interest to both academics and risk managers. This study compares the relative performance of the GARCH-type model combined with heavy-tailed distributions, namely; the Student-t distribution, Pearson type-IV distribution (PIVD), Generalized Pareto distribution (GPD) and stable distribution (SD) in estimating Value-at-Risk of FTSE/JSE all-share price index (ALSI) returns. Model adequacy is checked by using the Kupiec likelihood ratio test. The advantage of the proposed models lies in their ability to capture volatility clustering and the leverage effect on the returns, through the GARCH framework and at the same time model their heavy-tailed behaviour. The main findings indicate that the Asymmetric power ARCH (APARCH) model combined with heavy-tailed distributions performed well in modelling South African's market risk. Thus, APARCH model combined with heavy-tailed distributions provides a good alternative for modelling stock returns. The outcomes of this study are expected to be of salient value to financial analysts, portfolio managers, risk managers and financial market researchers, thus giving a better understanding of the South African market.

Keywords

Asymmetric volatility models; Value-at-Risk; Heavy-tailed distributions; FTSE/JES All share price index

1. Introduction

South Africa is one of the most diverse and promising emerging markets globally. It is the sixth most outstanding in the emerging economies category, with vast opportunities within its border. It is a gateway to the rest of the African continent (a market of more than one billion people) and is a key investment location. It is the economic powerhouse of Africa and forms part of BRICS group of countries which includes Brazil, Russia, India and China. South African stock market, Johannesburg Stock Exchange (JSE) is Africa's

largest stock exchange with more than 400 listed firms and offering a wide range of products. The South African stock market is significantly robust and is able to make the list of the first twenty largest stock markets in the world consistently (Hassan, 2013). This market value is unavoidably significant among world stock indexes, making it respond to the global economic meltdown surrounding emerging markets.

The FTSE/JSE All Share Price Index (ALSI) is designed to represent the performance of South African companies, providing investors with a comprehensive and complementary set of indices, which measure the performance of the major capital and industry segments of the South African stock market. It has 164 listed companies and it is about 99% of the full South African market capitalization value i.e. before the application of any investability weightings, of all ordinary securities listed on the main board of JSE, subject to minimum free-float and liquidity criteria. ALSI, as an equity index portrays the operational activities of a typical ordinary share in the South African market. The ALSI also evaluates the operationalization of the entire market (Makhwiting, 2014). The major volume of all securities listed on the JSE is an integral function of the market index because the share prices flow of the listed companies is what drives the market.

There are many types of empirical models which have been used to describe the stylized facts in stock returns. These include, ARCH (Engle, 1982), GARCH (Bollerslev, 1986), IGARCH (Engle and Bollerslev, 1986), EGARCH (Nelson, 1991), TARCH (Glosten et al., 1993a), APARCH (Ding et al., 1993), FIGARCH (Baillie et al., 1996), FIEGARCH (Bollerslev and Mikkelsen, 1996), FIAPARCH (Tse, 1998) and HYGARCH (Davidson, 2004). In order to obtain good estimates for risk management, the challenge is to choose the appropriate GARCH-type model which adequately captures volatility clustering and at the same type be able to capture the non-normality property of financial returns. Paolella (2016) used stable-APARCH model to model four stocks from DJIA index. Sin et al. (2017) used of the TGARCH combined with the generalized error distribution (GED) to model crude oil index. In the literature, there is no agreement of the type of the heavy-tailed distribution to be used in order to capture the non-normality of the residuals of the GARCH-type models. In this paper, we are interested in the relative performance of the asymmetric power auto-regressive conditional heteroscedastic (APARCH) model combined with heavy-tailed distributions, namely; generalized Pareto (GPD), Pearson type-IV (PIVD) and stable distributions (SD) in estimating the value-at-risk (VaR) for South Africa stock market.

We are not aware of any literature relating to an application of APARCH-GPD, APARCH- PIVD model and APARCH-SD model to the FTSE/JSE All Share Price Index. To the best of our knowledge, there are limited research on

combining dynamic volatility models with heavy-tailed distributions in modelling South African financial data. In this paper, we extend the work of Paolella (2016) by proposing an APARCH (1,1)-GPD model, APARCH (1,1)-PIVD model and compare it with the APARCH-SD model to the daily FTSE/JSE All Share Price returns. We estimate VaR and then select the more robust model using the Kupiec likelihood ratio test.

The rest of the paper is organised as follows. In section 2, we provide some background theory on APARCH(1,1), generalized Pareto distribution, Pearson type-IV distribution, stable distribution, VaR and backtesting. The data used in this study is described in Section 3. Section 4 presents the empirical results and discussions. Finally, Section 5 concludes this work.

2. Methodology

In this section, we present some background theory on the APARCH(1,1) model combined with generalized Pareto, Pearson type-IV and stable distributions. We also discuss VaR and backtesting procedures.

APARCH (1,1) model

Ding et al. (1993), introduced the APARCH model as an extension of the GARCH model. The APARCH generalized both the ARCH and GARCH models. The structure of the volatility equation is given by

$$\sigma_t^{\delta} = \omega + \sum_{i=1}^{p} \alpha_i (|X_{t-i}| + \gamma_i X_{t-i})^{\delta} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^{\delta} \qquad (1)$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $0 \leq \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \leq 1$. α_i and β_j are the ARCH and GARCH coefficients respectively and γ_i is the leverage coefficient. When γ_i is positive, it implies that the negative shocks has stronger impact on price volatility than the positive shocks. δ is a positive real number which functions as the symmetric power transformation of σ_t. Considering the case, where $\delta = 1$ for $p = q = 1$, then the volatility equation becomes:

$$\sigma_t = \omega + \alpha_1 (|Z_{t-1}| + \gamma_1 Z_{t-1}) + \beta_1 \sigma_{t-1} \qquad (2)$$

Generalized Pareto distribution

The two-parameter generalized Pareto distribution (GPD), with scale parameter β and shape parameter ξ, has the following distribution function

$$F_{GPD}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-\left(\frac{y}{\beta}\right)} & \text{if } \xi = 0 \end{cases} \qquad (3)$$

Where $y = x - \tau$ are the exceedances above the threshold τ and $y > 0$ when $\xi \geq 0$, $0 \leq y \leq -\beta/\xi$ when $\xi < 0$, and the scale parameter $\beta > 0$ (Tsay, 2013).

Threshold selection

In this paper, we utilize the mean excess plot and the parameter stability plot for threshold selections.

Pearson type-IV distribution

The generalized family of frequency curves, now known as the Pearsonian system of curves, was first developed by Karl Pearson (Cheng, 2011). The Pearsonian family includes members such as the normal, student-t, F, gamma, beta, inverse Gaussian, Pareto and Pearson type-IV distributions. The probability density function (pdf) of the Pearson's type-IV distribution (PIVD) is given by

$$f_{PIV}(x) = k\left[1+\left(\frac{x-\lambda}{a}\right)^2\right]^{-m} \times \exp\left[-\upsilon \tan^{-1}\left(\frac{x-\lambda}{a}\right)\right], \quad (4)$$

where $m > 1/2, \upsilon, a > 0, -\infty < x < \infty$, λ are real valued parameters and $k = \frac{2^{2m-2}|\Gamma(m-i\upsilon/2)|^2}{\pi a \Gamma(2m-1)}$ is a normalization constant that depends on m, υ and a. The pdf of the Pearson type-IV distribution is invariant under simultaneous change (a to $-a$, υ to $-\upsilon$).

Stable distribution

The stable distribution (SD) is a class of probability distributions described by four parameters namely: α an index of stability. Also referred to as the shape parameter in the literature with range $0 < \alpha \leq 2$, β the skewness parameter with range $-1 \leq \beta \leq 1$, $\gamma \geq 0$ the scale parameter, and $\delta \in \mathbb{R}$ a location parameter. These distributions are widely used in practice because they allow for skewness and heavy tails. Although many parametrization can be used to describe the characteristic function of a stable distribution, it does not have an analytical form in general. We follow the S_0-parametrization suggested by Nolan (2003) and say a random variable X follows a stable distribution if its characteristic function is given by

$$E(e^{iuX}) = \begin{cases} \exp\left(-\gamma^\alpha |u|^\alpha \left[1 + i\beta \tan\frac{\pi\alpha}{2} (\operatorname{sign} u)(|\gamma u|^{1-\alpha} - 1)\right] + i\delta u\right), & \alpha \neq 1 \\ \exp\left(-\gamma|u|\left[1 + i\beta\frac{2}{\pi}(\operatorname{sign} u)\log(\gamma|u|)\right] + i\delta u\right), & \alpha = 1 \end{cases} \quad (5)$$

The sign function used in equation (5) above is defined as

$$\operatorname{sign} u = \begin{cases} -1 & u < 0 \\ 0 & u = 0 \\ 1 & u > 0 \end{cases}$$

For the $\alpha = 1$ case, $x\log x$ at $x = 0$, is interpreted as $\lim_{x \downarrow 0} x\log x = 0$.

VaR and Backtesting

Value-at-Risk (VaR) has become a benchmark for evaluating market risks. This risk measure is used to assess the maximum possible loss for a portfolio over a specified time period (McNeil et al, 2005). There are two main approaches to calculating VaR for financial data. The parametric method and the non-parametric method (Brooks and Persand, 2000). In this paper, we estimate VaR using the proposed distributions (the parametric approach) and

compare them with the historical VaR values (a non-parametric approach).

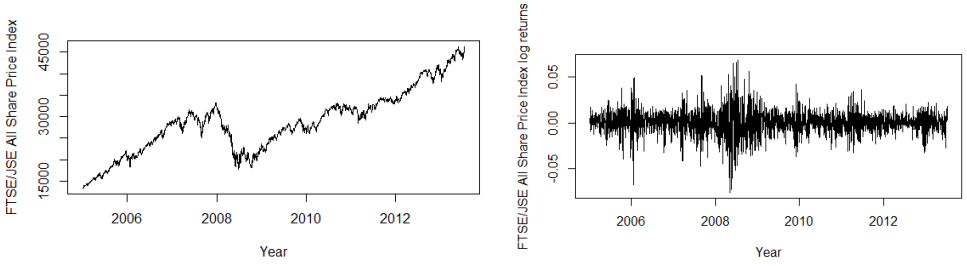
For a random variable X with distribution function F over a specified time period, the VaR (for a given probability p) can be defined as the p-th quantile of F, i.e., $\text{VaR}_p = F^{-1}(1-p)$, where F^{-1} is the quantile function (Tsay, 2013).

3. Result

In this paper, the data examined consist of the daily closing price of the all share index (ALSI) for the period 20 May 2005 to 31 May 2016 obtained from INET. We divide the data into in-sample dataset (20 May 2005 to 31 December 2013) and out-of-sample dataset (2 January 2014 to 31 May 2016). The in-sample data is used for the model estimation and for forecasting risk whilst the out-sample data is used for testing Value-at-risk (VaR) forecast. As a result, the estimation window has 2155 observations, the testing window has 602 observations, and thus the number of observations is 2757. Investors are interested in the return of their investment. We therefore, obtain the daily log returns (r_t) of the All Share Price Index. The log returns are given by

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

where r_t, is the natural logarithmic return of daily price of ALSI at time t, P_t is the daily closing price of ALSI at time t and P_{t-1} is the daily closing price of ALSI at time $t-1$. Figure 1(a) and 1(b) shows the time series and log returns plots of the in-sample data, respectively.



(a) (b)

Figure 1. Time series plot of (a) daily FTSE/JSE All Share Price Index (b) daily FTSE/JSE All Share Price Index log returns from 20 May 2005-31 December 2013 (in-sample data set).

The time series plot shows that the daily all share index has a trend and hence non-stationary in mean and variance. Using Figure 1(b) the log returns seem to be stationary but, the variance appears not to be constant over time indicating volatility clustering. In order to confirm the stationarity of the FTSE/JSE ALSI log returns, the augmented-Dickey-Fuller test is used to formally test for stationarity in mean and variance. The Augmented-Dickey Fuller statistic is -13.612 with p-value = 0.01< 0.05 thus, rejecting the null hypothesis at 5% significance level meaning that the log returns are stationary. The negative skewness is significantly different from zero and large excess

kurtosis clearly illustrates the non-normality (asymmetric property of the log returns) of the distribution. Since the p-value for Ljung-Box Q statistic is less than 0.05, we reject the null hypothesis of no presence of serial correlation.

The p-value for the ARCH Lagrange Multiple (ARCH LM) statistic is less than 0.05. Thus we reject the null hypothesis of absence of potential time varying volatility (no arch effect) up to lag 20. These findings led to the adoption of an asymmetric ARCH (APARCH) model as discussed in Section 2.

Asymmetric GARCH-type model fitting

In the first step, we fit the asymmetric GARCH type models to the returns and checks its adequacy since the returns have a significant skewness (asymmetric). The Table 2 shows the maximum likelihood parameter estimates and the standard errors in brackets of the asymmetric GARCH models with normal distribution innovation. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) model selection criteria are also reported in Table 1.

Table 1. ML Parameter estimates of asymmetric GARCH models

Parameter estimate	**EGARCH (1,1)**	**TGARCH (1,1)**	**APARCH (1,1)**
$\hat{\mu}$	0.0005 (0.0247)**	0.0005 (0.0195)**	0.0004 (0.0048)***
$\hat{\alpha}_0$	-0.1616 (0.0000)***	0.0000 (0.0897)*	0.0002 (0.0000)***
$\hat{\alpha}_1$	-0.1023 (0.0000)***	0.0105 (0.3905) *	0.0713 (0.0000)***
$\hat{\beta}_1$	0.9819 (0.0000)***	0.9057 (0.0000)***	0.9251 (0.0000)***
$\hat{\gamma}_1$	0.1369 (0.0000)***	0.1319 (0.0000)***	0.7932 (0.0000)***
δ	-	-	1.0000
AIC	-6.1498	-6.1459	-6.1533
BIC	-6.1367	-6.1327	-6.1402

Note: *, **, *** indicates p-value that is significant at 10%, 5%, and 1% level of significant respectively.

From Table 1, it is observed that the ML parameters estimates for the three asymmetric GARCH models fitted to the FTSE/JSE ALSI log returns are significant at least at 10% level of significance. The APARCH (1,1) model has the least AIC and BIC values and is selected as the best asymmetric GARCH-type model. The APARCH (1,1) model has successively captured the volatility clustering with Ljung-Box p-value$= 0.3367 > 0.05$ and ARCH-LM p-value$= 0.3743 > 0.05$ of the extracted standardized residuals. The model is found to be able to capture the asymmetry of the returns with p-value$= 0.2862 > 0.05$ of the sign bias statistic. To check for the non-normality of the standardized residuals, the Shapiro-Wilk test is employed. The standardized residuals are non-normal and this is confirmed by the Shapiro-Wilk test statistics with p-value $= 0.0000 > 0.05$. Thus, justifying using heavy-tailed distributions to model the extracted standardized residuals from APARCH model. In this study,

we fit the GPD, Pearson type-IV and stable distributions to the standardized residuals from the APARCH model.

Combining APARCH (1,1) with heavy-tailed distributions

APARCH (1,1)-GPD model

To fit the GPD model, we check whether the tail of the standardized residuals follow a Pareto distribution. It can be shown using the Pareto quantile plot that the tail of the data is almost a straight line confirming that the standardized residuals may follow a generalized Pareto distribution. The mean excess and the parameter stability plots were used to come up with a reasonable high threshold τ. The suitable threshold must lie where there is a positive change in the mean excess. We use the parameter stability plot to check the threshold where the parameters are most stable. The estimated parameters are more stable when $\tau \geq 1.7$. There are 635 observations above the threshold. Since the exceedances above the threshold cannot be assumed to be independent and identically distributed, declustering was performed.

We fit GPD (model 1) to the declustered exceedances. The maximum likelihood (ML) parameter estimates with standard errors in brackets are obtained as $\hat{\xi}$ = -0.0647 (0.1254) and $\hat{v}$ = 0.3793 (0.0701) with 635 observations in the tail. The probability plot and the quantile plot suggest that the exceedances seems to follow the GPD model. Thus, all the diagnostic plots suggest that the exceedances follows the GPD model.

APARCH (1,1)-PIVD model

The PIVD is fitted to the standardized residuals extracted from the APARCH (1,1) model with normal innovations. The parameters are estimated using the method of maximum likelihood. The maximum likelihood procedure is carried out using R package PearsonDS. The ML estimates of the Pearson type-IV distribution fitted to the standardized residuals of the APARCH(1,1) model with normal innovations are $\hat{m} = 12.66$, $\hat{v} = 11.7361$, $\hat{\lambda} = -2.2198$ and $\hat{a} = 4.2221$. The calculated AD statistic value is 0.2868 with a corresponding p-value of 0.9477. The the value of $\hat{m} = 12.6666 > 0.5$, thus satisfying the condition for a PIVD. The AD statistic is significant, thus the PIVD is a good fit of the standardized residuals extracted from the APARCH(1,1) model.

APARCH (1,1)- SD model

The stable distribution (SD) is also fitted to the extracted standardized residuals of the APARCH (1,1) model. The model is referred to as APARCH (1,1)-SD model. The ML parameter estimates of a stable distribution fitted to the standardized residuals of APARCH (1,1) model are $\hat{\alpha} = 1.9163, \hat{\beta} = -1.0000, \hat{\gamma} = 0.6784$ and $\hat{\delta} = 0.0778$. The calculated AD statistic value is 1.0652 with a corresponding p-value of 0.3248. The value of the index of stability ($\hat{\alpha}$) is 1.9163 which is less than 2. This suggested that the tail of the standardized residuals follows a Pareto law indicating the distribution is heavy-tailed and also has infinite variance. The stable skewedness ($\hat{\beta}$) is -1,

suggesting that the standardized residuals are skewed to the left. The AD statistics has a p-value $= 0.3248 > 0.05$, confirming that the stable distribution is a good fit for the standardized residuals.

VaR is calculated for each model and the models are backtested using Kupiec test. The p-values of the Kupiec test for both the in-sample dataset and out-of-sample dataset are summarized in Table 2.

Table 2. VaR backtesting for FTSE/JSE ALSI returns

	In Sample dataset			***Out Sample dataset***		
	Sample size = 2155			***Sample size = 602***		
	***p*-value of Kupiec test**			***p*-value of Kupiec test**		
Distr	**97.5%**	**95%**	**90%**	**97.5%**	**95%**	**90%**
Student *t*	0.0033	0.0003	0.0207	0.0436	0.0078	0.2551
GPD	**0.9862**	**0.5664**	0.2306	**0.4094**	**0.3265**	0.3921
PIVD	0.6890	0.4384	**0.3278**	0.0882	0.1667	**0.5642**
SD	0.1201	0.0000	0.0137	0.0882	0.0448	0.2551

The VaR estimates from the APARCH(1,1)-Student-t distribution produced the lowest p-value < 0.05 for the Kupiec likelihood ratio test statistic at almost all VaR levels. The best model for VaR estimation for the FTSE/JSE All share index returns differ at different VaR levels. We observe that at 97.5% and 95% levels, the APARCH(1,1) with GPD innovations produced the highest and significant p-values. This indicates that at these levels, the best VaR model is APARCH (1,1)-GPD model. While at 90% VaR level, the APARCH(1,1) model with PIVD innovations produced the highest p-value. We also observe the same performance of the models in the out-of-sample dataset. The APARCH (1,1)-SD model failed to adequately estimate VaR at 90% and 95% levels with p-value < 0.05 . In general, we conclude that the GPD and PIVD favourably capture the extreme risk in FTSE/JSE all share index returns.

4. Discussion and Conclusion

In this article, we examined the suitability of using APARCH (1, 1) framework combined with heavy-tailed distributions for modelling VaR for FTSE/JSE all share index returns. The APARCH framework was used to capture volatility and asymmetric characteristics exhibited by financial returns, while the heavy-tailed distributions are used to capture the heavy-tailed-ness of actual return distributions. The GPD, Pearson type-IV and the stable distributions are applied to the i. i. d. standardized residuals from the APARCH (1,1) model with normal innovations and VaR is calculated at different levels. Adequacy of the resulting VaR estimates were tested using the Kupiec likelihood ratio test. Backtesting using the Kupiec LR test has shown that the APARCH(1,1) with GPD governing the innovations is the most robust models at 97.5% and 95% level. At 90% level the APARCH(1,1) with PIVD governing

the innovations is the most robust model for estimating VaR for FTSE/JSE all share index returns. The backtesting procedure emphasized the superiority of the GPD and PIVD models over Student-*t* and stable models, thus providing a very good candidate as an alternative distributional scheme.

References

1. Cheng, R., 2011. Using Pearson type IV and other cinderella distributions in simulation. *Proceedings of the 2011 Winter Simulation Conference, University of Southampton*, 457-468.
2. Hassan, S., 2013. South African Capital Markets: An Overview. *In:* BANK, S. A. R. (ed.). Pretoria.
3. Makhwiting, M. R., 2014. *Modelling volatility and financial market risks of shares on the Johannesburg stock exchange.* Master of Science, University of Limpopo.
4. Paolella, M, S., (2016). Stable-GARCH models for financial returns: Fast estimation and tests for stability. *Econometrics*, 4(25), 1-28.
5. Nolan, J. P., 2003. *Modeling financial data with stable distributions*. *In:* RACHEV, S. T. (ed.) *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance.* Netherlands: Elsevier Science.
6. Tsay, R. S., 2013. *Analysis of Financial Time Series*, 3rd edition. New Jersey: Wiley & Sons.

Development of agricultural and rural statistics in the CIS Region

Andrey Kosarev
Commonwealth of Independent States (CIS)

Abstract

The agricultural production is a key supplier of food for consumer markets and for food security in many countries of the Commonwealth of Independent States (CIS). CIS-Stat, as coordinator of statistical activities in the CIS countries, pays much attention to the development of statistics in the framework of the implementation of the "Global Strategy to Improve Agricultural and Rural Statistics" in the CIS region. The implementation of the World Bank project "Development of Agricultural and Rural Statistics in the CIS Region" contributed a lot to the development of methodology and improving the statistical practices. Important methodological issues like agriculture census, sample approach in agriculture statistics, and others were developed basing on the international methodology standards.

Keywords

FAO; World Bank; methodology standards; agricultural census; sample surveys

1. Introduction

Work on the "Global Strategy to Improve Agricultural and Rural Statistics" (see [1]) started after the 2007 International Statistical Institute Conference on Agricultural Statistics. The agricultural production is a key supplier of food for consumer markets and for food security in many countries of the Commonwealth of Independent States (CIS). Thus participation of the CIS region in the implementation of the "Global Strategy" is one of priority activities for the CIS-Stat which is responsible for coordination of statistical activities in the CIS countries.

The World Bank project "Development of Agricultural and Rural Statistics in the CIS Region" (see [2]) played a crucial role in ensuring most efficient solutions supporting statistical development in the CIS countries. The development objective of the WB project is to create the framework for national and regional statistical systems which will allow the collection and use of statistical data that are needed in making decisions regarding the development of agriculture in the 21st century. Three pillars were indicated as specific objectives of the project:

- integration of agriculture into the national statistical system in order to meet the users' demand for consistent and compatible data by territories and over time;
- improvement of the sustainability of the agricultural statistical system through governance and statistical capacity building;
- identifying a minimum set of core data which will be collected by the countries to meet both current and emerging information requirements.

2. Methodology

The WB project enabled productive implementation of different kinds of activities, the key ones among them being wide expert discussions and development of specific methodological papers. Several CIS regional expert meetings were conducted during the period 2012-2018 in different CIS countries as well as in FAO Regional Office for Europe and Central Asia (Budapest). "Plan for the implementation of the Global Strategy in the CIS region" was elaborated at the expert meeting in 2012 and then approved by the highest regional statistical body – Council of Heads of CIS NSOs – in 2012. The discussions held during the following expert meetings were focused on most important issues related to the development of agricultural and rural statistics in the CIS region.

Following the CIS experts' priorities, the CIS-Stat developed several methodology papers discussing the key issues, adapting international methodology standards to specific features of CIS countries, analyzing the national CIS practices, and promoting implementation of most efficient solutions into the regional agricultural and rural statistics (see [3]…[10]). Proposals on the system and methodology of agro-environmental indicators for the CIS region are being developed currently in accordance with the international standards.

A special analytical **overview of agricultural statistics based on generalized practice of international standards** used all over the world was carried out by the CIS-Stat as a starting point in providing recommendations for CIS NSOs in the field of agricultural statistics (see [3]). The paper reviews the general issues of the organization of agricultural statistics in the EU countries (subjects, classifications, standards) in such areas as structural changes, land use, crop production, livestock, agri-environmental indicators, including organic agriculture, as well as FAOs contributions to the improvement of agricultural statistics.

An important task in developing agricultural statistics in the CIS region is ensuring methodological comparability of data produced by the NSOs. In this context the CIS-Stat carried out a special analytical work allowing reconciling different national data – **conversion keys for the main indicators of agricultural and rural statistics** (see [4]).

The objective of this work is to harmonize the methodological aspects of determining a minimum set of core data and indicators stated in the "Global strategy to improve agricultural and rural statistics" in the CIS countries in order to establish information resources in compliance with international standards and recommendations. The paper presents a comparative analysis of the methodologies applied by NSOs for collecting data to produce agricultural and rural statistics. A description is given of the capacity of the information resources in CIS countries to form a minimum set of core data and a set of indicators in the area of agricultural and rural statistics; including indicators for such industries as agriculture, forestry, and fishery and aquaculture, and also indicators addressing the issues of climate change, availability of land resources, agricultural impacts on the environment, and the economics of rural areas, which are stated in the Global Strategy. Proposals were made for the approaches to improve the methodology of statistical observation in the CIS countries in accordance with the international standards and recommendations. Also, proposals were developed to harmonize the methodologies for establishing a minimum set of core data and a set of indicators and adjustment procedures (conversion keys) to obtain comparable data for CIS countries on the indicators used to describe the conditions and development of agriculture and rural areas, as well as proposals for obtaining derived indicators to enable international comparisons.

The development of agricultural statistics in the CIS region is based on a balance approach, being considered a part of the general statistical system. In this context the CIS-Stat paid special attention to consistency between agricultural statistics and other branches of statistics, especially – national accounts. Methodological recommendations were prepared to assess **output and intermediate consumption for the agriculture production account** (see [7]).

The paper contains a broad overview of the sources of primary data necessary for estimating the output of agriculture, examines their features, relative advantages and limitations, and procedures for data processing. One of the tools for calculating agricultural output may be balance sheet of production and use of agricultural products; its scheme is described in the paper. These balance sheets enable balancing and harmonizing information on production and use of agricultural produce; they may be used as technical tools for calculating output of some groups of agricultural products. In a separate section the use of balance sheets for estimating output of livestock breeding is described. The use of balance sheets is particularly important for assessing production of enterprises in the informal sector, as data on their activities are collected mainly through sample surveys. Use of 2008 SNA concepts for describing and analyzing agricultural production and other aspects of economic activities in the agricultural sector will improve the quality

of the main macroeconomic aggregates and facilitate solving information problems in carrying out international comparison of GDP.

One of key topics in the agricultural statistics is the methodology for the **census of agriculture**. The CIS-Stat developed special works on this topic (see [9]).

At present in many CIS countries Censuses of agriculture have been conducted. However, they were organized in different years and they differ substantially in terms of observed variables. The objective of the paper document was to adapt the World Program for the Census of Agriculture 2020 (WCA-2020) developed under the leadership of FAO to the practices of agricultural statistics in CIS countries and to propose a unified approach to the programs for the agricultural censuses in the CIS. The objective of the work was to improve the quality and coverage of statistics compiled by the national statistical offices of CIS countries which describe the state and development of agricultural production, in terms of the source for the minimum set of core data and establishing a universal survey frame.

For the sake of better international comparability of census data, CIS-Stat emphasizes the importance of synchronizing the preparation and conduct of the agricultural censuses in CIS countries and harmonizing the observation variables to be included in the questionnaires.

A new Theme 15 included in WCA-2020: *Environment/greenhouse gas (GHG) emissions* requires understanding the substance of the calculations to estimate greenhouse gas emissions. The set of variables included in the program for the agricultural census may help countries to assess greenhouse gas emissions from different sources.

Variables and indicators for the census of agriculture in CIS countries were chosen with regard to the national practices and peculiarities of agricultural production in CIS countries; they comprise all core data and some additional variables.

If the census of agriculture that is being conducted for the first time, it is recommended to establish a survey frame on the basis of the agricultural statistical register (for legal units), local administrative data, and population census data (for physical persons). For a repeated census the updated survey frame of the previous census may be used.

The paper describes in detail the ways of conducting censuses of agriculture, their main characteristics, advantages and disadvantages, as well as requirements for their use; methods for estimating the total land area are also described.

The paper is focused also at the organization of monitoring in the course of an agricultural census, including methods of land surface remote sensing, and the compilation of aggregated tables on the basis of primary data from census questionnaires. After the census, it is necessary to introduce

adjustments into the time series; approaches to their recalculation are described in the paper.

One of the key tasks of the 2011 Busan Action Plan for Statistics is to improve coordination and cooperation between data producers and data users and to enable open access to statistical information. In this regard, it is recommended before the start of the census to develop a standard plan, methods and data dissemination venues, including access to anonymized microdata, and archiving formats for the agricultural census data.

Sample questionnaires for different types of agricultural producers and the layouts of summary tables are presented in the paper to assist in developing survey tools for agricultural censuses in CIS countries. A glossary of terms is prepared to assist census enumerators and statisticians in CIS countries.

The implementation of **sample approach** is one of key aspect in developing statistics in the CIS countries, especially in agriculture statistics. The CIS-Stat developed special methodological materials on this subject matter (see [6]).

The objective of this work is to improve the methodology of sample surveys of agriculture in the CIS countries and to propose such sampling methods that ensure maximum possible coverage of all types of agricultural producers with the account for their size, importance, location, and/or other characteristics and enable proper grossing up the sample survey estimates to the general population. It will allow fulfilling one of the main tasks of agricultural data collection as stated in the Global Strategy to Improve Agricultural and Rural Statistics.

The paper analyzes the international recommendations for the preparation and conduct of sample surveys in agriculture.

On the basis of the conducted research, approaches have been proposed to constructing a sample of small business entities, peasant (farmer) farms and personal subsidiary farms engaged in agricultural activities with the purpose to refine the existing methodologies for conducting sample surveys in agriculture in the CIS countries. The practical implementation of the proposed recommendations will allow exercising a more relevant approach to constructing territorial samples, taking into account international recommendations on this issue. This will contribute to improving the representativeness of data collected through sample surveys of agricultural activities and the comparability of data between the CIS countries and at the global level within the framework of agricultural statistics.

3. Result

At present, the main outcome of the work on the implementation of the Global Strategy in the CIS region is the development of methodological materials and an increased exchange of practical experience between the countries of the region through expert meetings. The effectiveness of this activity has been repeatedly noted by experts from CIS countries during the meetings. This work is also regularly reviewed and supported by the Council of Heads of CIS NSOs.

4. Discussion and Conclusion

At present, only intermediate results of the work on improving agricultural and rural statistics in the CIS countries have been obtained. CIS-Stat expresses its appreciation to the World Bank and FAO for the productive support of this work and plans to continue making efforts in this area.

References

1. Global Strategy to Improve Agricultural and Rural Statistics – FAO, 2010, http://www.fao.org/docrep/015/am082e/am082e00.pdf
2. Development of Agricultural and Rural Statistics in the CIS Region – World Bank, 2014, http://projects.worldbank.org/P150037?lang=en
3. Recommendations for CIS NSOs in the field of agricultural statistics based on generalized practice of international standards used all over the world – CIS-Stat, 2015, http://www.cisstat.com/gsagr/CIS_Agristat_Metodology_Recommendation_NSS_Generalized_Practice_of_Application_of%20International%20Standards.pdf
5. Development of conversion keys for the main indicators of agricultural and rural statistics in the CIS region in the framework of the "Global Strategy to Improve Agricultural and Rural Statistics" in accordance with the FAO methodology – CIS-Stat, 2015, http://www.cisstat.com/gsagr/
6. Methodology recommendations for computing agricultural producer price indices in case of seasonal production and utilization (processing) – CIS-Stat, 2016, http://www.cisstat.org/rus/CIS-PriceStat/CIS_PriceStat_02%2003%20Recommendations%20for%20seasonal%20adjustment%20of%20prices.pdf
7. Development of the methodology of sample statistical survey of agricultural activities of small businesses, farms and household plots – CIS-Stat, 2017, http://www.cisstat.com/gsagr/
8. Methodology recommendations to assess output and intermediate consumption for the agriculture production account – CIS-Stat, 2017, http://www.cisstat.com/gsagr/CIS_Agristat_Metodology_Recommendation_for_output_for_agriculture.pdf

9. Methodology recommendations to calculate production index for agricultural products – CIS-Stat, 2017, http://www.cisstat.com/gsagr/CIS_Agristat_Metodology_Recommendation_on_calculation%20of%20the%20industry%20of%20agricultural%20production.pdf
10. Methodology recommendations for improving the agricultural census program for the CIS countries – CIS-Stat, 2018, http://www.cisstat.com/gsagr/CIS_Agristat_Metodology_Recommendation_for_Improvement_Agricultural_Census_for_CIS.pdf
11. Methodology recommendations on the system of indicators for assessing the food security in the CIS countries – CIS-Stat, 2018, http://www.cisstat.com/gsagr/

Saving, borrowing and economic resourcefulness in Poland

Marek Kośny
Wroclaw University of Economics, Wroclaw, Poland

Abstract

Economic resourcefulness, as one of the important determinants of the ability to maintain economic security, indirectly affects many aspects related to the functioning of individuals and their households on the market. One of the important areas is making financial decisions related to the allocation of financial surpluses (saving) and taking loans. The analysis carried out on the basis of a dedicated study aims to identify various aspects of the processes of collecting savings and borrowing in Poland, and to assess to what extent these processes are conditioned by economic resourcefulness. The obtained results indicate the significance of economic resourcefulness for the majority of identified aspects of saving and borrowing, but at the same time, they show the specificity of these aspects. Resourcefulness, oriented at gaining a good financial situation and economic stability does not necessarily lead to strengthening individual retirement saving, what could be interpreted as a kind of trade-off between strategies build upon focusing on achieving well-being in the short (medium) and a long time horizon.

Keywords

Saving behaviour; Saving motives; Borrowing behaviour; Economic resourcefulness

1. Introduction

The management of financial resources includes a number of specific areas: in particular, the forms and means of obtaining funds, and sources of financing the expenditures, the adopted perspective (short- and long-term) and stability of income and expenditure over time. In the context of the analysis which will be presented later in the article, two issues are of particular importance. The first of them is saving behaviour - both in short and long term. The second is the behaviour in the area of borrowing. A similar characterization of these areas suggests the possibility of the occurrence of common determinants: identifying a set of features that influence behaviour in both areas will mean not only their formal similarity but also common behavioural determinants.

The literature review shows that both the concepts of saving and borrowing cover, however, a wide range of behaviours (see, for example,

investment and consumption debt). These differences result not only from the motivation of decision-makers (the subjective factor) but also from the economic meaning (objective factor). This internal diversity raises a number of basic questions. Firstly, is the formal differentiation reflected in the attitudes and behaviour of individuals – i.e. is the propensity to save (debt) depending on the motive, purpose, and manner of saving (borrowing)? Are people who save on a regular basis (collecting funds for the planned purchase of assets – the improvement and down-payment motive or for unforeseen expenses – the precautionary motive), will also be those who save for retirement purposes (the life-cycle motive or the bequest motive)? Secondly, what determines specific behaviour in terms of saving (borrowing) and are these the same factors, independent of the way (motive) of saving (borrowing)?

With regard to these questions, the purpose of the article is to identify various aspects of saving and borrowing and to indicate their determinants. The basic aspects characterizing the behaviour of households in the area of saving and borrowing were identified based on the results of the factor analysis for a dedicated study conducted on a group of Polish households. It allows for assigning households' behaviours to wider areas, which allows searching for common determinants for specific types of behaviour. In the second stage of the analysis, an attempt was made to assess to what extent the identified areas are conditioned by economic resourcefulness – an important factor in guaranteeing the economic security of households.

The article contributes to the literature by application of the recently introduced category of economic resourcefulness to the description of saving and borrowing behaviour in Poland. Obtained results show new aspects of short- and long term saving in Poland and suggest the existence of a kind of trade-off between concentration on the short (medium) and the long-term situation in the context of maintaining the desired standard of living.

2. Methodology

The empirical analysis was carried out on the basis of the data from a dedicated questionnaire survey which was intended to analyze various aspects of the functioning of individuals and households on the market. Within the survey respondents answered a number of questions regarding their market situation, saving, indebtedness, the situation on the labour market as well as many personal characteristics (noncognitive factors).

The study was conducted in Poland at the end of 2016 by a renowned research agency, using CAPI method. Sampling was based on the TERYT system, used for representative surveys by the Central Statistical Office in Poland. The group of respondents covered three cohorts aged 25-31, 32-38, 39-45. The age range of the cohort was established in relation to the political transformation that took place in Poland in the early 1990s of the previous

centuries. In order to obtain a homogeneous sample with relatively consistent experiences, all respondents aged 25-45 were required to have higher education. In total, 902 respondents participated in the study.

The analysis of such a defined data set was carried out in two stages. In the first step, data related to various aspects of saving and borrowing (separately for each area) were subjected to exploratory factor analysis (EFA), with principal component analysis and varimax rotation. The extracted factors were supposed to have eigenvalues higher than one. Additionally, the resulting set of factors had to explain a minimum of 75% of the variance of the original set of variables (factors explaining less than 10% of the total variance were not taken into account). The primary purpose of factor analysis was to identify latent constructs and assessment of dimensionality in the set of analyzed questions. In both analyzed areas – saving and borrowing – the number of concealed dimensions obtained was 3. The dimensions identified in this way meet the conditions set by Briggs and Cheek (1986) – they are both conceptually meaningful and empirically useful. In the second step, the emergent constructs were utilized in the regression analysis as independent variables what allowed to indicate key determinants of the aspects of saving and borrowing identified in this way.

The basic variable whose impact on saving and borrowing was verified is the respondent's economic resourcefulness, for which the structural equation model (MIMIC) proposed in Kośny and Piotrowska 2018 (one-generational model) was used. The theoretical values of the hidden variable – economic resourcefulness – were estimated on the basis of the estimates given there. Then the imputed values obtained in this way were used as an explanatory variable in the regression models estimated for particular aspects of saving and borrowing.

All analyzes and calculations presented in this article were carried out in IBM SPSS Statistics 24.

3. Result

The starting point for the analysis is the exploratory factor analysis (EFA), the results of which are presented in Table 1. For the principal component analysis, answers to 7 questions were used, which in the study concerned the issue of saving. Taking into account the content of the questions on the basis of which the analysis was carried out, emerging factors can be described in the following way.

Factor 1 can be characterized as the Ability to accumulate savings, because it shows the specific effects of the saving process, which is the savings accumulated – both pension savings, as well as those which, in the respondent's opinion, are not intended for this purpose. This distinction is significant because the question left the respondent free to define the nature

of savings, which in a broader context is a reference to the concept of mental accounts (see Shefrin and Thaler 1988).

Factor 2 can be described as Pension saving. It includes, of course, pension savings – which is a component shared with Factor 1. In addition, it refers to planning for saving for a pension. An interesting aspect of this factor is that it is not related to having savings, which – according to the respondent – are not directly related to pensions. This indicates the specific character of pension savings and another mechanism of their formation.

Factor 3 covers aspects related to the current saving for a specific purpose and the ability to keep the desired level of expenses, which is why it was described as Saving discipline. It is interesting, however, that the behaviours from this area form a separate factor, distinct from the two previously mentioned. An attempt to interpret it may be based on a reference to the saving motives described in Section 2. While these motives are directly referred to by Factor 2 (the life-cycle motive), Factor 3 seems to reflect the down-payment and improvement motives, nevertheless, references to other motives can also be found. However, of the most general character is Factor 1, which corresponds not only to the precautionary motive, the inter-temporal substitution motive but also to others, including the already mentioned life-cycle motive.

Table 1. Savings – factor's loading for the first three factors

Questions	**Factor's loadings**		
	Factor 1	**Factor 2**	**Factor 3**
Please specify how often in the last 6 months you have saved for significant expenses like a car, home or educational expenses?	0.076	0.154	**0.813**
Please specify how often in the last 6 months you managed to fit in your budget (expenditure plan)?	0.161	-0.077	**0.804**
Do you save for a pension in any form, in addition to compulsory contributions?	**0.445**	**0.562**	0.084
Do you have a plan on how to save for a pension?	0.034	**0.877**	0.048
Have you ever tried to estimate how much you would have to save in the period when you work to make you satisfied with your standard of living in retirement?	-0.073	**0.852**	-0.003
Does your household have savings?	**0.944**	0.037	0.126
What is the approximate value of your household savings?	**0.941**	0.010	0.148
Eigenvalue	2.419	1.749	1.063

Characteristic of identified aspects is an incentive for attempts to look for determinants. Such analysis was carried out using a regression model in which as independent variables were utilized the constructs that emerged from EFA. As the potential explanatory variables, the age (in years) and the level of income were adopted – according to a previously presented literature review.

As the key determinant, whose impact on particular aspects of saving was examined, we chose economic resourcefulness. Weighted OLS estimates are presented in Table 2.

Table 2. Determinants of factors identified for saving behavior

	Factor 1	**Factor 2**	**Factor 3**
Economic resourcefulness (variable imputed on the basis of SEM)	**0.598*** **	**-0.041**	**0.308*** **
Total monthly net income of the household (in PLN)	-0.005	-0.031	-0.041*
Age (in years)	**0.011** **	**0.030*** **	**0.004**
Constant	-3.183***	-0.700***	-1.392***
Adjusted R^2	0.315	0.030	0.080

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Estimates show a positive relationship between the age variable and the analyzed factors, the relationship being statistically significant for Factors 1 and 2. In the case of the income variable, the results were reversed – the coefficients were negative but significantly different from zero only for Factor 3. Economic resourcefulness has a positive effect on Factors 1 and 3.

These results mean that resourcefulness is an important determinant of the Ability to accumulate savings – both in the context of its effects (Factor 1) and internal discipline (Factor 3) – even though the analysis showed that these aspects are separate. At the same time, the Ability to accumulate savings increases with age and does not depend on the level of income, which is a direct reference to the Permanent Income Hypothesis and the Life Cycle Hypothesis (taking into account the age range of respondents). On the other hand, the discipline in current saving is a factor independent of age but conditioned by features characterizing the resourceful individuals. As the results show, this aspect of saving is particularly important for people with lower incomes, whose resourcefulness is manifested in the fact that they try – by saving – to limit a debt incurred for consumption purposes.

The positive relationship between Factor 2 and respondent age is completely natural and fully compatible with the Permanent Income Hypothesis and the Life Cycle Hypothesis. The result obtained for the variable economic resourcefulness is interesting, however. The lack of statistical significance for this explanatory variable means that pension saving is managed in a different way compared to other saving areas – to the extent to which it involves conscious planning of such savings. It is an important indication that in order to increase private savings in this area, specific incentives are needed, oriented at this type of savings, related in particular to the promotion of long-term savings planning, which is not directly related to the level of economic resourcefulness. This lack of relationship may – at least

partially – result from the specific situation of Poland, discussed at the beginning of this section. Relatively low income and the need to build the wealth without significant support of the generation of parents, causing concentration on the current situation, create at the same time the belief that there is no real opportunity to collect savings of significant value that would actually increase income during retirement. This result contradicts the predictions of Life Cycle Hypothesis, in which saving for retirement is the basic motive for saving, but complies – especially taking into account the age of respondents – with Carroll's buffer-stock model of savings (Carroll 1997).

The second of the characterized areas of behavior relates to borrowing. Due to the close relationship of saving and borrowing, the analysis of this area will be carried out according to the same scheme that was applied to saving processes. The results of exploratory factor analysis (carried out using the principal component method), which enable the identification of related behaviors, are presented in Table 3.

Table 3. Borrowing – factor's loading for the first three factors

Questions	**Factor's loadings**		
	Factor 1	**Factor 2**	**Factor 3**
Do you currently have loans or credits to pay off?	**0.930**	0.164	0.033
Do you currently have the mortgage to pay off?	**0.826**	0.000	0.006
How did you manage to cover expenses – did you get loans or credits?	0.067	0.019	**0.828**
How did you manage to cover expenses – did you not pay part of bills, loan installments?	-0.011	0.023	**0.835**
Please, assess to what extent the repayment of debts is burdensome for your household.	**0.944**	0.147	0.047
Do you have credit cards?	0.118	**0.912**	0.006
How often in the last 6 months you have used the entire credit limit on at least one credit card?	0.101	**0.915**	0.042
Eigenvalue	2.729	1.479	1.364

Factor 1 characterizes a "conventional" borrowing – in the form of loans or credits, including mortgage loans. A very interesting aspect is that only this component is related to the nuisance of repayment of liabilities, although this nuisance would be expected primarily in relation to Factor 3, or Factor 2. However, the results show that in the case of Factor 3 the problem is the overall level of income. And credit cards are perceived as a way of delaying payments (and transferring expenses between periods – Factor 2) and they are not directly related to inconveniences being a consequence of indebtedness.

Factor 2 concerns borrowing with credit cards. Credit cards are, on the one hand, a payment instrument, but they also serve as a source of credit. Due to the conservative credit policy of Polish banks, credit cards have not become so popular instrument in Poland as in many other countries, such as the United States. Holders of these cards are a group of people with a relatively more

stable financial situation (which is necessarily accompanied by a higher income level), and these cards are typically not a tool for emergency borrowing (see the relationship between Factor 2 and Factor 3) – even when respondents use the entire limit in the card.

Emergency borrowing is described by Factor 3. It concerns a situation when debt is a consequence of difficulties in covering current costs. Although this type of indebtedness is the most dangerous form of debt, it is not – as already mentioned – associated with a sense of nuisance. At the same time, credit cards are not a tool for incurring this type of debt.

The presented results raise the question about the nature of the excessive borrowing phenomenon in the context of Polish households. Excessive borrowing and insufficient savings can be seen as a product of bounded rationality (among the psychological mechanisms that can lead to this type of behavior is distinguished myopia, procrastination, optimism bias, "miswanting", and so-called cumulative cost neglect). The obtained results indicate that such behaviors are not necessarily related in Poland to forms of indebtedness perceived as riskier (emergency debt, credit card debt). The main burdens – including burdens that are perceived by the respondents as onerous - are associated with "conventional" debt.

Comparing the identified aspects of saving and borrowing, it is worth paying attention to the time horizon. While pension savings were separated as a distinct component, long-term loans (mortgages) came under one factor with other loans and credits.

Table 4. Determinants of factors identified for borrowing behavior

	Factor 1	Factor 2	Factor 3
Economic resourcefulness (variable imputed on the basis of SEM)	**0.172*** **	**0.188*** **	**-0.176*** **
Total monthly net income of the household (in PLN)	-0.056**	-0.033	0.010
Age (in years)	**0.016*** **	**0.004**	**0.006**
Constant	-1.091***	-0.847***	0.570**
Adjusted R^2	0.038	0.029	0.023

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Having identified aspects of indebtedness, the second stage of the analysis is an attempt to assess the significance of selected determinants. To the set of analyzed determinants – as in the case of saving – economic resourcefulness, age and income level were included. OLS regression results are presented in Table 4. From the point of view of the considerations being made, the dependence on the third of the analyzed variables – economic resourcefulness – is of key importance. Obtained results clearly indicate the significance of economic resourcefulness for all identified aspects of borrowing. However, it is worth paying attention to the sign of estimates. Resourcefulness is positively related to Factor 1 and Factor 2, and negatively to Factor 3. This means –

considering the definition of the economic resourcefulness – that not only saving, but also reasonable borrowing and the use of instruments such as credit cards can be a factor that has a positive impact on the financial standing. This is in line with the economic perception of borrowing (especially with regard to the Permanent Income Hypothesis), where borrowing is a natural activity stabilizing the level of consumption and allowing for interim financial transfers (acceleration of consumption). An undoubted problem, however, is excessive consumption, which is not justified in the level of income achieved (both temporary and permanent). In this situation, the indebtedness is not planned (built into the concept of the life cycle), but it is an ad hoc attempt to secure consumption needs as a result of current financial problems (the separation of these two aspects is indicated by the results of factor analysis). This type of behavior negatively affects the long-term building of wealth and the economic stability of the household, which is reflected in the negative impact of economic resourcefulness.

4. Discussion and Conclusion

The definition of the concept of economic resourcefulness indicates the possibility of identifying a certain set of features that are key to the "efficiency" of functioning on the market. Among them are not only aspects directly related to the economic dimension - no less important are personality traits, values in life and factors affecting self-confidence, feeling of "not being inferior". This category of economic resourcefulness proved to be helpful in explaining behaviour in the area of saving and borrowing. It is a step towards the identification of sets of significant features – instead of assessing the impact of respondents' characteristics separately, the impact of the compound category is assessed, which comprehensively describes the ability to cope with problems and challenges in the economic space.

The obtained results show a coherent picture of the behaviours of resourceful people. The resourceful people have a higher propensity to save, including short-term saving, oriented on the purchase of specific goods or services, which directly relates to a financial discipline. These people do not avoid borrowing – even in situations when it involves a significant burden on the household budget (which is very often the case in the case of mortgage loans). However, they are characterized by a lower tendency to indebtedness caused by financial problems which would indicate a limited ability to effectively manage household finances. An interesting exception – the area on which economic resourcefulness (defined in the manner adopted in this article) has no significant impact is the conscious planning of retirement savings. On one hand, it shows the specificity of such saving, which is crucial for building incentive systems aimed at promoting this type of saving. On the other hand, it shows that building pension security does not necessarily have

to be positively related to the pursuit of good financial standing and economic stability in the short and medium term. This suggests the existence of a separate mechanism for collecting this type of savings. This distinction is particularly visible in the effects of factor analysis. The existence of retirement savings is not a differentiating factor - they appear both in total savings (which indicate the ability to accumulate savings) and in retirement savings. This is due to the difficulty of defining which savings constitute a retirement savings – they can comprise all the savings, even if they were not collected for this purpose. The differences relate to having the plan for saving for retirement. This suggests that the features determining good functioning in the short and medium term do not necessarily assume solicitude for the more distant future, which is particularly visible among younger people (up to 45 years of age), who were included in this study. In a sense, it reflects a kind of trade-off between-strategy build upon the focus on achieving well-being in the short (medium) and long-term and refers to the Carrol's buffer-stock model of savings.

As suggested by the results obtained, the sample specification adopted in the analysis helped to show the characteristics of the analyzed group and eliminate the impact of many other potentially significant factors. However, it also constitutes a limitation and determines the direction of further research - to what extent the obtained results can be generalized to the whole population of Poland and other countries.

References

1. Briggs, S. R., Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality. Journal of Personality, 54, 107-147.
2. Carroll, C.D. (1997). Buffer-stock saving and the life cycle/permanent income hypothesis. Quarterly Journal of Economics, 112, 1-55.
3. Kośny, M., Piotrowska, M. (2018). Economic Resourcefulness: Definition and Modeling. Social Indicators Research, online first: https://doi.org/10.1007/s11205-018-2048-3.
4. Shefrin, H., Thaler, R.H. (1988). The behavioral life-cycle hypothesis. Economic Inquiry, 26, 609-643.

Economic Policy Uncertainty and Financial Market Volatility: Evidence from Japan

Takayuki Morimoto
Kwansei Gakuin University, Sanda, Japan

Abstract

In this study, we show a relationship between economic policy uncertainty and financial market volatility in Japanese financial market. Uncertainty is measured by the index of economic policy uncertainty (EPU) based on newspaper coverage, frequency newly developed by Baker et al. Volatility is calculated as a sum of squared intraday returns, which is known as the realized volatility (RV). The EPU and RV are combined with the mixed data sampling (MIDAS) approach in order to investigate how economic policy uncertainty shocks are associated with the Japanese financial market volatility. The result will contribute to financial market research and economic policy studies.

Keywords

Economic policy uncertainty index; Realized volatility; GARCH-MIDAS model; DCC-MIDAS model; Japanese financial market

1. Introduction

Asgharian et al. (2016) investigate US and UK stock market movements using the economic policy uncertainty indices of Baker et al. (2016) in combination with the mixed data sampling (MIDAS) approach. They find that the long-run US-UK stock market correlation depends positively on US economic policy uncertainty shocks while the US long-run stock market volatility depends significantly on the US economic policy uncertainty shocks but not on UK shocks while the UK depends significantly on both.

In this research, we follow Asgharian et al. (2016) and apply their method to Japanese stock market. Specifically, we investigate the relation between Nikkei225 which is the stock index for the Tokyo Stock Exchange (TSE) and individual stocks comprised in TOPIX100 which is composed of Top 100 stocks traded on TSE in light of economic policy uncertainty and stock market volatility. Uncertainty is measured by the index of economic policy uncertainty (EPU) based on newspaper coverage, frequency newly developed by Baker et al. (2016) Volatility is calculated as a sum of squared intraday returns, which is known as the realized volatility (RV). The EPU and RV are combined with the mixed data sampling (MIDAS) approach proposed by Ghysels et al. (2004, 2006) in order to investigate how economic policy uncertainty shocks are associated with the Japanese financial market volatility.

Meanwhile, Engle et al. (2013) use the MIDAS approach to link macroeconomic variables to the long-term component of volatility. They incorporate a mean reverting unit daily heteroscedastic volatility process with a MIDAS polynomial that applies to long-term macroeconomic variables, which is called the generalized autoregressive conditional heteroscedasticity model with MIDAS (GARCH-MIDAS) approach. Now we replace a macroeconomic variable with a monthly EPU in GARCH-MIDAS model following Asgharian et al. (2016). Furthermore, Colacito et al. (2011) introduce a novel component model for dynamic correlations which is called the dynamic conditional correlation model with MIDAS (DCC-MIDAS) approach. DCC-MIDAS model is a natural extension of GARCH-MIDAS model to DCC model advocated by Engle (2002). We also use DCC-MIDAS model to capture the dynamic correlation of volatilities between the market index and individual stocks in TSE.

The EPU index of Japan which can be downloaded on the web site: www.policyuncertainty.com is based on frequency counts of articles in Japan's newspapers, Asahi and Yomiuri. It counts the number of news articles containing the terms uncertain or uncertainty, and one or more policy terms. Policy terms are the Japanese equivalents of `tax', `policy', `spending', `regulation', etc. To capture `spending' by the government, they use a set of four terms: `saishutsu', `kokyo jigyohi', `kokyo toushi', and `kokuhi', see the web site for more details. Our specification employs monthly EPU index of Japan as an explanatory variable in the variance equation of a unit daily GARCH-MIDAS model, which we refer to the model as GARCH-MIDAS-EPU. In our empirical analysis, we first estimate the parameters the GARCH-MIDAS-EPU model pair of two stock returns. After that, we obtain the estimated DCC-MIDAS parameters with the standardized residuals from the GARCH-MIDAS-EPU model using the quasi-likelihood method.

2. Models

In this section, we briefly introduce GARCH-MIDAS-EPU and DCC-MIDAS models which are mentioned above, following Colacito et al. (2011), Asgharian et al. (2016) and Conrad et al. (2014). Let us assume that the vector of returns $\mathbf{r}_t = [r_{1,t},\ldots,r_{n,t}]'$ follows the process:

$$\mathrm{r}_t \sim N(\mu, H_t)$$
$$H_t = D_t R_t D_t$$

where μ is the vector of unconditional means, H_t is the conditional covariance matrix and D_t is a diagonal matrix with standard deviations on the diagonal.

Furthermore, we also assume that:

$$R_t = E_{t-1}\left[\xi_t \xi_t'\right]$$

$$\xi_t = D_t^{-1}(r_t - \mu)$$

where $E_t-1[\cdot]$ is the expectation at time $t-1$ given the observations until time $t-1$. Then we have $r_t = \mu + H_t^{\frac{1}{2}}\xi_t$ with $\xi_t \sim N(0, I_n)$. We refer to g_i and m_i as the short and long run variance components respectively for asset i and denote by N_v^i the number of days that m_i is held fixed. The superscript i indicates that this may be asset-specific and the subscript v differentiates it from a similar scheme that will be introduced later for correlations. In particular, while $g_{i,t}$ moves daily, $m_{i,\tau}$ changes only once every N_v^i days. We assume that for each asset $i = 1,\ldots,n$, univariate returns follow the GARCH-MIDAS process:

$$r_{i,t} = \mu_i + \sqrt{m_{i,\tau} \cdot g_{i,t}}\xi_{i,t}, \quad \forall t = \tau N_v^i, \ldots, (\tau+1)N_v^i$$

where $g_{i,t}$ follows a GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha_i - \beta_i) + \alpha_i \frac{(r_{i,t-1} - \mu_i)^2}{m_{i,\tau_t}} + \beta_i g_{i,t-1}$$

while the MIDAS component $m_{i,\tau}$ is a weighted sum of K_v^i lags of realized variances (RV) over a long horizon:

$$m_{i,\tau} = \bar{m}_i + \theta_i \sum_{l=1}^{K_v^i} \varphi_l\left(w_v^i\right) RV_{i,\tau-l}$$

where the RV involve N_v^i daily squared returns. Namely:

$$RV_{i,\tau} = \sum_{j=(\tau-1)N_v^i+1}^{\tau N_v^i} (r_{i,j})^2$$

where N_v^i could for example be a quarter or a month. The above specification corresponds to the block sampling scheme as defined in Engle et al. (2013), involving so called Beta weights defined as:

$$\varphi_l\left(w_v^i\right) = \frac{\left(\frac{1-l}{K_v^i}\right)^{w_v^i - 1}}{\sum_{j=1}^{K_v^i}\left(\frac{1-j}{K_v^i}\right)^{w_v^i - 1}}$$

where the parameters N_v^i and K_v^i are independent of i, i.e. the same across all series.

We use the two-step DCC-MIDAS model of Colacito et al. (2011) extended to allow for exogeneous variables influencing the long-run volatility and correlation as in Asgharian et al. (2016). The first step consists of estimating

separate GARCH-MIDAS models for the stock returns for day $i = 1,\ldots,N_t$ in month t as:

$$r_{i,t-1} = \mu + \sqrt{\tau_t g_{i,t}}\varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim N(0,1)$$

where the total stock variance $\sigma_{i,t}^2$ is separated into a short-run component $g_{i,t}$ and a long-run component τ_t such that $\sigma_{i,t}^2 = \tau_t g_{i,t}$. A GARCH (1,1) process describes the short-run component:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha\frac{(r_{i,t-1} - \mu)^2}{\tau_t} + \beta g_{i,t-1}$$

where $\alpha > 0$ and $\beta \geq 0, \alpha + \beta < 1$. The long-run component is described by a MIDAS regression where the lagged EPU shocks of the $EPU_t - k$ are included over $k = 1,\ldots,24$:

$$\tau_t = \theta_0 + \theta_1 \sum_{k=1}^{K} \varphi_k EPU_{t-k}$$

where the weighting scheme is described by a beta lag polynomial:

$$\varphi_l(w_v^i) = \frac{\left(1 - \frac{l}{K_v^i}\right)^{w_v^i - 1}}{\sum_{j=1}^{K_v^i}\left(1 - \frac{j}{K_v^i}\right)^{w_v^i - 1}}$$

where the parameter ϑ_1 measures the effects of the economic policy uncertainty shocks on the long-run volatility. We fix $w_1 = 1$ to ensure higher weights to the most recent observations as with Asgharian et al. (2016).

Colacito et al. (2011) propose the DCC-MIDAS model which is a natural extension of the GARCHMIDAS model to the Engle (2002) DCC model. Using the standardized residuals , it is possible to obtain a matrix whose elements are:

$$q_{i,j,t} = \bar{\rho}_{i,j,t}(1 - a - b) + a\xi_{i,t-1} + bq_{i,j,t-1}$$

$$\bar{\rho}_{i,j,t} = \sum_{l=1}^{K_c^{ij}} \varphi_l\left(w_r^{ij}\right) c_{i,j,t-l}$$

$$c_{i,j,t} = \frac{\sum_{k=t-N_c^{ij}}^{t} \xi_{i,k}\xi_{j,k}}{\sqrt{\sum_{k=t-N_c^{ij}}^{t} \xi_{i,k}^2}\sqrt{\sum_{k=t-N_c^{ij}}^{t} \xi_{j,k}^2}}$$

where we could have used simple cross-products of $\xi_{i,t}$ in the above formulation of $c_{i,j,t}$. The normalization allow us to discuss regularity conditions in terms of correlation matrices. Correlations can then be computed as:

$$\rho_{i,j,t} = \frac{q_{i,j,t}}{\sqrt{q_{i,i,t}}\sqrt{q_{j,j,t}}}$$

where we regard $q_{i,j,t}$ as the short run correlation between assets i and j, whereas $\rho_{i,j,t}$ is a slowly moving long run correlation. Rewriting the first equation of system as

$$q_{i,j,t} - \overline{\rho}_{i,j,t} = a\left(\xi_{i,t-1}\xi_{j,t-1} - \overline{\rho}_{i,j,t}\right) + b\left(q_{i,j,t-1} - \overline{\rho}_{i,j,t}\right)$$

conveys the idea of short run fluctuations around a time-varying long run relationship. The idea captured by the DCC-MIDAS model is similar to that underlying GARCH-MIDAS. In the GARCHMIDAS the short run component is a GARCH component, based on daily returns, that moves around a long-run component driven by realized volatilities computed over a monthly basis, see Colacito et al. (2011).

3. Empirical Analysis

We apply the DCC-MIDAS with GARCH-MIDAS-EPU model to Nikkei225 and TOPIX100 data listed on TSE from June 1988 to April 2016 in order to investigate the relation between economic policy uncertainty and financial market volatility in Japanese financial market. Here is an example of the results of our empirical analysis. Table 1 shows the result of GARCH-MIDAS-EPU for NK225 from January 1991 to April 2016. Figures below show the plots of estimated short- and long-run variances and correlations for Japan Tobacco (JT) Inc. (2914) and Nikkei225.

	Estimates	S.E.	t-stats	P-values
$\hat{\mu}$	0.0004	0.0002	2.4912	0.0127
$\hat{\alpha}$	0.1023	0.0065	15.6724	0.0000
$\hat{\beta}$	0.8785	0.0071	124.1653	0.0000
$\hat{\theta}$	0.0015	0.0005	2.6766	0.0074
$\hat{w}_2$	1.0076	0.0185	54.6014	0.0000
$\hat{m}$	0.0003	0.0000	10.3594	0.0000

Table 1 Result of GARCH-MIDAS-EPU for NK225 from January 1991 to April 2016

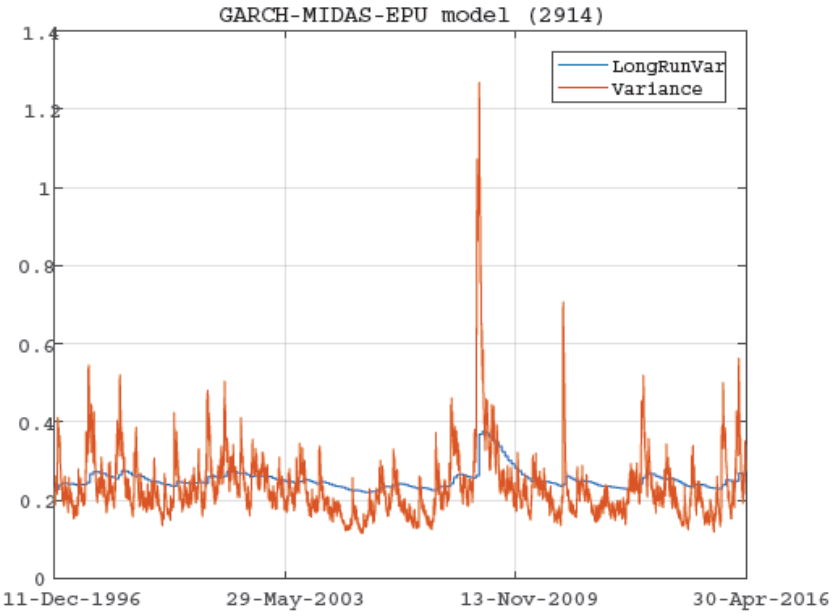


Fig. 1 Long and Short Run Variances (JT)

4. Conclusions

The result our empirical analysis will contribute to financial market research and economic policy studies.

Acknowledgements

The author would like to thank Yoshinori Kawasaki (Professor at the Institute of Statistical Mathematics) for his comments and discussions. This study is partly supported by the Institute of Statistical Mathematics (ISM) cooperative research program (2018-ISM-CRP-2010) and JSPS KAKENHI Grant Number 18K01554.

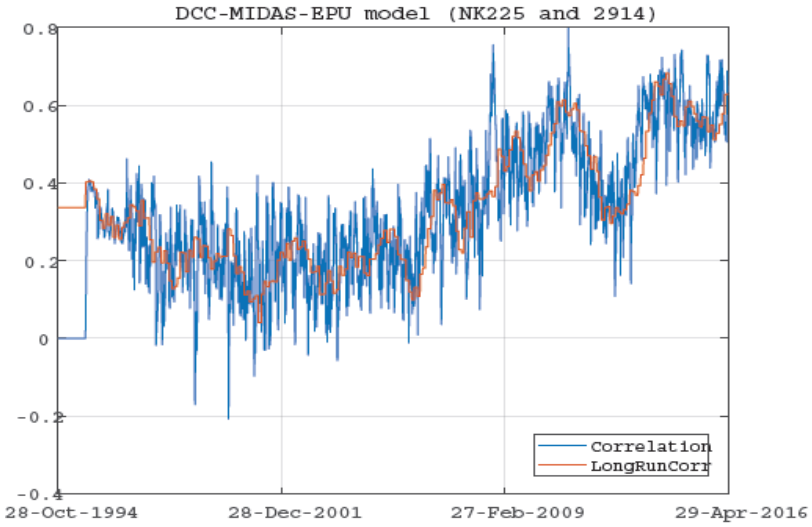


Fig. 2 Long and Short Run Correlations (NK225 and JT)

References:

1. Asgharian, H., Christiansen, C., Gupta, R., and Hou, A. J. (2016). Effects of Economic Policy Uncertainty Shocks on the Long-Run US-UK Stock Market Correlation (October 3, 2016). Available at SSRN: https://ssrn.com/abstract=2846925 or http://dx.doi.org/10.2139/ssrn.2846925.
2. Baker, S.R., Bloom, N., and Davis, S.J. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131: 1593-1636.
3. Colacito, R., Engle, R.F., and Ghysels, E. (2011). A Component Model for Dynamic Correlations. Journal of Econometrics 164: 45-59.
4. Conrad, C., Loch, K., and Rittler, D. (2014). On the macroeconomic determinants of long-term volatilities and correlations in US stock and crude oil markets. Journal of Empirical Finance 29: 26-40.
5. Engle, R. (2002). Dynamic conditional correlation - a simple class of multivariate GARCH models. Journal of Business and Economic Statistics 20: 339-350.
6. Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. The Review of Economics and Statistics 95: 776-797.
7. Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models, CIRANO Working Paper 2004s-20.
8. Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. Journal of Econometrics 131: 59-95.

Combining LASSO and Liu type Estimator in the Linear Regression Model

M. Kayanan[1, 2], P. Wijekoon[3]

[1]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka

[2]Deparment of Physical Science, Vavuniya Campus of the University of Jaffna, Vavuniya, Sri Lanka

[3]Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka

Abstract

The Ordinary Least Square Estimator (OLSE) has been widely used to estimate unknown parameters in the linear regression model. Since OLSE produces high variance on the estimates when multicollinearity exists among the predictor variables, the Ridge Estimator (RE) is introduced as an alternative estimator. However, RE yields heavy bias in the high dimensional linear regression models, and it also produces irrelevant predictors to the estimated model. Hence, the Least Absolute Shrinkage and Selection Operator (LASSO) has been used to ensure the variable selection as well as to handle the multicollinearity problem simultaneously. It is noted that LASSO failed to outperform RE when high multicollinearity exists among the predictor variables. Further, the LASSO estimator is unstable when the number of predictors is higher than the number of observations. Hence, the Elastic net (Enet) estimator is introduced to address this problem by combining LASSO and RE. Since Liu Estimator (LE) is an alternative estimator for RE to address multicollinearity problem, the objective of this study was to propose Liu type Elastic net estimator by combining LASSO and LE. Then, we compared the prediction performance of the Liu type Elastic net (LEnet) estimator with the Elastic net and LASSO estimators in Root Mean Square Error (RMSE) sense using the real-world examples. The results showed that LEnet outperforms the other two estimators in RMSE sense.

Keywords

Multicollinearity; Variable selection; Liu estimator; LASSO; Elastic net

1. Introduction

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observations on the predictor variable, $\boldsymbol{X}$ is the $n \times p$ matrix of observations on p non stochastic regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vectors of unknown parameters, $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of disturbances, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$.

The Ordinary Least Squares (OLS) is the usual approach to estimate the unknown parameter vector $\boldsymbol{\beta}$, and the Residual Sum of Squares (RSS) of the model (1) takes the form

$$RSS = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}). \quad (2)$$

By minimising RSS, the Ordinary Least Squares Estimator (OLSE), which is the Best Linear Unbiased Estimator (BLUE) for $\boldsymbol{\beta}$, is defined as

$$\boldsymbol{\beta}_{OLSE} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}. \quad (3)$$

It is well-known that OLSE is unstable and produces estimates with high variance when columns of $\boldsymbol{X}$ are multicollinear. A general approach to handle this problem is to introduce a penalty in RSS, which produces bias but reduces the variance of the estimators.

The Ridge Estimator (RE) was proposed by Hoerl & Kennard (1970) by introducing L2-norm of $\boldsymbol{\beta}$ as a penalty in RSS as below:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{RE} &= argmin_{\boldsymbol{\beta}}\left\{(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}})'(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}}) + k\sum_{j=1}^{p}|\beta \quad \hat{}_j|^2 \right\} \\ &= (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y}\end{aligned} \quad (4)$$

where $k > 0$ is the shrinkage parameter. RE helps with obtaining less variance of the estimates by shrinking the regression coefficients toward zero. However, it has two significant issues in high dimensional linear models as it introduces heavy bias when the number of predictors is high, and it may shrink irrelevant regression coefficients, but they are still in the model. As a remedial solution to this problem, Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO) by introducing L1-norm of $\boldsymbol{\beta}$ as a penalty in RSS, and it is defined as

$$\widehat{\boldsymbol{\beta}}_{LASSO} = argmin_{\boldsymbol{\beta}}\left\{(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}})'(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}}) + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \right\} \quad (5)$$

where $\lambda > 0$ is the shrinkage parameter. LASSO handles both multicollinearity and variable selection simultaneously in the high dimension linear regression model. However, LASSO is unstable when the number of predictors p is higher than the number of observations n. Further, the prediction performance of RE dominates LASSO if there exist high multicollinearity among predictors.

To handle this problem, Zou & Hastie (2003) proposed Elastic net (Enet) estimator by combining RE and LASSO, and it is defined as

$$\widehat{\boldsymbol{\beta}}_{Enet} = argmin_{\boldsymbol{\beta}}\left\{(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}})'(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}}) + k\sum_{j=1}^{p}|\hat{\beta}_j|^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|\right\} \quad (6)$$

Liu (1993) proposed Liu Estimator (LE) as an alternative estimator to RE to handle the multicollinearity problem. They have shown that LE outperforms RE under certain conditions. This work is motivated us to combine LE and LASSO estimators and name it as Liu type Elastic net estimator (LEnet). Further,

we analysed the performance of LEnet with LASSO and Enet estimators in Root Mean Square Error (RMSE) sense using the real-world examples.

2. Methodology

2.1. Method to obtain LASSO solutions

The solution of LASSO has been obtained using a modified version of the Least Angle Regression (LARS) algorithm (Efron et al. (2004)), and it is outlined as below:

Step 1: Centre the response variable $\boldsymbol{y}$ that has to mean zero, and standardise the predictors $\boldsymbol{X}$ that has mean zero and unit norm.

Step 2: Start with all estimates of the coefficients $\widehat{\boldsymbol{\beta}}$ to be equal to 0 with the residual $\boldsymbol{r}_0 = \boldsymbol{y}$.

Step 3: Find the predictor x_{j1} most correlated with r_0.

$$x_{j1} = \max_j \left| Cor\left(x_j, r_0\right) \right| \qquad j = 1, 2, \ldots, p.$$

Step 3: Move the estimate of $\hat{\beta}_{j1}$ from 0 towards the OLSE coefficients until some other predictor x_{j2} has as large a correlation with the current residual as x_{j1} does. At this point instead of continuing in the direction based on x_{j1}, LAR proceeds in the direction of equiangularity between the two predictors x_{j1} and x_{j2}.

Step 4: A third variable x_{j3} eventually earns its way into the most correlated (active set), and then LARS proceeds equiangular between x_{j1}, x_{j2} and x_{j3}. Continue adding variables to the active set in this way moving in the direction defined by the least angle direction.

On this step, the coefficient estimates are updating using the following formula:

$$\boldsymbol{\beta} ji = \boldsymbol{\beta} j(i-1) + \alpha i \boldsymbol{u} i \qquad (7)$$

where α_i is a value between [0, 1] which represents how far the estimate of moves in the direction before another variable enters the model and the direction changes again, and $\boldsymbol{u}_i$ is the equiangular vector.

The direction $\boldsymbol{u}_i$ is calculated using the following formula:

$$\boldsymbol{u}_i = \boldsymbol{E}_i (\boldsymbol{E}_i' \boldsymbol{X}' \boldsymbol{X} \boldsymbol{E}_i)^{-1} \boldsymbol{E}_i' \boldsymbol{X}' \boldsymbol{r}_{i-1} \qquad (8)$$

where $\boldsymbol{E}_i$ is the matrix with column (e_{j1} , e_{j2} ,… . , e_{ji}), and e_j be the j^{th} standard unit vector in $\mathbb{R}^p$.

Then, choose α_i as given below:

$\alpha_i = \min\{\alpha_i \in [0,1]$: ($\alpha_i = \alpha_{ji}^{+}$ or $\alpha_i = \alpha_{ji}^{-}$ for some j such that $\boldsymbol{\beta}_{j(i-1)} = 0$) or ($\alpha_i = \alpha_{ji}^{*}$ for some j such that $\boldsymbol{\beta}_{j(i-1)} \neq 0$)} (9)

where

$$\alpha_{ji}^{\pm} = \frac{Cor\left(\boldsymbol{r}_{i-1}, \boldsymbol{x}_{ji}\right) \pm Cor\left(\boldsymbol{r}_{i-1}, \boldsymbol{x}_j\right)}{Cor\left(\boldsymbol{r}_{i-1}, \boldsymbol{x}_{ji}\right) \pm Cor\left(\boldsymbol{H}_i \boldsymbol{r}_{i-1}, \boldsymbol{x}_j\right)} \quad ; \boldsymbol{H}_i = \boldsymbol{X}\boldsymbol{E}_i \left(\boldsymbol{E}_i^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{E}_i\right)^{-1} \boldsymbol{E}_i^T \boldsymbol{X}^T$$

and

$$\alpha_{ji}^{*} = -\frac{\widehat{\boldsymbol{\beta}}_{j(i-1)}}{\boldsymbol{u}_i}$$

If $\alpha_i = \alpha_{ji}^{*}$ for some j such that $\boldsymbol{\beta}_{j(i-1)} \neq 0$, then $\boldsymbol{E}_i$ is the matrix formed by removing the column e_j from $\boldsymbol{E}_{i-1}$. Note that this is the modification done in LARS algorithm to obtain LASSO estimates. Then the residual $\boldsymbol{r}_i$ related to the current step is calculated as

$$\boldsymbol{r}_i = \boldsymbol{r}_{i-1} - \alpha_i \boldsymbol{X}\boldsymbol{u}_i$$

and then, move to the next step where j_{i+1} is the value of such that $\alpha_i = \alpha_{ji}^{+}$ or $\alpha_i = \alpha_{ji}^{-}$.

Repeat this step until $\alpha_i = 1$.

2.2. LARS-EN algorithm

The LARS-EN algorithm (Zou & Hastie (2003)) is used to obtain Elastic net estimates, and it is also a modified version of the LARS algorithm. In LARS-EN algorithm, the equiangular vector $\boldsymbol{u}_i$ of the LARS algorithm in the equation (8) is replaced by incorporating RE as follows:

$$\boldsymbol{u}_i = \boldsymbol{E}_i(\boldsymbol{E}_i'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I})\boldsymbol{E}_i)^{-1}\boldsymbol{E}_i'\boldsymbol{X}'\boldsymbol{r}_{i-1}, \quad (10)$$

and the rest of the steps are similar to the algorithm status above.

2.3. LARS-LEnet algorithm

According to Liu (1993), the LE is defined as

$$\boldsymbol{\beta}_{LE} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X} + d\boldsymbol{I})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (11)$$

where $0 < d < 1$ is the shrinkage parameter.

Now we propose LARS-LEnet algorithm by incorporating LE in the LARS algorithm. Here we modify the equiangular vector $\boldsymbol{u}_i$ of the LARS algorithm in the equation (8) as:

$$\boldsymbol{u}_i = \boldsymbol{E}_i(\boldsymbol{E}_i'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})\boldsymbol{E}_i)^{-1}(\boldsymbol{E}_i'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{d}\boldsymbol{I})\boldsymbol{E}_i)(\boldsymbol{E}_i'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{E}_i)^{-1}\boldsymbol{E}_i'\boldsymbol{X}'\boldsymbol{r}_{i-1},$$

(12) and all other steps described in section 2.1 are the same.

2.4. Performance evaluation

The performance of LEnet, Enet and LASSO estimators were compared using Root Mean Square Error (RMSE) sense, which is the expected prediction error. The RMSE is defined as

$$RMSE = (\boldsymbol{y}_{new} - \boldsymbol{X}_{new}\boldsymbol{\beta})'(\boldsymbol{y}_{new} - \boldsymbol{X}_{new}\boldsymbol{\beta}) \quad (13)$$

where $(\boldsymbol{y}_{new}, \boldsymbol{X}_{new})$ denotes new data that are not used to obtain the coefficient estimates $\boldsymbol{\beta}$.

In this study, we considered two real-world examples, namely the Prostate Cancer Data (Stamey et al. (1989)), and the UScrime dataset (Venables & Ripley (1999)), to compare the performance of the three estimators LEnet, Enet and LASSO.

In the Prostate Cancer Data, the predictors are eight clinical measures: log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log capsular penetration (**lcp**), Gleason score (**gleason**) and percentage Gleason

score 4 or 5 (**pgg45**). The response is the log of prostate specific antigen (**lpsa**). The dataset has 97 observations. Stamey et al. 1989 have examined the correlation between the level of prostate specific antigen and those eight clinical measures. Further, Tibshirani (1996), Efron et al. (2004) and Zou & Hastie (2003) have used this data to examine the performance of LASSO, LARS algorithm and Enet estimators. This data set is attached with "lasso2" R package, and we have used 50 observations to fit the model, and 47 observations to calculate the RMSE.

The UScrime dataset has 16 variables with 47 observations, and it is attached with "MASS" R package. This data contains the following columns: $\boldsymbol{M}$ (percentage of males aged 14--24), $\boldsymbol{So}$ (indicator variable for a Southern state), $\boldsymbol{Ed}$(mean years of schooling), $\boldsymbol{Po1}$(police expenditure in 1960), $\boldsymbol{Po2}$ (police expenditure in 1959), $\boldsymbol{LF}$ (labor force participation rate), $\boldsymbol{M}.\ \boldsymbol{F}$ (number of males per 1000 females), $\boldsymbol{Pop}$ (state population), $\boldsymbol{NW}$ (number of non-whites per 1000 people), $\boldsymbol{U1}$ (unemployment rate of urban males 14--24), $\boldsymbol{U2}$ (unemployment rate of urban males 35--39), $\boldsymbol{GDP}$ (gross domestic product per head), $\boldsymbol{Ineq}$ (income inequality), $\boldsymbol{Prob}$ (probability of imprisonment), $\boldsymbol{Time}$ (average time served in state prisons), $\boldsymbol{y}$ (rate of crimes in a particular category per head of population). The variable $\boldsymbol{y}$ is considered as a dependent variable, the variable $\boldsymbol{So}$ is ignored since it is categorical. Venables & Ripley (1999) have examined the effect of punishment regimes on crime rates using this dataset. For the analysis, we have used 40 observations to fit the model, and 7 observations to calculate the RMSE.

The RMSE for different values of the shrinkage parameter (k/d) was calculated, and the shrinkage parameter (k/d) was chosen between (0, 1) for simplicity. Further, we used K-fold cross-validation to find the optimal values of λ, k and d.

3. Result and Discussion

The Estimated RMSE values of LASSO, Enet and LEnet for the Prostate Cancer Data and UScrime data are displayed in Table 1 and Table 2, respectively.

Table 1. Estimated RMSE values of the estimators for Prostate Cancer Data

	LASSO estimator								
λ	1.417	1.417	1.417	1.417	1.417	1.417	1.417	1.417	1.417
RMSE	23.114	23.114	23.114	23.114	23.114	23.114	23.114	23.114	23.114
	Enet estimator								
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
λ	1.599	1.595	1.592	1.588	1.585	1.582	1.579	1.576	1.574
RMSE	21.183	21.178	21.173	21.169	**21.164**	**21.160**	**21.157**	**21.153**	**21.150**

	LEnet estimator								
d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
λ	1.574	1.577	1.580	1.583	1.586	1.589	1.592	1.596	1.599
RMSE	**21.150**	**21.153**	**21.157**	**21.161**	21.165	21.169	21.174	21.178	21.183

Table 2. Estimated RMSE values of the estimators for UScrime data

	LASSO estimator								
λ	1194	1194	1.417	1194	1194	1194	1194	1194	1194
RMSE	48249	48249	48249	48249	48249	48249	48249	48249	48249
	Enet estimator								
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
λ	1024	1021	1020	1043	1330	1267	1241	1024	1136
RMSE	52101	51619	51164	53755	**30229**	**29840**	**31492**	**27200**	**33610**
	LEnet estimator								
d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
λ	1185	1212	1341	1479	1046	1048	1024	1192	1193
RMSE	**33484**	**25737**	**30227**	**29231**	53766	54257	51531	47925	48092

According to Table 1 and Table 2, we can observe that LEnet outperforms the other two estimators when d < 0.5.

Table 3 and Table 4 show the cross-validated RMSE values of LASSO, Enet and LEnet for the Prostate Cancer Data and UScrime data, respectively.

Table 3. Cross-validated RMSE values of for Prostate Cancer Data

Estimators	**Number of Variables Selected**	**Optimal Shrinkage parameter Values**	**RMSE**
LASSO	5	λ = 1.498	23.114
Enet	7	k = 0.8 and λ = 1.499	21.153
LEnet	7	d = 0.17 and λ = 1.498	**21.152**

Table 4. Cross-validated RMSE values of for UScrime Data

Estimators	**Number of Variables Selected**	**Optimal Shrinkage parameter Values**	**RMSE**
LASSO	11	λ = 1158	820239
Enet	12	k = 0.96 and λ = 1143	573295
LEnet	12	d = 0.10 and λ = 1158	**569234**

According to Table 3 and Table 4, we can observe that LEnet produces minimum RMSE compared to the other two estimators.

4. Conclusion

In this study, we introduced LEnet estimator by proposing LARS-LEnet algorithm, and we showed that LEnet outperforms LASSO and Enet in the RMSE sense. Simulation study will be employed in future studies to support our results.

References

1. Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 55–67.
2. Tibshirani, R. (1998). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
3. Zou, H., & Hastie, T. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B, 67, 301-20.
4. Liu, K. (1993). A new class of biased estimate in linear regression. Communication in Statistics - Theory and Methods, 22, 393–402.
5. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.
6. Stamey, T., Kabalin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E., and Yang, N. (1989),
   "Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate ii. Radical Prostatectomy Treated Patients," Journal of Urology, 16, 1076–1083.
7. Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition.
   Springer.

Spatial analysis for forced displacement and war actions: Colombian case

Carlos A Mantilla Duarte
Universidad Industrial de Santander, Bucaramanga, Colombia

Abstract

From 2011 to 2016, Colombia Government and Colombian Revolutionary Armed Forces held dialogues in order to end the armed conflict in Colombia. Spatial relationships can be explained using spatial data analysis methodologies to compare spatial configuration before and after dialogs. This relationships was analyzed using spatial statistical methodologies like **Moran's I** (Global and Local) and **Spatial Regression** comparing likelihood functios from **SAR**, **SEM** and **SDM** models. The **SDM Model** fitted better in both cass. It's necessary to analyze missing data to adjust Spatial Panel Data Models or complement with functional data analysis.

Keywords

Spatial Analysis; Spatial Models, Spatial Correlation; Changes in Spatial Relationships

1. Introduction

Since 2011, Colombia government began negotiations with Colombian Revolutionary Armed Forces (a.k.a. FARC) with the objective to end an armed conflict with more than 50 years of duration. Mantilla (2017) describes spatial relationships between some common factors about armed conflict in Colombia. But, the spatial configuration of the variables shows some changes that are not described in the work of Mantilla (2017).

This paper research a different approach to describe the changes that the variables show using **Global and Local Spatial Correlation Moran's Indexes** and Spatial Regression likelihood function from **SAR**, **SEM** and **SDM** models. The first part contains a summary of methods used in the analysis. After this, the paper exhibit the results of analysis and, finally, exposes a short discussion about the models fitted and contains conclusions,

2. Methods

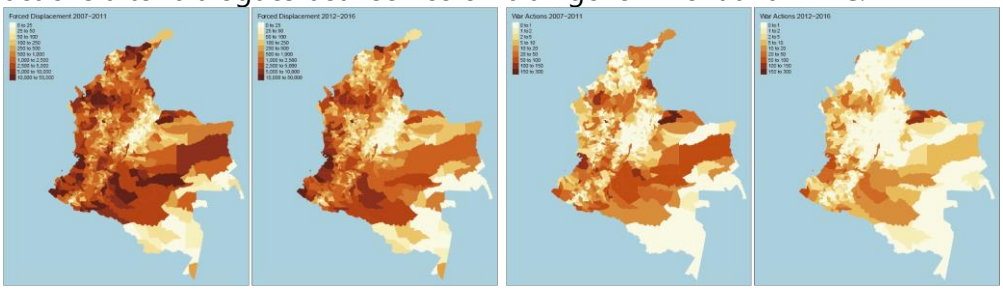
The selected variables are based on the work of Mantilla (2015, 2017) and Mantilla & Angulo (2017). The data set corresponds to open data by Office for the Coordination of Humanitarian Affairs United Nations in Colombia (OCHA-UN), The Observatory of Human Rights in Colombia, the Directorate of Criminal Investigation and INTERPOL of the National Police (Colombia) - DIJIN

- and The Unit for Comprehensive Care and Reparation for the Victims of the Colombian Armed Conflict. Selected variables was analyzed using **Global and Local Spatial Correlation Moran's Indexes** and Spatial Regression likelihood function from **SAR**, **SEM** and **SDM** models.

Data was separated in two groups: Before dialogs (2007 - 2011) and after dialogs (2012 - 2016)[3]. Because of data structure, it was not possible analyzed under spatial panel data methodologies.

3. Results

Figure 1(a) shows the changes about spatial distribution for forced displacement. In Figure 1(b) it's possible to observed changes about war actions after dialogues between colombian government and FARC.



(a) Forced Displacement (b) War Actions

Figure 1: Spatial Plots for Forced Displacement - War Actions

Usin *Moran's I* (Bongiovanni, 2008) it's possible to identify the location of spatial conglomerates. *Moran's I* is calculate like in equation 1

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - \overline{y})\,(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{1}$$

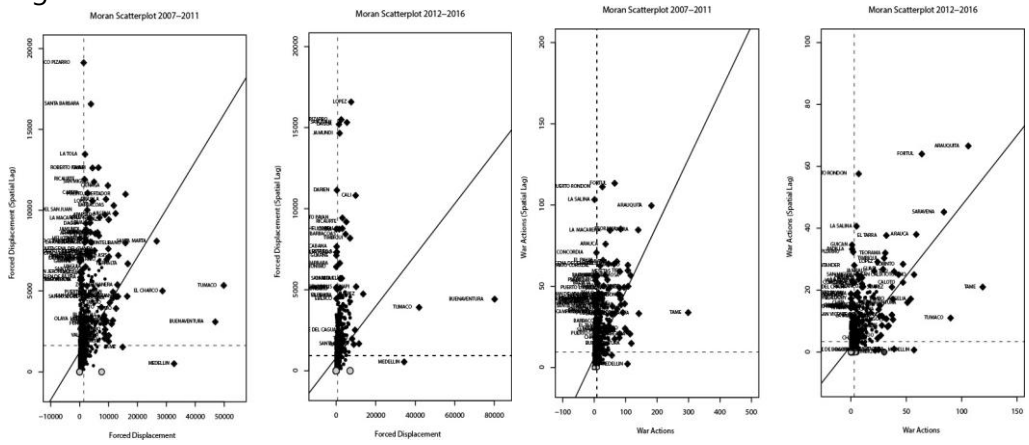
Where, w_{ij} represents Spatial Weight for the couple *i,j*. Most spatial weights matrices **W** are based on some version of a connectivity matrix **C**. **C** is an $n \times n$ binary matrix, where $i = \{1,2,...,n\}$ and $j = \{1,2,...,n\}$ are the units in the system (cities in international system). Entry $c_{ij} = 1$ if two units $i \neq j$ are considered connected, and $c_{ij} = 0$ if they are not. Results for Moran's I are shown In Table 1. The evidence suggests the presence of spatial correlation in all data groups. Note Moran's I for displacement is low but p-value suggests spatial autocorrelation true.

[3] Since the end of 2011, the Colombian government began negotiations with the FARC to end an armed conflict

Table 1: Spatial Correlation (Moran's I)

Data	Moran I	Expectation	Variance	St Dev	P-Value
Desplacement 2007-2011	0,2950	-0,0009	0,0003	17,2460	<2,2e-16
Desplacement 2012-2016	0,1586	-0,0009	0,0002	11,0280	<2,2e-16
War Actions 2007-2011	0,4076	-0,0009	0,0003	23,4990	<2,2e-16
War Actions 2012-2016	0,4537	-0,0009	0,0003	26,1590	<2,2e-16

Figures 2(a) and 2(b) show Local Spatial Correlation Scatter Plot for colombian cities. This index was calculated using Equation 2 (Zhukiv, 2010) and allows to identify the spatial clusters of territorial entities (municipalities) with high or low values for $z-$ *scores*.



(a) Forced Displacement

(b) War Actions

Figure 2: Moran Scatterplot for Forced Displacement - War Actions

$$I_i = \frac{(y_i - \overline{y}) \sum_{j=1}^{n} w_{ij} (y_j - \overline{y})}{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}} \tag{2}$$

Equation 3 is used to calculate $z-$ *scores* for local spatial autocorrelation (Zhukov, 2010):

$$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{E[I_i]}} \tag{3}$$

where

$$E[I_i] = -\frac{\sum_{i=1, j \neq 1}^{n} w_{ij}}{n-1}$$

and

$$V[I_i] = E\,\mathrm{h}I_i^{2\mathrm{i}} - E[I_i]^2$$

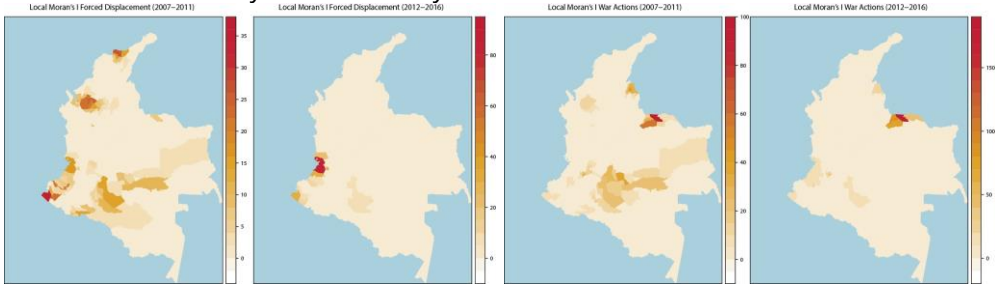
Figures 3(a) and 3(b) show spatial clusters with high values for $z - scores$. Note significatives changes about Forced Displacement. It's possible to observe 6 clusters in period 2007 2011 and only 3 clusters in period 2012 - 2016. On the other hand, war actions show changes in the clusters observed, going from 10 during the period 2007 - 2011 to 5 in the period 2012 - 2016. With this results it's possible to propose some spatial regression models: SAR, SEM and SDM.

4. Discussion and Conclusions

This paper consider the following models for Spatial Analysis about colombian forced migrations: Spatial Autoregressive Model (SAR), Spatial Error Model (SEM) and Spatial Durbin Model (SDM). Table 2 shows differents parameters for the models estimates. Previously, data was separated in two group: before and after of dialogues to understand changes in spatial relationships between variables in study.

Evidence suggests that spatial effects can't be modeled by SAR model in first group and SEM models in both. The value for the Moran's I in the SDM model suggests spatial randomness in the errors in both. For SAR model in group after dialogues the value for the Moran's I in

with more than 50 years of history.



(a) Forced Displacement (b) War Actions

Figure 3: Moran Local Plot for Forced Displacement - War Actions

Table 2: Spatial Models Forced Displacement - War Actions

Data	Model	Coefficients			Spatial Correlation Residuals				AIC
		Coef	Estimate	Prob	Moran's I	Expectation	Variance	Prob	
2007 2011	SAR	(Intercept)	154,071	0,086	0,0497	-0,0009	0,0003	0,0015	20862,2
		War Act	93,632	0,000					
		ρ	0,289	0,000					
	SEM	(Intercept)	442,449	0,001	-0,0296	-0,0009	0,0003	0,0015	20862,2
		War Act	107,382	0,000					
		λ	0,415	0,000					
	SDM	(Intercept)	298,139	0,001	-0,0291	-0,0009	0,0003	0,9515	20863,8
		War Act	108,811	0,000					
		lag.w.a.(θ)	-50,209	0,000					
		ρ	0,415	0,000					
2012 2016	SAR	(Intercept)	153,629	0,081	0,0540	-0,0009	0,0002	0,0000	20942,7
		War Act	164,905	0,000					
		ρ	0,192	0,000					
	SEM	(Intercept)	236,980	0,049	-0,0048	-0,0009	0,0002	0,6170	20942.73
		War Act	191,862	0,000					
		λ	0,331	0,000					
	SDM	(Intercept)	289,770	0,001	-0,0040	-0,0009	0,0002	0,5905	20929,8
		War Act	209,265	0,000					
		lag.w.a.(θ)	-113,425	0,000					
		ρ	0,316	0,000					

the SDM model suggests spatial randomness too. The SDM results in a slightly better fit.Using Akaike's Information Criteria (AIC) to compare models, it can to seen that the SDM model results in a slightly better fit. The model that fits better is observed in Equation 4.

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \epsilon \tag{4}$$

where, $\beta = \begin{pmatrix} 298.139 \\ 108.811 \end{pmatrix}$, $\theta = -50.209$, $\rho = 0.415$ $\mathbf{W}$ is spatial weights matrix, $\mathbf{X}$ is the covariables matrix and $\alpha\iota_n = (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\epsilon$

For second data group $\beta = \begin{pmatrix} 298.770 \\ 209.265 \end{pmatrix}$, $\theta = -113.425$, $\rho = 0.316$ $\mathbf{W}$

Conclusions:

- SEM models can't be explain relationships between this particular variables.
- SDM models results in a slightly better fit.
- Data structure suggests to prove other forms for models in addition to SDM Models like Spatial Panel Data Models or Spatio-Temporal Data Models.
- it is necessary to analyze missing data to implement additional models.
- Next step consists in to analyze data using functional data methodologies.

References

1. Baronio, A., Vianco, A. & Rabanal C. (2012). Una Introducci´on a la Econometr´ıa Espacial. Dependencia y Heterogeneidad. Recovered from: http://http://www.econometricos.com.ar/.
2. Bongiovanni, R. (2008). Econometr´ıa Espacial Aplicada a la Agricultura de Precisi´on. Actualidad Econ´omica - Ao XIX - N 67.
3. Lovelace, R. and others (2015). Introduction to Visualising Spatial Data in R. Recovered from: https://cran.r-project.org.
4. Mantilla, C. (2015). Factorial Analysis for Forced Migration in Colombia. International Statistical Institute (ISI) - 60th World Statistics Congress ISI2015. Recovered from: http://www.isi2015.org/.
5. Mantilla, C. (2017). Spatial Analysis for Common Factors in Forced Migrations. Colombian Case. International Statistical Institute (ISI) - 61th World Statistics Congress ISI2017. Recovered from: http://www.isi2017.org/.
6. Mantilla, C. & Angulo, J. (2017). Modelamiento de Datos Espaciales y Datos Espacio-Temporales: An´alisis Espacial y Espacio-Temporal del Desplazamiento Forzado (Tesis de Maestr´ıa). Universidad de Granada. Granada, Spain
7. Pebesma, E. (2012). Spacetime: Spatio-Temporal Data in R. Recovered from: https://www.jstatsoft.org.
8. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Zhukov, Y. (2010). Applied Spatial Statistics in R. Recovered from: http://scholar.harvard.edu/zhukov/classes/ spatial-statistics-r.

An attempt to determine a confidence interval for the economic growth rate-Case of the Moroccan economy[4]

Ahmed Oulad El Fakir
Haut-Commissariat of Planning

Abstract

In Morocco, several departments work out economic forecasts and publish figures which are sometimes far from each other. These departments are the Haut-Commissariat of Planning (HCP), the Ministry of Economy and Finance (MEF), Bank Al-Maghrib (BAM which is the Moroccan Central Bank), the Research Centers (including the Moroccan Center of the Conjoncture -or CMC- which is close to the private sector's corporates) and the International Financial Institutions (including the International Monetary Fund -IMF- and/or the World Bank). But, giving an estimate point for the economic growth rate is always misleading and, usually, has some negative sides because it's read from different sides. That's why it seems better to determine a confidence interval for this rate and, for us, it will be a good idea to help to avoid some misleading interpretations, mainly those which are political. The aim of this paper is to provide a confidence interval for the economic growth rate. This is intended to avoid political speculation for the determination of economic growth in Morocco.

Keywords

Economic budget; forecast; point estimate; confidence interval

---

The growth of the economy is the concern of each government because it represents the degree of evolution of this country's wealth. This growth essentially reflects GDP growth, which is the value added's growth. So, to help economic operators making the right decisions in the right time, forecasts are made to save money, reduce the costs of an operation, have visibility on an uncertain future, … etc.

But, giving an estimate point for the economic growth rate is always misleading and, usually, has some negative sides because it's read from different sides. That's why it seems better to determine a confidence interval for this rate and, for us, it will be a good idea to help to avoid some misleading interpretations, mainly those which are political.

[4] The technical work of this paper is in progress. It's a part from a doctorate thesis in economics.

Who develops economic forecasts in Morocco?

In Morocco, several departments work out economic forecasts and publish figures which are sometimes far from each other. These departments are the Haut-Commissariat of Planning (HCP), the Ministry of Economy and Finance (MEF), Bank Al-Maghrib (BAM which is the Moroccan Central Bank), the Research Centers (including the Moroccan Center of the Conjoncture -or CMC- which is close to the private sector's corporates) and the International Financial Institutions (including the International Monetary Fund -IMF- and/or the World Bank).

The production of these forecasts by several departments is laudable, but sometimes it is confusing and misleading for economic agents. Some specialists speak about a "war of figures" which affects the credibility of the statistical system of our country. In this case, which department to trust: the MEF? the HCP? The BAM? the CMC? the IMF? or the World Bank? etc.

In the HCP, the economic forecasting process is as follows: towards the summer of the current year (year t), the HCP prepares its exploratory economic budget for the coming year (year t + 1). It submits this report to the Ministry of the Economy and Finance, which is supposed to take it into account when preparing the budget for the coming fiscal year (accompanied by an economic and financial report supporting the measures to be taken, taking into account the national and international contexts).

For example, for 2014, the HCP has already given a forecast of 2.4%. The government, meanwhile, drafted the finance law on the basis of growth rate of around 4.2%. For its part, the CMC forecast 2.7% as a maximum growth rate for GDP in Morocco. World Bank forecasts were estimated at 3.6%. Who is right and who is wrong? How to send signals of confidence to the national and foreign economic operators regarding the soundness of our economy?

This contrast between the figures announced by different departments reflects the lack of coordination and the gap between the departments responsible for defining and adopting economic policies (government departments) and those responsible for macroeconomic and forecasting studies, which are generally non-governmental departments and/or independent of government (such as the HCP, the CMC, the IMF and the World Bank). Indeed, reading the figure announced by the government, a number of questions arise: how the government was able to base its 2014 budget on the assumption of an ambitious economic growth rate of 4.2% while the economy's growth is driven by domestic demand, which was hampered by a wage freeze, a rise in prices following decompensation and a rainfall deficit marking the beginning of this agricultural year (internally) and by anchoring external demand essentially by partners who have not yet emerged from a global economic crisis? What will happen to this domestic demand face to a higher taxation (direct and / or indirect) and a withdrawal of several public

investments? Does the announcement of this 4.2% growth forecast in 2014 not require additional funding to be sought here or elsewhere? Does it not reflect an excess of confidence on the part of the government (this is reminiscent of the 7% growth rate announced by the ruling party during the legislative campaign of 2012)?

This difference in forecasts stems - in part - from the different assumptions adopted by each party, in addition to the ingredients and mathematical models used to obtain these forecasts. This difference in assumptions reflects the absence of an international consensus or even between the national experts of the same country on a given economic situation or a particular methodology. Moreover, no one can be exhaustive in determining the internal and / or external factors that can affect the economic growth of a country.

Do forecasts meet our needs? Are these forecasts so useful even if they do not meet our needs and are often far from reality? Do economists and/or policy makers not feel ridiculed when the announced forecasts differ from the realized numbers? Should we continue to produce these forecasts in the same way that we are currently developing them? These are questions that have emerged after examining the table that follows and which gives the forecasts of the growth rate in Morocco between 2004 and 2018:

Evolution of expected and realized economic growth rates in Morocco

Years	Forecast economic growth rate (in %)	Realized economic growth rate (in %)	Absolute error of forecast	Relative error of forecast (in %)
2004	3,3	4,8	-1,5	-0,3
2005	2,6	3	-0,4	-0,1
2006	5,2	7,8	-2,6	-0,3
2007	3	2,7	0,3	0,1
2008	6,1	5,6	0,5	0,1
2009	6,7	4,8	1,9	0,4
2010	4,1	3,6	0,5	0,1
2011	4,6	5	-0,4	-0,1
2012	2,4	2,7	-0,3	-0,1
2013	4,6	4,4	0,2	0,0
2014	2,5	2	0,5	0,3
2015	3,7	4,5	-0,8	-0,2
2016	2,6	1,1	1,5	1,4
2017	3,5	4,1	-0,6	-0,1
2018	2,9	3,1[5]	-0,2	-0,1

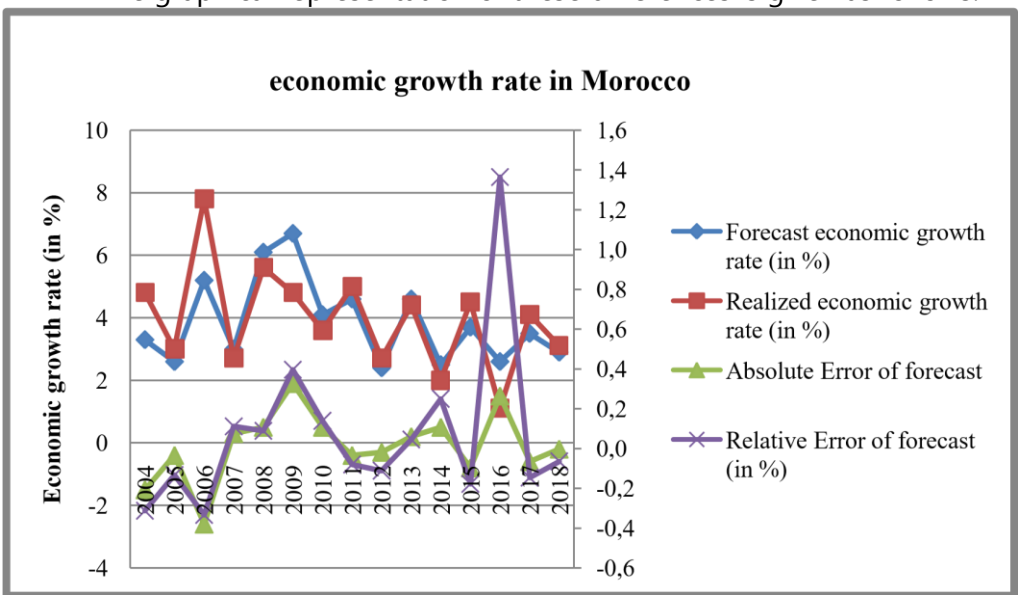
Source: Economic Budget Reports[6], HCP

[5] Provisional figure.

[6] The exploratory economic budgets prepared in June of the current year give the outlook for the economy for the coming year. These exploratory economic budgets are expected to be modified as part of the preparation of the projected economic budgets which are prepared in

This table shows that the forecasts prepared by the HCP are of good quality and close to reality. This is confirmed by the relative error which is very low since its extreme values are only of the order of 0.3% below (for the years 2004 and 2006) or 1.4% above (only for the year 2016).

The graphical representation of these differences is given as follows:



Absolute and relative errors committed in forecasting

It should be noted that the greatest differences were recorded for the years 2004, 2006 and 2016. Thus, for the year 2006, the estimated growth rate was 5.2% but the circumstances have been different since this year saw a good agricultural year and also had the positive effects of the Early Voluntary Retirement announced in 2005. In contrast, the growth rate forecast in 2016 was overestimated and it was "called back to the order" from 2.6 % to 1.1%. Does this difference in results reflect the limit of the forecast exercise?

Everyone knows that an economic forecast is obtained using a mathematical model linking the set of macroeconomic variables in equations. But, these ingredients (variables and equations linking these variables) differ from one organism to another according to one's own vision to explain an economic phenomenon to come and prepare the recipe to face it. Moreover, human behavior is difficult to model and predict because it is subject to a behavioral bias that is difficult to master and thus makes economic life very uncertain. This explains the panoply of results obtained to explain the same variable (the growth rate of GDP).

January of each year after the drafting of the Finance Law by the Ministry of Economy and Finance and its adoption by the Two-Chamber Moroccan Parliament.

Thus, despite this intellectual effort, we must not be too confident in forecasting models since they cannot incorporate "surprise" variables or events that may occur along the way. In addition, we need to involve everyone in the economic decision-making process to help obtain good forecasts - not subject to political speculation - in terms of both short-term forecasting and medium/long term economic projections. Finally, the forecasts must be developed in the form of an interval and avoid giving oneoff forecasts that have shown that they are never reached.

Theoretical considerations

After the previous introduction, the aim of this paper is to provide a confidence interval for the economic growth rate. This is intended to avoid political speculation for the determination of economic growth in Morocco.

The theoretical work is in progress. It will be based on some papers that deal with this issue and this paper is a part of a thesis in progress to obtain a Doctorate Degree.

References

1. ABDELKHALEK, Touhami & DUFOUR, Jean-Marie (2006). Confidence regions for calibrated parameters in computable general equilibrium models. Annales d'économie et de statistique; N° 81.
2. DAWKINS, C. & SRINIVASAN, T. N. & WHALLEY, J. (2001). Calibration.
3. In J. J. Heckman and E. E. Leamer (eds); Handbook of econometrics; volume 5; North-Holland; Amsterdam.
4. DUFOUR, J. M. (1989). Non-linear hypotheses, inequality restrictions and non-nested hypotheses: Exact simultaneous tests in linear regressions. Econometrica; N° 57; pp: 335-355.
5. DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. Annals of mathematical statistics; N° 28; pp: 181-187.
6. HARRISON, Glenn W. (1989). The sensitivity analysis of applied general equilibrium models: A comparison of methodologies. Technical report; Department of Economics; University of New Mexico.
7. Haut-Commissariat au Plan. Reports on "Economic budget". Different issues.
8. HOOVER, Kevin D. (1991). Calibration versus estimation: Standards of empirical assessment in the new classical macroeconomics. Research program in applied macroeconomics and macro policy; Working paper series N° 72; Institute of governmental affairs; University of California; Davis; January.
9. HWANG, Jiunn T. and ULLAH, Aman (1991). Confidence sets centered at James-Stein estimators-A surprise concerning the unknown variance case. Research report; N° 8909.

10. PAGAN, Adrian R. and SHANNON, J. (1985). Sensitivity analysis for linearized computable general equilibrium models. In PIGGOT, J. and WHALLEY, J. (eds): New developments in applied general equilibrium analysis; Cambridge University press; Cambridge; UK.
11. PAGAN, Adrian R. and SHANNON, J. (1989). How reliable are ORANI conclusions? Economic record; N° 63; pp: 33-45.

Reduced social accounting matrix for Mozambique

Eliza Mónica A. Magaua
National Institute of Statistics of Mozambique

Abstract

In order to measure the economic activity, it is usual to rely on the national accounts data, which in principle, follows the international recommendations written in the System of National Accounts, SNA. As a way to obtain one more instrument to analyse the economic activity of a country and also which allows for police impact analysis, the social accounting matrix, SAM, is the one that provides the answer. For this paper, we propose it in its reduced form, and the data is regarded the year 2016. The essence of a SAM in this work is to be found by the transactions and transfers between different institutional sectors.

Keywords

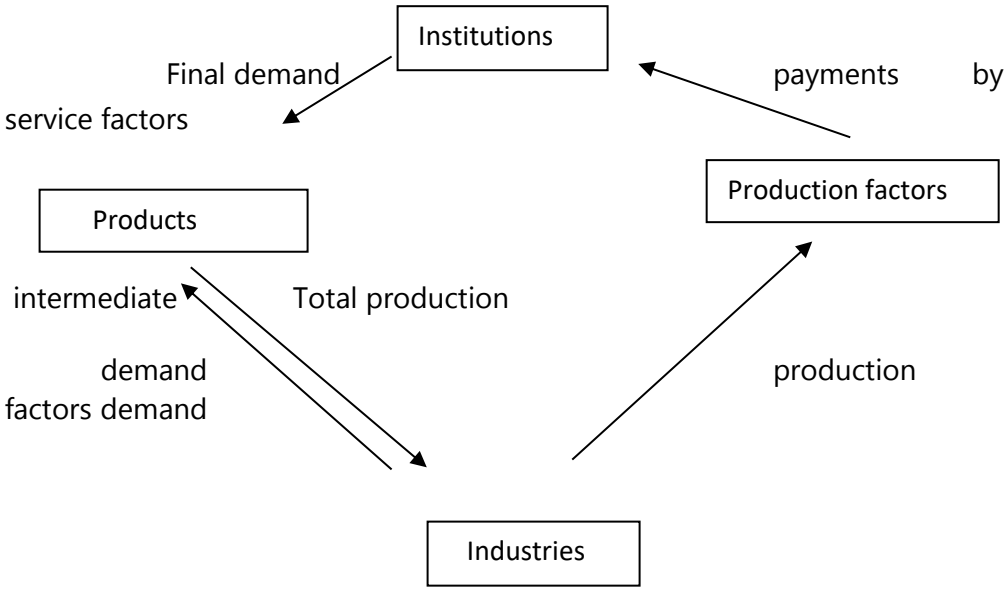
Economic activity; social accounting matrix; national accounts

1. Introduction

The social accounting matrix, SAM, is a matrix representation of the income circular flow. It captures all the real transfers among industries and institutional sectors namely, non-financial corporations, financial corporations, government, households and non-profit institutions serving households and its interaction with the rest of the world sector. With a SAM it is possible to distinguish the commodities, which in broad terms are called goods and services, from the industries, production factors and institutional sectors. A reduced SAM that we are proposing for Mozambique, will help in quality control in the national accounts, as long as they are relatively weak and incomplete. The designation reduced is due to the fact that it will be presented in aggregated terms; the details not going to be at the level of institutional units such as enterprises. The base for the construction of a SAM is the Economic Integrated Accounting that represents a sequence of economic accounts, starting from the current accounts ending with accumulation accounts. This sequence belongs to the standard SNA, is produced by the national accounts and allows for a comprehensive description of an economy emphasizing the distributives aspects. National accounts constitute a product of official statistics so that accuracy in its compilation is required. In fact, the economic activity, the main source of income and wealth is the origin of all flows that composes the network links that is established in the economy. It requires intermediate inputs that are used during the productive process in

order to obtain other goods and services and constitutes flow of funds for the market.

Figure 1. Income circular flow



source: Pyatt (1991)

Main objective

To construct a reduced social accounting matrix

Specific objectives

- To develop a social accounting matrix for the year 2016
- To provide a formal demonstration of similarities and differences between data systems
- To underline the informal activities in official statistics

2. Revision of literature

Recent studies suggest that the traditional national accounts, are insufficient to better describe the economic performance and wellbeing of population. There is a growing dissatisfaction with the national accounting practice that deals meanly with aspects related to economic growth. For the case of Mozambique, and other less developed countries, the exclusive concentration on the economic growth it is due in part by the lack of good basic statistics, and on the other side, the lack of better instruments for cross checking and consequently quality control. In the paste decades, the economic world is suffering from changes due to globalization, changes in the characteristics describing wellbeing, among others. In order to follow these

changes, the performance indicators have to achieve more dimensions of analysis as a way to provide better insights for users. According to Stiglitz, Sen and Fitoussi (2008), Report by the Commission on the measurement of economic performance, there are some weaknesses and limitations of GDP, so that it is necessary to look at other dimensions of human development such as well-being, environment and sustainability, among others. In this work, we are not trying to fulfill the majority of those concerns, but would like to contribute with an additional way to describe the economy, which not only looks at economic performance measured by the real Gross Domestic Product, GDP, but also presents some social economic aspects of the economy, via interactions among institutional sectors and that illustrates the distribution and redistribution of income. According to Pyatt G. (1991), a SAM provides a straightforward way of capturing the details of the income circular flow. In supporting this sentence, it is important to refer that other systems like for example the T-accounts cannot provide what is happening within the economy. Traditionally, policy analysis makes use of statistics in order to take decisions. The UNSD handbook on the Use of Macro Indicators for policy analysis, emphasizes the role of macro accounting as an instrument rather than as a data set. With this insight, we feel that it is important not to rely only on real GDP for decision taken but also on SAM, since we can derive from it, several macroeconomic indicators such as gross national product, GNP, disposable income, YD, savings, operating surplus among others. According to Santos, S. (2017), is proposed as a framework for the study of the activity of a country. This activity which she considers socioeconomic activity involves monetary flows measured by the national accounts. One way to expose the potentialities of a SAM, the author suggests the possibility of constructing networks for the links of the above-mentioned flows. Supporting the above-mentioned author's ideas, we elected a SAM as an additional instrument for analyzing the economic activity with focus in distributive aspects.

3. Methodology

The methodology we are going to use in this project starts with the literature revision, as a mean to support the use of macro indicators in this instrument. The following step is to present the first picture of the reduced SAM and describe in broad terms all the transactions that are taking place among the institutional sectors within and between this and the rest of the world. As we previously mentioned, for the construction of a SAM there is an interaction of many statistics. This interaction provides the possibility to detect data inconsistency, as long as the national accounts in Mozambique, although are being produced for more than two decades, are not very well developed. Thus, the identification of data inconsistency will result in upgrade the quality of official data. Consequently, we consider this project relevant as long as with

an updated SAM, some aspects not very well covered in the traditional accounts are taking into account, and in this manner we can contribute for the improvement of different aspects of socioeconomic activity in the country. Additionally, this instrument can support for future research in related areas.

Table 1. SAM in big categories

		Payments for:				
		Activities	Products	Production Factors	Institutions	Total
Revenue received by:	Activities		Production of goods and services (sales)			
	Products	Intermediate goods			Final Consumption	
	Production factors	Value added generated in the productive process				
	Institutions			Distribution of income factos to the institutions		
	Total					

The author based on the Pyatt, G. and Round, J. (1985). A detailed description of the cell will be found in the annex of this paper, which will be found in the complete version of this proposal. In the above table, one can see that SAM distinguishes between activities which are entities that do production and products which represents the market product. The payments done by the activities are done for the goods and services accounts. These consist of products produced either domestically or imported. Under the category of institutions, we can find the non-financial corporations, financial corporations, households, non-profit institutions serving households and public administration, which for the purposes of this paper we are not going to separate the central government from the local government. Additionally, in the table we could understand that SAM contains a lot of production factors that earn incomes from its use in the production process and, pay revenues to the enterprises, households, government and the rest of the world.

References

1. Pyatt, G. (1991). Fundamentals of Social Accounting. Economic System Research, Vol.3, No.3,1991
2. Santos, S. (2017). "Identificação e Construção de Cenários Macroeconómicos para o Estudo de Impactos de Medidas de Política Económica: Uma abordagem matricial com simulação a Moçambique". Working Paper 163/2017/CESA (Centre for African, Asian and Latin American Studies) /CSG (Research in Social Sciences & Management), – ISEG (School of Economics and Management) / Universidade de Lisboa
3. Stiglitz, E., Sen, A. and Fitoussi, J. (2009). "Report by the commission on the Measurement of Economic Performance and Social Progress". Commission on the Measurement of Economic Performance and Social Progress.
4. United Nations Statistics Division. Use of Macro Accounts in Policy Analysis. Studies in methods series F no.81

Data mining of mobility table Based on community discovery methods

Xu Sun[1], Xiao-hui Li[2]
[1]School of statistics, Dongbei University of Finance and Economics, Dalian, China
[2]College of public administration and humanities, Dalian Maritime University, Dalian, China.

Abstract

Based on community discovery methods, a new approach to the modeling social mobility data is presented. Community detection algorithm to identify communities of social classes within which social classes share members at above expected rates. This approach, when applied to mobility data, may be used to substantially improve the fit of models of social mobility. To illustrate, the community effect model of social mobility is analyzed using data from the General Social Survey.

Keywords

Intergenerational mobility tables; Log-linear model; Community detection; Eigenspectrum decomposition

1. Introduction

Intergenerational mobility is an important perspective of social mobility analysis, and have a variety of log-linear at their disposal with which to analyze the structures and patterns embedded within mobility tables (e.g.,Hout, 1983). In empirical analysis, in many cases, the structure in mobility tables is so sufficiently complicated that parsimonious models do not capture the observed patterns. In such circumstances, there are ever more complicated models that may be fit to the data. For example, Moses & Holland (2010) compared 12 statistical strategies which included significance tests based on four chi-squared statistics proposed for selecting log-linear models. Tibshirani (2011) proposed Lasso (Least aboslute shrinkage and selection operator) method for estimation in generalized regression model. Yuan et al. (2011) purposed an automatic data mining method of contingence table based on multinomial processing tree model. Likewise, if a preferred model (e.g., quasi-symmetry) does not fit the data, one can estimate a correspondence analysis on the residuals to "see" the associations left over in the data (Falguerolles & Leeuw, 1989). Melamed (2015) drawn on the idea that mining the residuals and uses community detection methods to "see" the associations left over in the data. These methods provided a good-fitting log-linear model, however, the results are often particularly complicated and an understanding of the underlying mobility processes may be obscured by the complexity of the model.

In this paper, a novel automatic log-linear modeling method is introduced for understanding and fitting mobility processes. Eigenspectrum decomposition community detection approach (Newman, 2010) is used in association analysis. By thinking about class categories in a mobility table as nodes, and people as the weighted relations between them, novel insights are drawn that illuminate the association between rows and columns of social mobility tables. A community detection analysis identifies which categories share members at rates above chance, and which should belong together. First of all, community membership may be included in the "independence" model without interaction. Subsequently, if the log-linear model include community membership is ill-fitting, then the second community discovery was initiated to remain after the log-linear model has been estimated. If data still is ill-fitting, then implement the third and fourth times until the good-fitting is achieved. The results from the community detection analysis parsimoniously inform the analyst of associations that remain after the log-linear model has been estimated.

The approach described above offers a couple of advantages relative to other similar methods. Falguerolles & Leeuw (1989) and Melamed (2015) analysis of the residuals from an ill-fitting log-linear model, suffers from uncertainty of the results. Because the different initial model of mobility processes (e.g., quasi-independence, quasi-symmetry, unidimensional social distance etc.) which does not fit the data will affect the final modeling results. The modeling results of Moses & Holland (2010) or Tibshirani (2011) or Yuan et al. (2011) are particularly complicated to understanding the underlying mobility processes. The strength of the novel modeling method that is detailed below is that an objective function is maximized to identify the "best" way to combine categories, and subsequently a single within-community term may improve any log-linear model fit. The results are therefore certainty, clear, automatic, and relatively straightforward.

Below, this new approach to interpreting social mobility that draws from recent advances in community detection algorithm in social networks. Then the approach is applied to social mobility tables that were derived from the General Social Survey (GSS; Smith, Marsden, Hout, &Kim, 2005). The community structures of multiple models of social mobility respectively is identified for female, male, and all respondents, the results of which reveal interesting substantive findings. Last, the proper model for female, male, and all respondents respectively are provided.

2. Communities of intergenerational mobility tables

Within network science, a mode refers to a set of objects for which relations may be measured. A person-to-person network is a one-mode network, while a person-to-groups network is a two-mode network. An intergenerational

mobility table may be understood to be a two-mode network, where one mode is parental social class and the other mode is respondent social class. What are shared between these two modes are members, or the count of people who have a given social class and have parents of a given social class. It is often assumed that the same social classes are represented for both parents and respondents, which implies that both modes share the same number of nodes, or in this case, categories of social class. Newman (2010) provided an approach for the analysis of social networks is the identification of communities or cohesive subgroups. Here, a community refers to a subset of nodes which share relations at above expected rates. Community detection has a rich history in computer science and the social sciences (Wasserman & Faust, 1994), but until Girvan and Newman (2002) this problem is brought to the attention of the general scientific community.

In this paper, We use eigenspectrum decomposition approach Newman's (2006a,b) because it is easily applicable to intergenerational mobility tables, has been generalized to multi-mode networks, such as mobility tables, and is highly efficient and accurate relative to other solutions to the community finding problem. Newman's (2006a) eigenspectrum approach is elegantly simple. One begins with a relational matrix that is denoted by *A*, defines a matrix of expected cell counts that is denoted by *P* (an "independence" model), subtracts *P* from *A* to yield *B*, which is called the residuals matrix, and finally one computes the eigenspectrum decomposition of the residuals matrix. The eigenspectrum of the residual matrix, *B*, sheds light on the structure of *A* (Newman, 2006a,b). Newman has shown that the signs of the entries in the eigenvector associated with the largest eigenvalue partition the nodes into an optimal two community split. Subsequent splits into more than two communities may be determined by examining the signs of the entries in the second leading eigenvector, and so on.

Another development with respect to community structures was to define the quality function that is unfortunately also called modularity Newman &Girvan (2004). Modularity (denoted by *Q*) indicates the strength of, or variance explained by, a community structure discovered by the community finding algorithms. That is, it provides a bench-mark with which to compare possible solutions to the community structure. Larger values of *Q* indicate that larger shares of the relations in the data are within communities; hence larger values indicate a better fit to the data. The specific formula for modularity was generalized for the eigenspectrum approach by Newman (2006b). Thus, the eigenspectrum decomposition of the residual matrix, *B*, can be used to identify possible solutions to the community structure, and the quality function modularity can be used to identify which solution is "best".

The formula for the modularity function is the intuitive. Define a number-of-communities by number-of-communities matrix, which is denoted *E*. The

jj,th entry in *E* is the proportion of ties in the network that go from the community in row *i* to the community in column *j*. In this case, each entry is the proportion of respondents in the *jj*,th cell. The diagonal elements in *E* define the share of within-community ties. To compare this to the overall distribution of both within- and between-community ties, the modularity function is defined as $Q = Tr(E) - ||E^2||$, where *Tr*( ) is the trace of matrix *E* and || || indicates the sum of the elements of matrix.

Intergenerational mobility tables are typically square because they have the same social classes for parents and respondents. As such, the table is amenable to spectral partitioning as discussed above. There is, however, a problem that is the resulting community structure would only row (partition parental social class) because it does not recognize column (respondent social class) is a distinct mode of the data. Fortunately, standard multi-mode generalizations allow row and column (parental social class and respondent social class) to be dually represented in the same community partition. To do so, one simply includes the mobility table and its transpose in a larger block off-diagonal matrix (Wasserman & Faust, 1994). Denote a mobility table as M. The spectral decomposition of the following matrix would allow the rows and the columns to be dually represented in sub-sequent community partitions: $Z = \begin{matrix} 0 & M \\ M & 0 \end{matrix}$. Here M^T is notation for the transpose of matrix M and 0 is notation for a matrix of zeroes.

When the genspectrum decomposition is applied to typical social networks (e.g., person-to-person networks), the expected network is invariably an "independence" model; that is, each entry in the expected network is the row sum times the column sum divided by the number of relations in the network. More generally, the approach taken herein is to remove the effects of "independence" model of mobility compute the community structure of residual variation, and determine the effects of community membership on social mobility patterns by parameterizing a within-community effect in the design or "model" matrix. In every empirical case examined, the effect of community membership is quite large while being very parsimonious (i.e., 1 degree of freedom).

3. Data, modeling and results

We use data which the GSS collected information on father's occupational status-1988-2010, and thehe sample space was restricted to respondents who were ages 25-64 and were in the labor force. Melamed (2015) used this data, and 13746 respondents (include 6864 female respondents and 6919 male respondents). Respondent and parental occupational classifications were recorded in 1980 census codes by the GSS, and were then transformed into social classes using a standard measure (Erickson & Goldthorpe, 1992).

Table 1. Social class positions used in the analyses

I	Professionals, administrators, officials, and managers, higher level
II	Professionals, administrators, officials, and managers, lower level
III	Routine non-manual and service workers, higher and lower levels
IV	Self-Employed, with or without employees
V	Technical specialists and supervisors of manual workers, skilled manual workers
VI	Semiskilled and unskilled manual workers, nonfarm and farm

Table 1. illustrates the six class positions, and Table 2. presents the social mobility tables that are analyzed in this paper. Below models and community structures will be presented for female, male, and all respondents.

Table 2. Social mobility tables

	Female respondents						Male respondents						All respondents					
	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
I	245	405	196	142	28	89	404	241	42	199	144	135	245	405	196	142	28	89
II	133	257	148	84	12	58	176	193	36	118	105	135	133	257	148	84	12	58
III	24	64	45	22	12	30	37	36	19	22	37	35	24	64	45	22	12	30
IV	260	524	358	272	72	243	314	290	68	493	281	345	260	524	358	272	72	243
V	213	416	388	139	47	208	209	261	81	196	409	331	213	416	388	139	47	208
VI	207	458	466	126	82	391	216	216	88	177	319	511	207	458	466	126	82	391

Beginning with the female respondents, the community structure analysis proceeds as follows: first, Afrom above is defined as the observed mobility table, and then P is defined as the fitted values from an "independence" model that are estimated. The matrix B is defined as A-P, then B is compiled into the block-off-diagonal matrix Z. Following Newman (2006a), taking the eigenspectrum decomposition of Z sheds light on the community structure of A net of P. Specifically, the signs of the entries in the eigenvector associated with the largest eigenvalue indicate the optimal two community split, and subsequent eigenvectors may similarly be used to determine more than two communities. Using the quality function modularity (Q), the locally defined

optimal community solution may be obtained based on any number of leading eigenvectors (see Table 3).

Table 3. Three community solution for female respondent

The first community detection			The second community detection			The third community detection		
Number	structure	Q	Number	structure	Q	Number	structure	Q
2	$+_s$-$+_s$-	0.0100	2	$+_s$-$+_s$-	0.0062	2	$+_s$-$+_s$-	0.0034
3	$+_s$-$+_s$--$+_s$-	0.0100	3	$+_s$-$+_s$--$+_s$-	0.0062	3	$+_s$-$+_s$--$+_s$-	0.0032
	$++_s$+-$_s$-	0.0137		$++_s$+-$_s$-	0.0066		$++_s$+-$_s$-	0.0036
4	$+_s$-$++_s$-+-$_s$-$+_s$-	0.0084	4	$+_s$-$++_s$-+-$_s$-$+_s$-	0.0062	4	$+_s$-$++_s$-+-$_s$-$+_s$-	0.0032
	$+_s$-$+_s$--$+_s$-$+_s$-	0.0100		$+_s$-$+_s$--$+_s$-$+_s$-	0.0062		$+_s$-$+_s$--$+_s$-$+_s$-	0.0032

From the community structure, if the respondent and her father's social class is in the same community, the interaction between them is considered to exist, then the design matrix for the community effect is defined (see Table 4.)

Table 4. The design matrix for the community effect and the estimate value

	λ_{ij}^{RC1}	λ_{ij}^{RC2}	λ_{ij}^{RC3}
design matrix	$\begin{bmatrix}1&0&0&0&0\\1&0&0&0&0\\0&1&0&1&1\\0&0&1&0&0\\0&1&0&1&1\\0&1&0&1&1\end{bmatrix}$	$\begin{bmatrix}0&1&1&0&0\\0&1&1&0&0\\1&0&0&0&0\\0&0&0&1&1\\1&0&0&0&0\\0&0&0&1&1\end{bmatrix}$	$\begin{bmatrix}0&0&1&1&0\\0&0&1&1&0\\0&0&1&1&0\\0&1&0&0&0\\0&0&1&1&0\\1&0&0&0&1\end{bmatrix}$
Parameter estimate value	0.5149	0.2608	0.1122

The model with community Effects is parsimonious, the community effect only used 1 degree of freedom, and the DF the new model is 24. Meanwhile, $\cdot^2$ and L^2 substantial decline, it implied that the community structure maybe implemented to improve independence model fit (see Table 5.). But, in one case, including the community effect in the log-linear model does not result in a model that fits the observed mobility patterns. So, the second community detection for the the new residual matrix B, B is defined as A-P, A is still the observed mobility table, and then P is defined as the fitted values from the model with the first community Effects that are estimated. After the second community detection, the model fit is improve a great deal, but the log-linear model with two community effect parameters is ill-fitting for the observed mobility table. So the third community detection is implemented.

After three community detection, the good-fitting model is got.

Table 5. Model fit statistics for the log-linear models (Results for female respondents)

number of interaction parameters in log-linear model	x^2	p-value	L^2	p-value	DF
0	331.3807	0.0000	332.7932	0.0000	25
1	116.1002	0.0000	115.5011	0.0000	24
2	47.1577	0.0021	56.6776	0.0025	23
3	30.5735	0.1052	30.5982	0.1048	22

4. Discussion and conclusion

For male respondents, modeling use three times community detection and for all respondents, modeling use five times community detection. These models only include 3 or 5 interaction parameters, is parsimonious. The strength of the novel modeling method that is detailed below is that an objective function is maximized to identify the "best" way to combine categories, and subsequently a single within-community term may improve any log-linear model fit. The results are therefore certainty, automatic, straightforward, and relatively clear.

By thinking about a mobility table as a network, and introduced a community detection algorithm, improvement in model fit have been demonstrated. The use of the eigenspectrum decomposition for community detection has a few benefits (eg. Newman(2006a) pointed out the eigenspectrum approach is efficient, and is also relatively accurate). But other similar algorithms also could be used in the method detailed above. Such as partitional clustering (Porter et al., 2007), or k-cliquebased approaches (Palla et al., 2005), may yield solutions that are just as useful as the eigenspectrum decomposition.

References

1. Erickson, R., & Goldthorpe, J. H. (1992). The constant flux: A study of class mobility in industrial societies. Cambridge: Oxford University Press.
2. Falguerolles A D, Leeuw J D. (1989). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Log-Linear Analysis. Journal of the Royal Statistical Society, 38(2):249-292.
3. Girvan M , Newman M E J . (2001). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821-7826.
4. Hout M . (1983). Mobility tables. Beverly Hills Calif.
5. Melamed, David. (2015). Communities of classes: A network approach to social mobility. Research in Social Stratification and Mobility, 2015, 41:56-65.
6. Moses T, Holland P W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. British Journal of Mathematical & Statistical Psychology, 63(3):557.
7. Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113.
8. Newman, M. E. J. (2006a). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103(23), 8577–8582.
9. Newman, M. E J.(2006b). Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74(3):036104.
10. Porter, M. A., Mucha, P. J., Newman, M. E. J., & Friend, A. J. (2007). Community structure in the United States House of representatives. Physica A, 386, 414–438.
11. Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlap-ping community structure of complex networks in nature and society. Nature, 435(7043), 814–818.
12. Tibshirani R. (2011). Regression shrinkage and selection via the lasso: a retrospective . Journal of the Royal Statistical Society, 73(3):273-282.
13. Wasserman, S., Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge.
14. Yuan Y, Huan Qi, Xiang-En Hu. (2011). Data Mining of Contingence Table Based on Multinomial Processing Tree Model. Computer Engineering, 37(11):10-12.

Integration of statistics on gender and sustainable development through capability building

Josefina V. Almeda, Ana Julia J. Macaraig
Philippine Statistical Research and Training Institute
Quezon City, Philippines

Abstract

The 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs) adopted in 2015, envisions a progressive and sustainable world that leaves no one behind. In all the 17 goals, the aim to attain gender equality and women empowerment is essential. The integration of statistics on Gender and Sustainable Development is part of the capacity building program of the Philippine Statistical Research and Training Institute (PSRTI). As the mandated agency tasked in the Philippines to conduct statistical research and training, the PSRTI is expected to continuously upgrade the quality of statistical personnel and expand the manpower base that will undertake statistical work and contribute to the improvement of statistical activities. The capability building on Statistics for Gender and Sustainable Development among personnel in the Philippine Statistical System (PSS) and the public make use of a training strategy that expound the importance of appreciating basic statistics when dealing with SDG indicators, with particular focus on SDG Goal 5: Achieve gender equality and empower all women and girls. The capability building programs also attempt to emphasize that gender equality should cut across all persons - women and girls, men and boys - because by definition, gender pertains to characteristics and perception between masculinity and femininity, these include biological sex, sex-based social structures, or gender identity. In ensuring the rights of all persons across all the goals, policies and decisions have to be evidenced-based, complete with statistics that are of high quality and impeccable integrity.

Keywords

sustainable development; gender statistics; gender equality; capability building

1. Introduction

The Sustainable Development Goals (SDGs), otherwise known as the Global Goals, are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity, that no one will be left behind. There are 17 Goals that resulted on the successes of the Millennium Development Goals, while including new areas such as climate change, economic inequality, innovation, sustainable consumption, peace and justice,

among other priorities. These goals are interconnected, where the key to success of one will involve tackling issues more commonly associated with another. It is on this premise that the goal of attaining gender equality and women empowerment can be associated with. Although there may be that prevailing idea that the topic of gender, whether on equality or its development, is exclusively for women and girls.

Gender is a whole lot more. It is not only for equality for women or girls; it is also for the men or boys as well. The need for high quality statistics that will echo the basis for policy and decision-making are important. Equally important is the integration of statistics in capacity building programs that are well understood and easily appreciated.

The Philippine Statistical Research and Training Institute (PSRTI) is the mandated agency tasked to conduct statistical research and capability building endeavors, including those topics dealing on Statistics for Gender and Sustainable Development. The clientele or targeted participants of these trainings are personnel in the Philippine Statistical System (PSS) and the public. The PSS is composed of national government agencies that deal with statistics/statistical data, either as a data producer or data user. Its main task is to deliver quality statistical information to the public.

The PSRTI is expected to continuously upgrade the quality of statistical personnel and expand the manpower base that will undertake statistical work and contribute to the improvement of statistical activities. Every year, the agency offers regular training courses on Statistics from the most fundamental to the more advanced tools.

In developing the training materials on Statistics for Gender and Sustainable Development, PSRTI collaborated with statistical experts, data producers, and data users. The purpose of the training is to teach concepts and methods encompassing social concerns regarding gender issues, and link them to the technical and methodological aspects of statistics production and use. Specifically, the training objectives are (i) awareness of the Sustainable Development Goals with focus on Gender Equality; (ii) incorporation of gender perspective in all stages of the data production process; (iii) identification of gender issues (iv) integration of statistics on gender and sustainable development; and (v) formulation of policies.

The capability-building course on Statistics for Gender and Sustainable Development is a 5-day training from 8:00 a.m. to 5:00 p.m. (40 hours). The target-training participants are policy-makers, planners, gender experts, the public, national and international development agencies, NGOs, research institutes, and media practitioners. Knowledge of basic computer operations is required and data management using MS Excel® is a pre-requisite to this course.

2. Methodology: Development of Training Materials

Based on the inputs of experts, data producers, and data users, the course outline for the capability building on Statistics and Gender and Sustainable Development covers the following:

I. Introduction
- 1.1 Sustainable Development Goals with Focus on Gender Equality
- 1.2 Key Concepts of Gender, Sex, Statistics, Variables, and Indicators

II. The Production Process of Gender Statistics
- 2.1 Identifying Gender Issues
- 2.2 Identifying Gender Indicators
- 2.3 Assessing Availability of Gender Indicators and Statistics
- 2.4 Presenting Available Gender Indicators and Statistics - Engendering Presentation and Interpretation of Indicators and Statistics
- 2.5 Modifying Existing Surveys and Administrative Reporting Systems to Make them Gender Responsive
- 2.6 How Surveys are Designed and Planned

III. Sectoral Gender Statistics

Lecture notes are in power point presentations. Each presentation begins with the learning objectives of a particular topic. It also includes the use of relevant animated presentations with high quality graphics interchange format (gif) images from Presenter Media that are free to download and use. These animated images are meticulously chosen to fit or be appropriate to the given topic or discussion set for the day.

Examples of free GIFs from Presenter Media:


Gender equality | United Nations | Investigation | Education

Before presenting the actual notes, a quotation that is appropriate on the topics of gender and statistics is given. Such quotations by famous personalities contribute to getting the attention of participants and make the topic more interesting and relevant. Here are some good quotations on statistics:

"Statistics is the grammar of science." by *Karl Pearson*

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." by *H.G.Wells*

"Statistics are no substitute for judgment." by *Henry Clay*

"It is a capital mistake to theorize before one has data." by *Sir Arthur Conan Doyle*

On gender equality, here are some good quotations:

"Gender equality is more than a goal in itself. It is a precondition for meeting the challenge of reducing poverty, promoting sustainable development and building good governance." by *Kofi Annan*

"Any serious shift towards more sustainable societies has to include gender equality." by *Helen Clark*

"It is time that we all see gender as a spectrum instead of two sets of opposing ideals." by *Emma Watson*

"No country can truly develop if half of its population is left behind." by *Justine Greening*

There is clear distinction of important concepts like gender and sex by providing definitions, graphics, and illustrations. Sex is the biological and physiological characteristics that define women and men. It is permanent and the categories are Female-Male. On the other hand, gender refers to roles, behavior, and activities assigned by society to women and men and boys and girls. It can change over time with changes in culture and others categorize it as Feminine – Masculine or Lesbian, Gay, Bisexual, Transgender, and Queer.

The common notion about the term 'gender statistics' surfaced during capability trainings. Many participants think that these only refer to numerical figures about women. The purpose of distinguishing the terms gender and sex is to stress that policy and research interest is usually in gender, not sex, **but** examination of data by sex, is the means to making gender-based analyses. Cross-classification of data by sex, presents information separately for men and women, boys and girls. In addition, sex-disaggregated data reflect roles, real situations, general conditions of women and men, girls and boys in every aspect of society.

The lecture presents the different SDGs with gender-related indicators. To appreciate the gender-related sustainable development goal indicators, the lecture provides actual data using infographics that are meant to explain the information simply and clearly.

Examples of Gender-related indicators data:

For Goal 1: End poverty in all its forms everywhere

- 4.4 million more women than men live in extreme poverty globally

- Women aged 25-34 are more vulnerable to poverty because of the struggles of combining childcare and other unpaid work without earning an income.
- In all countries, single mothers are more likely to live in poverty than others.
- More women than men live on less than USD1.90 per day.

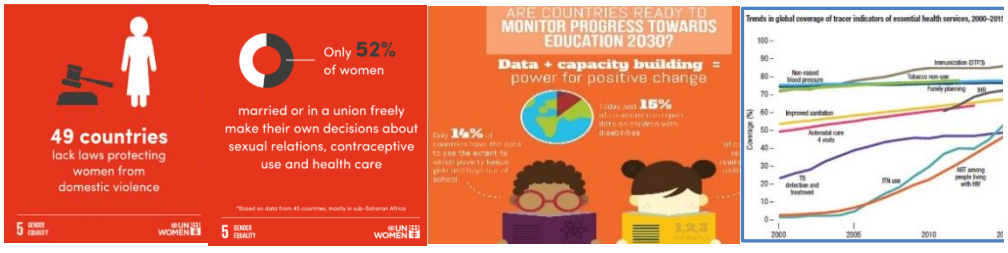
For Goal 3: Ensure healthy lives and promote well-being for all at all ages

- In 2015, there were 216 deaths per 100,000 live births (a drop of 44% from 1990); however, in 2013, over 40% of all pregnant women were still not receiving early antenatal care
- Incidence rate of tuberculosis have declined by 19% over the 16-year period from 2000 – 2016. However, while progress is impressive, it is still not fast enough to close persistent gaps and drug-resistant TB is a continuing threat.

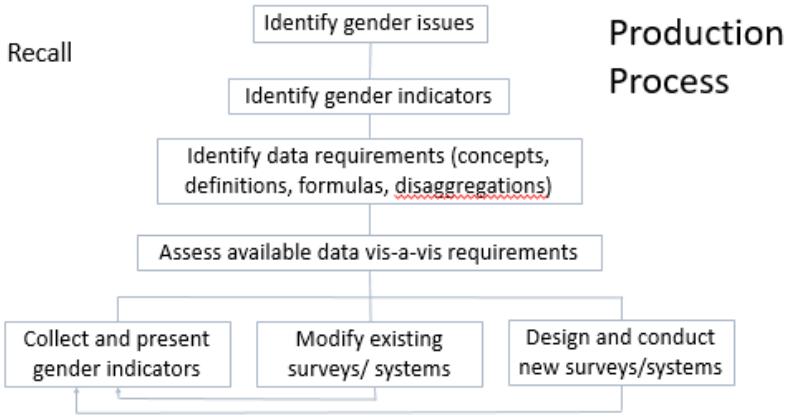
For Goal 5: Achieve gender equality and empower all women and girls

- Every year, 15 million girls under the age of 18 are still forced into marriage.
- One in 5 women and girls age 15-49 globally report that within the last 12 months, they experienced physical or sexual violence at the hands of an intimate partner.

Examples of Gender-related indicators on Infographics:



There are notes on the roles of women and men in the society, the difference between production and reproduction, the gender allocation of resources (access vs control), how gender needs are addressed (practical vs strategic), the gender equality points of views, and the gender issues.



The lecture includes the production process of gender-related indicators as given in the flowchart below.

Before presenting the topic on identifying gender indicators, the term 'variable' is defined and further enhanced by its two types: qualitative and quantitative variables. This is followed by the discussion of the different levels of measuring the variable namely, nominal, ordinal, interval, and ratio. The term 'indicator' is likewise defined, with several SDG indicators used as illustrative examples. The indicator may come in the form of ratio, rates, proportions, percentages, mean, and median.

Discussion of the different methods of data collection include surveys (personal interview, self-administered questionnaire, and time-use surveys), use of administrative data, and registration. For each method, the advantages and disadvantages are given.

On the topic of gender analysis, it is explained that this is the systematic collection and examination of information on gender differences and social relations in order to identify, understand and address inequalities based on gender. There is discussion of the different summary measures like rates, ratios, proportions, percentages, mean, median, standard deviation, coefficient of variation, and skewness. Each formula has an explanation but the emphasis is on the properties of the summary measures and interpretation of results. It distinguishes proportions from percentages since most of the SDG indicator formulas are multiplied by 100 but are called proportions and not percentages. There is provision of examples of MS Excel outputs calculating the different summary measures.

The last topic dwells on the different guidelines for presenting gender statistics. It begins with organizing the data by constructing the frequency distribution and histogram using MS Excel. Focus is on knowing the shape of the data set and its implications. There is discussion of several types of charts like line, vertical and horizontal bars, pie, pictograph, and maps. Emphasis is on presentation guidelines, focus of the chart, and data requirement.

3. Conduct of the Capability Training

Each capability-training begins with levelling of training expectations on the course, resource speaker, and co-participants. A pre-test follows to determine the level of knowledge of the participants on the course coverage. It is a 25 item multiple-choice type of exam with four possible answers. No participant's name appears on the exam paper but codes are used instead, for post-test matching purposes.

Each participant has a desktop or a laptop for his/her own referencing/viewing of the lecture notes and workshop activities. There is also provision of flash drives containing pdf copies of all the lectures. Training modes are lecture, discussion, workshop, presentation, and commenting on. For each main topic, participants work on exercises or workshops. There are hands-on exercises using MS Excel in the computation of summary measures and construction of charts. For the workshops, the grouping of the participants is according to offices/agencies or by sector. Each group need to identify a gender issue, identify data sources and indicators, compute for some summary measures, construct charts, and finally crafting a possible policy addressing the gender issue.

The final part of the training is the workshop presentation of each group. The resource person and another expert comment on the output after each presentation.

4. Training Evaluation and Next Steps

The training ends with a post-test and an evaluation of the training course, resource persons, facilities, and food. Each participant receives a certificate of completion if he or she has attended 90% of the total training hours. Otherwise, only a certificate of attendance is received.

The PSRTI had conducted three capacity building on Statistics for Gender and Sustainable Development in 2017, and another one for 2018. A total of 80 participants attended these trainings, with 30 men and 50 women. The agencies and sectors they represent are diverse, ranging from government agencies, academe, private sector, local government units, and non-government organizations. The level of satisfaction in the training program, resource person, and over-all conduct of each capacity-building course has a rating ranging from satisfactory to very satisfactory. There were recommendations on lengthening the time allocation for workshops as most participants find the exchange of opinions and ideas relevant to their own situations and challenges they face in terms of gender discrimination and inequalities.

The results of test scores on the pre-test and post-test were significant at 0.05 level of significance. The post-test scores were relatively higher than the

pre-test scores which implies that knowledge in gender and statistics improved.

Development and gender equality is fundamental to delivering on the promises of sustainability, peace and human progress. Achieving gender equality and women's empowerment is integral to each of the 17 goals. Only by ensuring the rights of women and girls, and in the case of the capacity building conducted by the PSRTI, the rights of all persons, across all the goals will mean equal justice and inclusion, economies that work for all, and sustaining a shared environment now and for future generations.

PSRTI intends to align itself with the achievement of the global goals by pursuing more capability building initiatives not only on Statistics on Gender and Sustainable Development but also in terms of inclusion of gender statistics in the different training modules being conducted by PSRTI. The focus on statistics that are well presented and appreciated may give rise to better gender-sensitive policies and programs. It will also be helpful in the monitoring and evaluation stages of indicators dealing with sustainable development and gender equality. The numbers will tell the stories, these will point to the next steps, to the assurance that gender equality is possible because the integrated statistics are appreciated and understood.

References

1. *Statistics for Gender and Sustainable Development* (2017). Philippine Statistical Research and Training Institute
2. *Harmonized Gender and Development Guidelines for Project Development, Implementation, Monitoring, and Evaluation* (2010). National Economic and Development Authority Philippine Commission on Women Official Development Assistance Gender and Development Network
3. *The Harmonized Gender and Development Guidelines (HGDG)* (2003). National GAD Resource Program Philippine Commission on Women National GAD Resource Program
4. http://www.unwomen.org

Estimation using probability proportional to aggregate size sampling in heterogeneous populations

Daniel David M. Pamplona
University of the Philippines, Philippines

Abstract

Estimation using Probability Proportional to Aggregate Size (PPAS) is compared with traditional design-unbiased techniques under different population scenarios. The study considers both standard error and relative bias of total estimates for comparison. Heterogenous populations were simulated by exploring varying behaviours of an auxiliary variable and its relationship with the target variable. Results show that the optimality of PPAS estimates improve as the linear association between the target variable and auxiliary variable increase. Furthermore, PPAS estimates are more stable under large variability in population.

Keywords

sampling rate; covariate effect; model fit; auxiliary variable; standard error; absolute percentage error

1. Introduction

Estimation methods for the total in survey sampling have developed over the years. Among these methods are design-unbiased and model-assisted techniques. Design-unbiased methods generate estimates based on the sampling distribution induced by the sample selection procedure. In other words, the method of sample selection determines the confidence in the estimates produced. This method, however, works best only when the sampling procedure has been religiously implemented, which in most cases, pose a challenge due to many practical reasons such as: unavailability of respondents, logistical limitations, absence of population frame, etc. Model-assisted estimation is a procedure of generating estimators with an aid of a model, usually in linear form. Inferences made about the population is still based on the sampling method used, but the estimation still works even if the model does not fit the data well. Aside from the target variable alone, this method partly relies on other information from the population to motivate the estimate.

Given the two methods of interest, several attempts have been made to find the criteria for comparing the various strategies while attempting to obtain optimal results from a sample survey. In this paper we explore on the use of the model-assisted estimation using Probability Proportional to

Aggregate Size Sampling (PPAS) in estimating the population total as compared to the designunbiased estimation using Simple Random Sampling Without Replacement (SRSWOR) and Probability Proportional to Size: Systematic (PPSS).

This study aims to identify the population characteristics where optimality of estimates is achieved using PPAS as compared to SRSWOR and PPSS. Data sets were simulated to explore on the different behaviours of the population of interest. Comparison of estimates were made by comparing bias and precision of estimates. Variance estimation is done with nonparametric bootstrap to address the issue of negative estimated variance.

2. Methodology

2.1 Simulation Study

To evaluate the performance of PPAS estimates under varying conditions, a simulation study was conducted. Each scenario postulates a linear model:

$$y = bx + k\varepsilon, \quad \varepsilon \sim N(0,1)$$

For this equation, the following quantities are made to vary: covariate effect (*b*), standard deviation of the auxiliary variable (sd(*X*)), multiplier (*k*) on the error term, and sampling rate. These variations aim to capture the different patterns of linear association between the target and auxiliary variable.

The covariate effect (*b*) are set to two values: 1.5 and 5 to reflect low and high covariate effect. The auxiliary variable X is randomly generated from a normal distribution with mean 50 and standard deviations 5, 10, and 40. Error terms are generated from the standard normal distribution with multipliers (*k*) set to 5, 10, 20. These values induce varying strengths of linear association between X and Y. A similar approach was used by Barrios & Kwong (2010) in simulating the different model fit for linear and nonlinear relationships between the target and auxiliary variable to capture to strong, average, and weak linear relationships, respectively. Also, as (*k*) increases, the model fit suffers because of large prediction errors. Lastly, in a population of N=1000, the random samples are drawn given the sampling rates: 1%, 5%, and 10%.

Table 2.1 Coefficient of Variation of Y = 1.5X + k*e and Pearson r Across Simulation Settings

sd(X)	5			10			40		
k	5	10	20	5	10	20	5	10	20
r	0.848	0.633	0.395	0.953	0.847	0.633	0.997	0.988	0.953
CV(Y)	0.124	0.170	0.286	0.218	0.248	0.339	0.817	0.827	0.860

Table 2.2 Coefficient of Variation of Y = 5X + k*e and Pearson r Across Simulation Settings

sd(X)	5			10			40		
k	5	10	20	5	10	20	5	10	20
r	0.982	0.935	0.801	0.995	0.982	0.935	0.999	0.998	0.995
CV(Y)	0.105	0.112	0.132	0.207	0.212	0.222	0.814	0.815	0.819

2.2 Method of Evaluation

For each combination of model restriction and sampling rate, the point estimate and standard error is generated for the SRSWOR and PPAS sampling methods, while PPSS estimates are only generated for sampling rate 1%. Standard errors serve as basis of efficiency of estimates, while the point estimates are further used to calculate the **absolute percent difference** defined by

$$PD_{est} = \left| \frac{\hat{T}_{est} - T}{T} \right| * 100\%$$

where T is the true population total, and T_{est} is the estimator of sampling procedure. The estimates are expected to be unbiased in the long run, but a single sample may incur discrepancy between the true value and the estimate. The better estimator corresponds to a lower PD_{est} value.

3. Results

3.1 Estimation of Population Total

The estimates of population total using SRSWOR, PPAS, and PPSS incorporating the three sampling rates (1%, 5%, 10%). There is notable bias in the estimates, but a clearer comparison is presented using the PD_{est} in tables 3.3-3.5.

Table 3.1 Estimates of Population Total by Sampling Design and Sampling Rate

Covariate Effect (b)	Model Fit (k)	Variance of X	Population Total	10%		5%		1%		
				SRS-WOR	PPAS	SRS-WOR	PPAS	SRS-WOR	PPAS	PPSS
1.5	5	5	75206.7	75189.0	75786.7	75461.0	76241.6	77595.0	76556.5	73612.0
		10	75332.7	75430.0	75820.0	75471.0	76183.6	79319.0	75780.9	77010.0
		40	76088.9	76873.0	76778.0	75536.0	77347.2	89659.0	78352.0	75657.0
	10	5	75287.3	75138.0	76447.3	75910.0	77357.1	78467.0	77986.9	72099.0
		10	75413.4	75379.0	76387.9	75921.0	77115.2	80191.0	76309.7	78768.0
		40	76169.6	76822.0	77547.7	75986.0	78686.0	90531.0	80695.7	75305.0
	20	5	75448.6	75036.0	77768.5	76810.0	79588.1	80211.0	80847.8	69071.0
		10	75574.6	75277.0	77523.7	76821.0	78978.3	81934.0	77367.3	82283.0
		40	76330.9	76720.0	79087.2	76885.0	81363.7	92275.0	85383.1	74602.0
5	5	5	250500.8	250751.0	251080.7	250486.0	251535.6	256617.0	251850.6	248906.0
		10	250920.9	251553.0	251408.2	250522.0	251771.8	262361.0	251369.1	252598.0
		40	253441.7	256364.0	254130.7	250737.0	254699.9	296829.0	255704.7	253010.0
	10	5	250581.4	250700.0	251741.4	250936.0	252651.1	257488.0	253281.0	247393.0
		10	251001.5	251502.0	251976.0	250971.0	252703.4	263233.0	251897.9	254356.0
		40	253522.3	256313.0	254900.5	251186.0	256038.7	297701.0	258048.4	252658.0
	20	5	250742.7	250597.0	253062.6	251835.0	254882.2	259232.0	256141.9	244365.0
		10	251162.8	251399.0	253111.8	251871.0	254566.5	264977.0	252955.5	257871.0
		40	253683.6	256210.0	256439.9	252086.0	258716.4	299445.0	262735.8	251955.0

3.2 Standard Error of Estimates

An optimal estimate is characterized by small variability, hence, the lower the standard error the more efficient the estimate. As expected, the estimates improve as sampling rate increases, but under 1% sampling rate,

PPS shows more stable estimates. When the linear model is induced with higher error, PPAS and PPS estimates are comparable. As for the 5% and 10% sampling rates, PPAS show superior estimates compared to SRSWOR. Furthermore, as additional variation in the population is introduced by increasing variation in X, SRSWOR estimates greatly suffer, but the precision of PPAS estimates remain roughly the same. What appears to affect PPAS estimates is the error in the model. When standard deviation of X is fixed to 5, it can be noted that PPAS estimates become more unstable as model error increases. If the model does not fit the data well, ratio or regression estimation might not increase precision for estimated means and totals (Lohr, 2010). In fact, at k = 20 (poor model fit), standard errors of SRSWOR and PPAS estimates are roughly similar, but under similar model fit, a stronger covariate effect improved the precision of PPAS estimates.

Table 3.2 Standard Error of Estimates by Sampling Design and Sampling Rate

Covariate Effect (b)	Model Fit (k)	Std. Dev of X	10%		5%		1%		
			SRS-WOR	PPAS	SRS-WOR	PPAS	SRS-WOR	PPAS	PPSS
1.5	5	5	852.82	455.83	1164.40	652.64	2246.50	1424.01	2020.00
		10	1477.4	449.96	2144.6	653.82	4498.2	1426.86	1141.8
		40	5528.10	429.93	8356.70	587.88	18265.00	3518.70	773.55
	10	5	1183.90	911.66	1485.20	1305.26	2384.90	2848.02	4039.90
		10	1705.60	899.92	2328.80	1307.66	4493.00	2853.72	2283.60
		40	5621.40	859.90	8396.00	1175.80	18150.00	7037.30	1547.10
	20	5	1998.70	1823.32	2378.70	2610.53	3102.80	5696.04	8079.80
		10	2367.80	1799.80	2970.50	2615.30	4769.80	5707.40	4567.20
		40	5909.60	1719.70	8578.50	2351.50	17993.00	14074.60	3094.20
5	5	5	2361.90	455.80	3509.20	652.63	7546.20	1424.01	2020.00
		10	4619.90	449.96	6968.20	653.80	15200.00	1426.90	1141.80
		40	18282.00	429.93	27832.00	587.88	61196.00	3518.65	773.55
	10	5	2524.80	911.66	3620.00	1305.26	7482.10	2848.02	4039.90
		10	4723.70	899.92	7018.50	1307.65	15092.00	2853.72	2283.60
		40	18337.00	859.86	27835.00	1175.80	61057.00	70737.30	1547.10
	20	5	3034.90	1823.30	4054.00	2610.53	7530.70	5696.03	8079.80
		10	5049.50	1799.80	7239.90	2615.30	14964.00	5707.40	4567.20
		40	18480.00	1719.70	27873.00	2351.50	60800.00	14074.00	3094.20

3.3 Average Absolute Percentage Difference of Estimates

The PD_{est} provide a standard measure to compare observed bias of estimates across model restrictions. The average of this measure can be computed to produce an estimate of the relative bias across model settings: variance of auxiliary variable, model fit, and covariate effect.

Table 3.3 summarizes the average absolute percentage difference for SRSWOR, PPAS, and PPSS estimates across the different variations in the auxiliary variable (X). It is given that as variation in X becomes larger so does the variation in Y. Under 1% sampling rate it can be noted that PPAS and PPSS estimates are generally better than SRSWOR across different variations in the auxiliary variable particularly at large values of sd(X). At higher sampling rates

(5% and 10%), SRSWOR appear to generate less bias across variations of auxiliary variable, particularly at small values of sd(*X*). This is not surprising since SRSWOR estimates work best under homogenous populations.

Table 3.3 Average Absolute Percent Difference of the Estimates Across Variations in X

sd(X)	10%		5%		1%		
	SRS	PPAS	SRS	PPAS	SRS	PPAS	PPSS
5	0.162	1.168	0.592	2.084	3.716	2.718	3.210
10	0.186	0.979	0.493	1.710	5.829	0.901	3.371
40	0.941	1.371	0.719	2.504	18.360	4.504	0.860

Table 3.4 shows the average absolute percentage of SRSWOR, PPAS, and PPSS estimates across different error multiplier (*k*). It is given that as (*k*) increases the model fit suffers and so does modelassisted estimation techniques. At 1% sampling rate, PPAS produces comparable results with PPSS and superior results when compared to SRSWOR estimates. At 5% and 10% sampling rates, PPAS estimates generate more bias as misspecification increases. Conversely, the relative bias of SRSWOR estimates remain roughly the same. This is expected since auxiliary information does not affect selection of SRSWOR samples.

Table 3.4 Average Absolute Percent Difference of the Estimates by Different Model Fit

k	10%		5%		1%		
	SRS	PPAS	SRS	PPAS	SRS	PPAS	PPSS
5	0.412	0.652	0.473	1.165	7.806	1.509	1.518
10	0.408	1.006	0.469	1.801	9.078	2.323	2.128
20	0.433	2.009	0.921	3.596	10.423	4.637	4.248

Table 3.5 summarizes the average absolute percentage of SRSWOR, PPAS, and PPSS estimates across varying covariate effect. A high covariate effect (*b*) increases the magnitude of linear association between the auxiliary and target variable. For higher covariate effect (*b*=5), PPAS estimates generate lower Average PD_{est}. This is because the increase in magnitude of Y dominated the error settings. Also, it was noted previously in Table 3.2 that increasing the covariate effect improved the efficiency of PPAS estimates, under similar model fit. In other words, an increase in covariate effect decreased bias and improved precision of the estimate, particularly under good model fit.

Table 3.5 Average Absolute Percent Difference of the Estimates by Covariate Effect

b	10%		5%		1%		
	SRS	PPAS	SRS	PPAS	SRS	PPAS	PPSS
1.5	0.415	1.803	0.797	3.228	10.148	4.163	3.814
5	0.444	0.542	0.406	0.971	8.456	1.252	1.147

4. Conclusions

The findings of the study are summarized as follows:

1. PPAS estimates are more precise than SRSWOR and PPSS, particularly in populations with less variability.
2. When sample size is small, bias of PPAS and PPSS estimates are roughly the same.
3. SRSWOR estimates have lesser bias then PPAS and PPSS, especially in populations with small variability.
4. The optimality of PPAS estimates improve as the linear association between the target variable and auxiliary variable increase.
5. PPAS estimates are more stable under large variability in population as compared to SRWOR.

The findings above may only reflect the simulated data and may not necessarily be true for other random generators. It is advisable to verify these findings by recreating the data using different random seeds used in the study. A similar study may also be conducted to explore on other non-linear relationships between the target and auxiliary variable.

References

1. Barrios E. & Kwong, A. H. (2010) *Nonparametric Model-Based Estimation in Data Mining*. 11th National Convention on Statistics. EDSA Shangri-La Hotel.
2. Efron B. (1992) *Bootstrap Methods: Another Look at the Jackknife*. In: Kotz S., Johnson N.L.
3. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY
4. Gauran, I. & Poblador, M. (2012) *Sampling with Probability Proportional to Aggregate Size using Nonparametric Bootstrap in Estimating Total Production Area of Top Cereals and Root Crops across Philippine Regions*. The Philippine Statistician Vol. 61, No. 1, pp. 87-108
5. Lohr, S. 2010. *Sampling Design and Analysis, 2nd Ed*. Boston: Brooks/Cole, p. 147.

Measure of multidimensional poverty Robustness of indices to weighting schemes

Khalid Soudi
High Commission of Planning, Rabat, Morocco

Abstract

This paper analyzes the robustness of multidimensional poverty indexes by considering various weighting schemes. In particular, it establishes comparison between three approaches for measuring multidimensional poverty, namely: fuzzy logic method, Alkire and Foster (AF) method and Bourguignon & Chakravarty method. By combining two approaches: A quantitative approach of factor analysis and a qualitative one taking into account the perception of deprivation by the poor, this paper develops an empirical approach to determine dimensions of multidimensional poverty. By referring to the analytical frame of capabilities as developed by A. Sen, a list of central and basic functioning was chosen. The results show that all poverty indices estimated by using the AF weighting scheme are significantly higher than those estimated using the other pre-determined normative weights. The stochastic dominance of the curves of poverty allowed us to confirm the robustness of these results. Moreover, if the items of deprivation are not structured by dimension, the results show, whatever selected weighting schemes, an overestimation of the indices of multidimensional poverty, in particular those obtained through the AF weighting scheme. This pattern was consistent for all weighting schemes and impacts both the headcount of poverty as well as the multidimensional poverty index (MPI). The results of Bourguinon and Chakravarty method show that, whatever is the adopted weighting scheme, the indices of multidimensional poverty are lower than those obtained from the AF approach. The differences become more important when adopting the weighting scheme suggested by AF. Besides, the differences observed become more pronounced if we don't structure the items of deprivation by dimension. Also, the multidimensional index of poverty obtained according to fuzzy logic method not only shows an important sensibility to the weighting design but also it remains widely lower than those obtained according to the approach A.F.

Keywords

weighting schemes; multidimensional poverty; robustness of indices.

1. Introduction

In recent years, despite the many methods available for measuring multidimensional poverty, some important methodological problems remain

unsettled. They relate mainly on four points: (i) the identification of the relevant dimensions of deprivation; (ii) the aggregation of the dimensional indications; (iii) the choice of the weighting scheme; and (iv) the determination of the poverty line by dimension. Each of these methodological items widely affects the targeting of the poor and the choice of relevant economic and social policies especially those aiming to reduce poverty and inequalities.

The main purpose of this work is to assess the robustness of the multidimensional indices poverty according to different weighting schemes. So, it is necessary to make the tour of the methodological frames, with a particular focus on the approach of the capabilities of A. Sen, which establishes an adequate frame to measure the multidimensional poverty. To implement this approach, three methodologies of measurement will be applied while adopting a range of weighting schemes: (i) the fuzzy logic method; (ii) the Bourguinon and Chakravarty method; and (iii) the AF method.

2. Weighting schemes: presentation of the main statistical methods

Every measure of multidimensional poverty sets somehow a weight to each well-being dimension. However, the weighting scheme can vary in its specification and the way it affects the estimation of weighted indices. From then on, the robustness of the multidimensional indices to a range of weighting schemes continues to be a serious challenge.

In this regard, it is important to analyze the sensibility of the poverty indices according to various weighting schemes. Since 1988, different statistical weighting schemes have been designed to facilitate the summing of dimensional indices in a composite index. Generally, three statistical functions of weighting can be distinguished, such as: The specifications of Desai & Shah (1988), Cerioli & Zani (1990) and Betti & Verma (1998) and Betti & al (2007).

Although there are several possible formulations of these types of functions, we present below some that are usually used to determine the weighting coefficients:

i) The function of normalized weighting proposed by Cerioli and Zani (1990): $\mathbf{w_{CZ}^{j}} = \frac{\mathbf{ln}(\sum_{\mathbf{i}} \mathbf{f(a_i)}/(\sum_{\mathbf{i}} \mathbf{f(a_i)x_{ij}}))}{\sum_{\mathbf{j}} \mathbf{ln}(\sum_{\mathbf{i}} \mathbf{f(a_i)}/(\sum_{\mathbf{i}} \mathbf{f(a_i)x_{ij}}))}$ with f $(\mathbf{a_i})$ is the weight attached to the observation of the sample $\mathbf{a_i}$,x_{ij} ϵ [0, 1] denotes the value of a particular deprivation item j.

This formulation shows that the weight attributed to the factor j is an inverse function of its degree of deprivation.

Ceriolis and Zani also developed another not logarithmic format:

$$w_{CZ-alt}^{j} = \frac{(\sum_i f(a_i)/(\sum_i f(a_i)x_{ij})}{\sum_j(\sum_i f(a_i)/(\sum_i f(a_i)x_{ij}))}$$

ii) The function of normalized weighting proposed by Desai and Shah (on 1988)

$$w_{DS}^{j} = \frac{1 - \left((\sum_i f(a_i) x_{ij}) / \sum_i f(a_i) \right)}{\sum_j \left(1 - \left((\sum_i f(a_i) x_{ij}) / \sum_i f(a_i) \right) \right)}$$

In this specification, although the weighting function grants more weight for the deprivation, the fact of not considering a logarithmic function allows to grant more importance for indices of deprivation translating less frequent symptoms of poverty. This approach tends to converge the weights of items measuring the deprivation. Unlike the approach of Cerioli and Zani which overrepresented the weight of the least wide-spread deprivation.

iii) Weighting scheme of Betti & Verma

This weighting scheme is based on two principles: i) the weight to be attributed measures the intensity of deprivation of an attribute; its value is an inverse function of the degree of deprivation of this attribute for the population; ii) the index measuring this weight aims to reduce the over-representation due to the risks of the high correlation between the attributes and to the redundancy of the information. So, the suggested function removes from the calculation items bringing the same information by eliminating their weights.

$$w_{BV}^{j} = \frac{w_a^j \times w_b^j}{\sum_{m=1}^{M} w^m} \quad \text{with} \quad w_a^j = \frac{\left(\sum_i f(a_i)\,(x_{ij} - \bar{x}_j)^2\right)^{\frac{1}{2}}}{\left(\sum_i f(a_i)\right)^{\frac{1}{2}} \bar{x}_j} \text{ and}$$

$$w_b^j = \left(1 + \sum_{m=1}^{M} \rho_{jm} \cdot \mathbf{I}(\rho_{jm} < \rho_h)\right)^{-1} \times \left(\sum_{m=1}^{M} \rho_{jm} \cdot \mathbf{I}(\rho_{jm} \geq \rho_h)\right)^{-1}$$

w_a^j depends on the distribution of the attribute j in the population, and w_b^j depends on the correlation between x_j and other attributes, it measures the average correlation of item j with all the other items. The more it increased the lower is the weight of the attribute j. ρ_{jm} represents the level of correlation between two attributes j & m, I(.) is an indicator function and ρ_h is a pre-determined cutoff correlation level
(in our case $\rho_h = 0.5$).

iv) The third weighting scheme combines the advantages of Cerioli-Zani method and Betti-Verma method : It attributes an upper weight for the least wide-spread deprivation, and limites the influence of the correlation and the redundancy of the information on the weighting. Within the framework of this study, we adopted the following both specifications:

$$W_{BV-CZ}^{j} = \frac{\left(\ln(\sum_i f(a_i) / (\sum_i f(a_i) x_{ij}))\right)}{\left(1 + \sum_{m=1}^{M} \rho_{jm} \middle| \mathbf{I}(\rho_{jm} < \rho_h)\right)\left(\sum_{m=1}^{M} \rho_{jm} \middle| \mathbf{I}(\rho_{jm} \geq \rho_h)\right)}$$

This function would allow to attribute a weight to the item j by combining the methods of Betti & Verma and Cerioli & Zani.

Similar to the specification BV-CZ, the second specification would allow us to adjust the function of the weight proposed by Desai-Shah by introducing

the measure of the average correlation which grants less weight for the strongly correlated attributes.

$$W^{j}_{BV-DS} = \frac{(1 - (\sum_i f(a_i) x_{ij}) / \sum_i f(a_i))}{\left(1 + \sum_{m=1}^{M} \rho_{jm} \middle| \mathbf{I}(\rho_{jm} < \rho_h)\right)\left(\sum_{m=1}^{M} \rho_{jm} \middle| \mathbf{I}(\rho_{jm} \geq \rho_h)\right)}$$

v) **Linear weighting scheme of A.F**: this weighting scheme gives equal weighting to each dimension and each item in every dimension.
$W^{j}_{A.F} = \frac{1}{D.K_j}$ with D: number of dimensions and K the number of items in the dimension containing the item j.
Definition of the space of deprivation: central and basic functioning and indices of measure

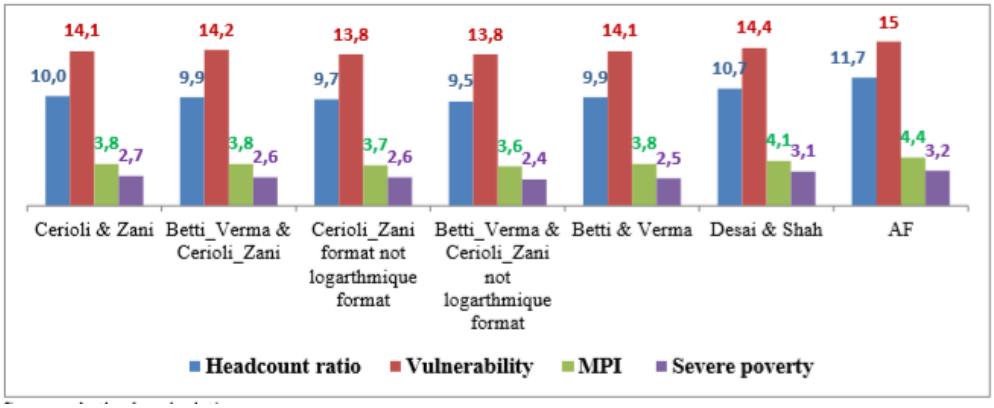
Dimension	**central and basic functioning**	**indicators of measure**
Education	Schooling for adults	The number of members of the household having completed 5 years of schooling.
	Schooling of children	Number of chidren of school age attending the school.
Maternal health and children's nutrition	medicale consultation	Sick individuals who have access to a medical consultation.
	Deliveries under medical supervision	Deliveries occured in watched environment
	Children nutrition	Number of undernourished children
Economic power	Income inequality	Income per capita
Labour market participation	To be put in job	Number of unemployed by household
Food consumption	Provide red or white meats and fish	per capita average spending for meats and fishes
Housing conditions	Sanitary equipments	Sanitary equipments (toilet, washbasin, bath/shower, sewers)
	Domestic equipments	Domestic equipments (Tv, radio, telephone, refrigerator, bicycle, moped, car, tractor)
	Electricity	Access to electricity
	Water	Access to drinking water

Source: the national survey on living conditions, 2007, HCP, Morocco

3. Summary of the results: measurement indices of poverty according to approaches of measure and weighting schemes

By adopting AF method, the results show that all indices of poverty (headcount ratio of poverty, multidimensional poverty index (MPI), vulnerability and severe poverty) estimated according the AF weighting scheme are significantly higher than those estimated using the other pre-determined normative weights. The stochastic dominance of the curves of poverty allows us to confirm the robustness of these results.

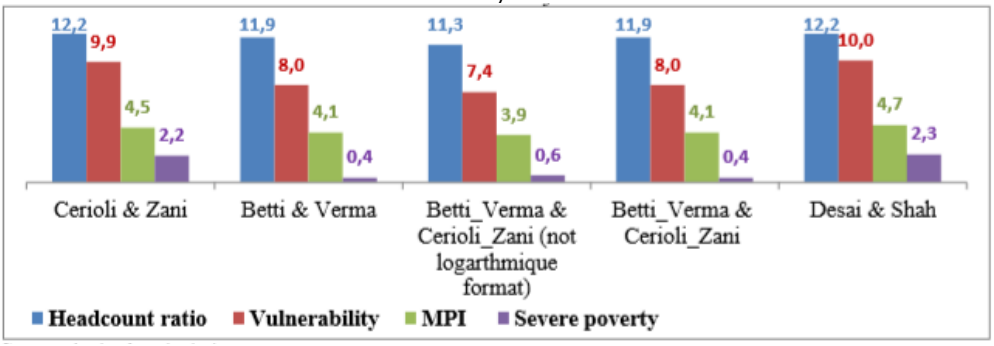
Graph 1: Multidimensional indices of poverty according to AF method : sensibility to weighting schemes. Factors structured by dimension (%)



Source: Author's calculation

If we don't structure the items by dimension we'll have an overvaluation of the index of multidimensional poverty, in particular when using the AF weighting scheme. This tendency concerns all the weighting schemes, the headcount of poverty and the multidimensional poverty index (MPI). In light of these results, it turns out important to rethink the broad configuration of the deprivation items by dimension.

Graph 2: Multidimensional indices of poverty according to the approach AF Factors aren't structured by dimension

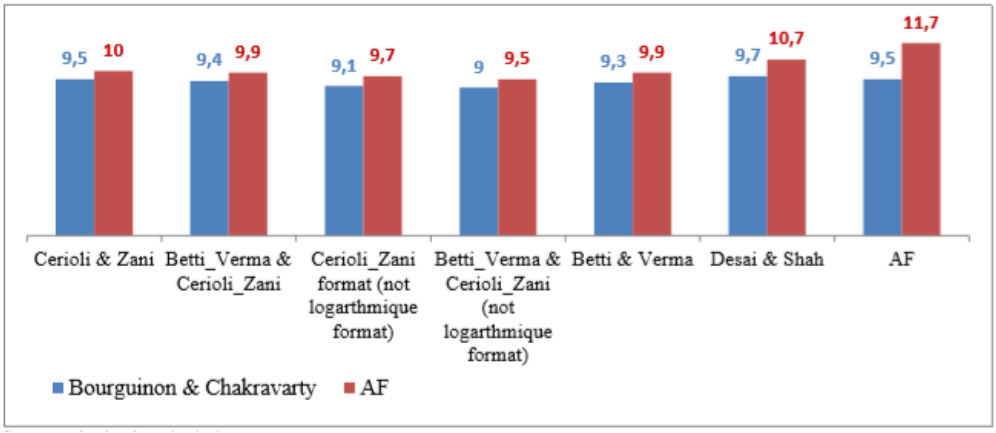


Source: Author's calculation

By structuring the items of deprivation by dimension, the results of the fuzzy logic approach show that, regardless of weighting scheme, the indices

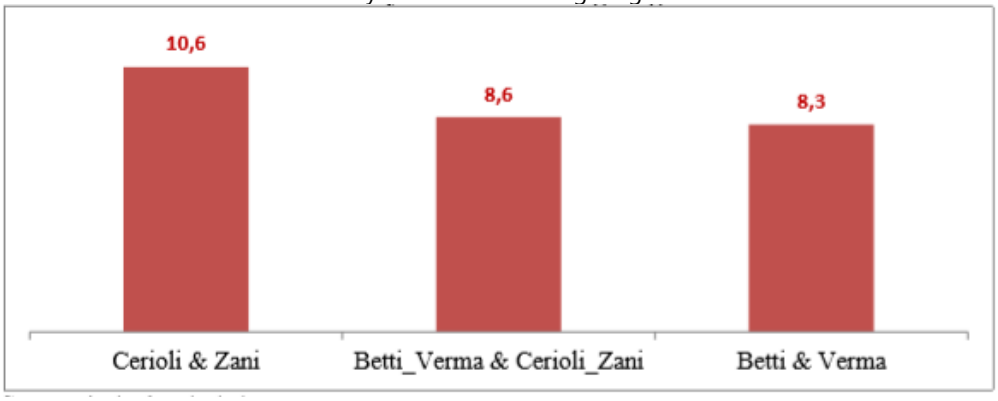
of poverty are lower than those emanating from the AF approach. When the items of deprivation are not structured by dimension the observed differences become more significant. Besides, these differences also show the tendency pronounced by the approach AF to overestimate the incidence of poverty in comparison with the fuzzy logic approach. Straightaway, by referring to the approach of Bourguinon and Chakravarty (2002), the multidimensional indices of poverty obtained while showing a sensibility important for the weighting schemes, they remain lower than those obtained according to the AF method.

Graph 3: Comparison of headcount ratio between AF method and Bourguinon & Chakravarty method. Factors structured by dimension (%)



Source: Author's calculation

Graph 4: Headcount ratio of poverty[7] according to fuzzy logic method
Sensibility of measure to weighting schemes



Source: Author's calculation

[7] According to fuzzy logic method, a household is considered poor if it accumulates at least 30 % of the deprivations among the dimensions determining the well-being.

4. Conclusion

This paper studied the sensitivity of multidimensional indices of poverty to different weighting schemes. Our approach explained why multidimensional poverty measure depends not only on the approach but also on the weighting scheme adopted. Testing of pre-determined normative weighting schemes showed that the linear weighting overestimates poverty indices. Whatever the weighting scheme adopted, an important difference is observed in the estimation of poverty indices depending on whether the items of deprivation are organized or not by dimension. The most important difference concerns the AF approach. This tendency concerns all the multidimensional indices of poverty. The stochastic dominance of the curves of poverty allowed us to confirm the robustness of these results.

The results of Bourguinon and Chakravarty method show that, regardless of the weighting scheme, the indices of multidimensional poverty are lower than those obtained from the AF approach. The differences become more significant by adopting the weighting scheme proposed by AF. Besides, the differences noticed become more important if we don't structure the items of deprivation by dimension. Also, the multidimensional indices of poverty obtained according to fuzzy logic method not only shows an important sensibility for the weighting schemes but remains widely lower than those obtained using the A.F. approach.

In addition, this work shows that a better targeting of the poverty is not only conditioned by a determination of the relevant dimensions of the poverty but also by an adequate choice of the weighting scheme and the measurement approach.

References (some papers)

1. Alkire S. et J. Foster (2009), « Counting and Multidimensional Poverty Measurement », OPHI, WP No.32
2. Atkinson A.B. (2003), «Multidimensional Deprivation: Contrasting Social Welfare and Counting Approaches » Journal of Economic Inequality, 1, 51-65.
3. Betti G. & Verma V. K. (1999), «Measuring the degree of poverty in a dynamic and comparative context: a multi-dimensional approach using fuzzy set theory », Working Paper 22, Dipartimento di Metodi Quantitativi, Università di Siena.
4. Betti G., Cheli B., Lemmi A. & Verma V. K. (2007), «The fuzzy set approach to the multidi- mensional poverty: The case of Italy in the 1990s', in N. Kakwani & J. Silber (eds.), Quantitative Approaches to Multidimensional Poverty Measurement », Palgrave Mac Millan, New York, 30-48. Boniface E. (2000), « Inégalité, pauvreté et bien-être social: fondements analytiques et normatifs », Ouvertures Economiques, De Boeck Université.

5. Bourguignon F., S. Chakravarty, (2003), « The measurement of multidimensional poverty », Journal of Economic Inequality 1, 25-49.
6. Cerioli A. & Zani S. (1990), «A fuzzy approach to the measurement of poverty », in C. Dagum & M. Zenga (eds.), Income and Wealth Distribution, Inequality and Poverty, Springer Verlag, Berlin.
7. Cheli B., Lemmi A**.,** (1995). « A total fuzzy and relative approach to the multidimensional analysis of poverty », *Economic Note*, Vol. 24, 1, pp. 115-134.
8. Chiappero-Martinetti E. (2000) « A Multidimensional Assessment of Well-Being based on Sen's Functioning Approach », *Rivista Internazionale di Scienze Sociali, Università Cattolica Del Sacro Cuore*, n°2, Milano.
9. Dagum, C., Gambassi, R., Lemmi, A**.,** (1992). «New approaches to the measurement of poverty», *Poverty measurement for economies in transition in eastern European countries,* Polish Statistical Association and Central Statistical Office, Warsaw.
10. Desai M. & Shah A. (1988), «An econometric approach to the measurement of poverty », Oxford Economic Papers, 40(3):505-522.
11. Desai M. (1995), « poverty, Famine and Economic Development », Edward Elgar, Aldershot. (1995), «Poverty and Capabilities: Towards an Empirically implementable Measure », in the selected Essays of Meghnad Desai, Volume 2, *Economist of the Twentieth Century Series*, Aldershot, UK, Elgar, 1995.
12. Duclos J.-Y., D. Sahn and S.D. Younger (2006), «Robust Multidimensional Poverty Comparisons », Economic Journal, 116, 943{968.
13. Sen A. (1976) « Poverty: an ordinal approach to measurement », *Econometrica*, vol 44, pp 219-231. (1981) «*Poverty and famines: an essay on entitlements and deprivation* », Clarendon Press, Oxford, Royaume-Uni.
14. Soudi, K. (2009), « Dynamiques de la pauvreté 1985-2001 : rôles de la croissance et de l'inégalité »,26ème Congrès de l'UISSP, Marrakech.
15. Tsui K-Y. (2002), «Multidimensional Poverty Indices», *Social Choice and Welfare* 19: 69-93.

Fuzzy clustering in a reduced subspace

Paolo Giordani, Maria Brigida Ferraro, Mario Fordellone, Maurizio Vichi
Sapienza University of Rome, Rome, Italy

Abstract

A general method for two-mode simultaneous reduction of observation units and variables of a data matrix is introduced. It consists in a compromise between the Reduced *K*-Means (R*K*M) and Factorial *K*-Means (F*K*M) procedures. Both methodologies involve principal component analysis for variables and *K*-Means for observation units, even though R*K*M aims at maximizing the between-clusters deviance without imposing any condition on the within-clusters deviance, while F*K*M aims at minimizing the within-clusters deviance without imposing any condition on the between one. It follows that R*K*M and F*K*M complement each other. In order to take advantage of both methods a convex linear combination of the R*K*M and F*K*M loss functions is used. Furthermore, the fuzzy approach to clustering is considered because of its flexibility in handling the real-world complexity and uncertainty.

Keywords

Subspace clustering; Factorial *K*-Means; Reduced *K*-Means; Linear convex combination; Fuzzy approach to clustering

1. Introduction

Clustering is the process of discovering a partition of *I* observation units in a limited number of groups or clusters *K* (< *I*) such that observation units belonging to the same cluster are similar according to a certain criterion. When *J* quantitative variables are observed on the set of observation units, the most common choice is to consider the (squared) Euclidean distance in order to compute the dissimilarities between pairs of observation units. The probably most famous clustering algorithm involving the squared Euclidean distance is the *K*-Means (*K*M) algorithm (MacQueen, 1967). It provides a partition of the observation units into *K* clusters, summarized by *K* centroids, in such a way to minimize the within-cluster sum of squares.

The Euclidean distance is usually evaluated by considering all *J* variables. This is inadequate when there exists a subset of variables, which does not play a relevant role to properly recover the cluster structure and actually, tends to mask it. To overcome this problem, several strategies can be adopted. Roughly speaking, they consist in reweighting the variables in such a way to increase or decrease their role in the clustering process. For every variable, the higher

the associated weight is, the more the variable plays a relevant role. If the weight of a variable is equal to zero, then the variable is discarded. There are various ways to give weights to the variables. A common one is based on Principal Component Analysis (PCA), i.e., by considering the component loadings. In this case, the principal components span a low dimensional space of order Q ($< J$) where the observation units are projected. The partition is carried out by clustering the observation units in terms of their coordinates on such a low-dimensional space, i.e., in terms of the component scores. For this reason, we refer to as *subspace clustering*.

In the naïve approach to subspace clustering, the data reduction and the clustering steps are done sequentially. In other words, firstly, PCA is applied to the data, then the clustering method is run on the resulting component scores. Such an approach is usually known as *tandem analysis* (Arabie & Hubert, 1994). Although it is very intuitive, its use is not recommended because the principal components are not optimal in the clustering sense. In fact, as is well-known, they maximize the total sum of squares and therefore may lead to a low-dimensional configuration of the observation units such that the taxonomy is obscured. For more details, the interested reader may refer to, for instance, De Sarbo et al. (1990) and De Soete & Carroll (1994).

In order to address the clustering problem in a reduced subspace simultaneously, at least two proposals can be used. These are the Reduced *K*-means (R*K*M) analysis suggested by De Soete & Carroll (1994) and the Factorial *K*-means (F*K*M) analysis suggested by Vichi & Kiers (2001). Both methods detect a partition of the observation units in K clusters by assuming that centroids lie in a subspace of variables. Although they are based on the same assumption, as we shall see, they present distinctive features.

In this paper, we are going to propose a new clustering method in a reduced subspace exploiting the potentialities of R*K*M and F*K*M. For this purpose, a linear convex combination of the R*K*M and F*K*M loss functions will be used. Furthermore, in order to enlarge the applicability of our proposal, the clustering problem is approached from the fuzzy point of view (Zadeh, 1965). In contrast with the standard approach where the observation units either belong or not to the clusters and every observation unit can be assigned to one and only one cluster, the fuzzy approach allows to assign the observation units to the clusters with the so-called fuzzy membership degrees ranging in the interval [0, 1], where 0 means complete non-membership and 1 complete membership, and such that, for each observation unit, the sum of the fuzzy membership degrees is equal to one.

The paper is organized as follows. In the next section, R*K*M and F*K*M are recalled and the new proposal is introduced. In Section 3 the results of the application of the new clustering procedure to real data are reported. Some final remarks in Section 4 conclude the paper.

2. Methodology

In this section we start by reviewing R*K*M and F*K*M. Later, the new clustering procedure is illustrated in detail.

a. Reduced *K*-Means (R*K*M):

Let **X** be the data matrix of order ($I \times J$) containing the scores of I observation units with respect to J variables. The Reduced *K*-means (R*K*M) analysis (De Soete & Carroll, 1994) can be formulated as:

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{E} \quad (1)$$

where **U**, of order ($I \times K$), is the membership matrix with elements equal to 0 or 1 expressing for each observation units the membership to one of the K clusters. Note that **U** is row-stochastic, that is, its row-wise sum is equal to 1. **A** is the component weight matrix of order ($J \times Q$). It is column-wise orthonormal, i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$, being $\mathbf{I}_Q$ the identity matrix of order Q, and every column expresses the weights of the variables on the corresponding component. Finally, **F** is the centroid matrix of order ($K \times Q$) such that every row refers to a cluster centroid. The centroids lie in the reduced subspace spanned by the columns of **A**. Finally, **E** is the residual matrix having the same order of **X**. The optimal parameter matrices **U**, **F** and **A** are obtained in the least square sense by minimizing the residual sum of squares:

$$f_{RKM} = \| \mathbf{E} \|^2 = \| \mathbf{X} - \mathbf{UFA}' \|^2, \quad (2)$$

being $\| \cdot \|$ the Frobenius norm of matrices. Suitable Alternating Least Squares (ALS) algorithms can be adopted for the minimization of (2). The R*K*M solution is not unique. Equally fitting solutions can be found up to rotational indeterminacy for the weights in **A** and the centroids in **F** and cluster label switching. Given an orthonormal rotation matrix **R** of order ($Q \times Q$) and a permutation matrix **P** of order ($K \times K$), letting $\mathbf{A}^* = \mathbf{AR}$, $\mathbf{U}^* = \mathbf{UP}$ and $\mathbf{F}^* = \mathbf{P}'\mathbf{FR}$, we have $\mathbf{U}^*\mathbf{F}^*\mathbf{A}^{*\prime} = \mathbf{UFA}'$.

When $Q = J$, i.e., when the variable space is not reduced through PCA ($\mathbf{A} = \mathbf{I}_J$, where $\mathbf{I}_J$ is the identity matrix of order J), R*K*M boils down to the standard *K*M algorithm.

b. Factorial *K*-Means (F*K*M):

The R*K*M loss function in (2) is a proxy of the within-cluster sum of squares in the reduced space. In fact, it is the sum of the squared distances between the observation units in the (J-dimensional) observed space and the centroids in the (Q-dimensional) reduced space. This represents a sort of idiosyncrasy because it appears more reasonable to compute the within-cluster sum of squares in the reduced space by considering not only the centroids but also the observation units lying in the reduced space. This motivation leads to the so-called Factorial *K*-Means (F*K*M) procedure developed by Vichi & Kiers (2001). The F*K*M model is expressed as

$$\mathbf{XAA}' = \mathbf{UFA}' + \mathbf{E}. \quad (3)$$

As for R*K*M, parameters can be addressed by the least square estimation method. In particular, the optimal parameter matrices are found by minimizing

$$f_{FKM} = \| \mathbf{E} \|^2 = \| \mathbf{XAA}' - \mathbf{UFA}' \|^2 = \| \mathbf{XA} - \mathbf{UF} \|^2. \quad (4)$$

For this purpose, an ALS algorithm can be used. The F*K*M solution is found up to rotational indeterminacy for the weights and the centroids and cluster label switching. If $\mathbf{A} = \mathbf{I}_J$, F*K*M coincides with *K*M.

c. Comparison between F*K*M and R*K*M:

Vichi & Kiers (2001) and Timmerman et al. (2013) investigate the F*K*M and R*K*M procedures from a theoretical and a practical point of view. Their findings are summarized in this subsection. First of all, Vichi & Kiers (2001) show that R*K*M may fail when a large amount of variance pertains to directions orthogonal to the one relevant for clustering purposes. It implicitly suggests that the two methods model the data in different ways. Timmerman et al. (2013) extensively analyze this point by defining two types of residuals, namely, *subspace residuals* and *complement residuals*. The former ones denote the residuals lying on the subspace spanned by the columns of **A**. The latter ones refer to the residuals lying on the complement of this subspace, i.e., those lying on the subspace spanned by the columns of $\mathbf{A}^{\perp}$, being $\mathbf{A}^{\perp}$ a column-wise orthonormal matrix of order ($J \times J - Q$) such that $\mathbf{A}'\mathbf{A}^{\perp} = \mathbf{0}_{Q \times J - Q}$, where $\mathbf{0}_{Q \times J - Q}$ is the matrix of zeroes of order ($Q \times J - Q$). Real life data usually contain both kinds of residuals. The performances in recovering the clusters of F*K*M and R*K*M are related to the relative sizes of such two kinds of residuals. Specifically, F*K*M performs better than R*K*M when the complement residuals are smaller than the subspace ones. Conversely, R*K*M outperforms F*K*M when the subspace residuals are small if compared to the complement ones.

d. Fuzzy Reduced and Factorial *K*-Means (FRF*K*M):

Taking into account that the objectives of F*K*M and R*K*M are different and every method aims at minimizing a specific kind of residuals, our idea is to exploit the potentialities of both methods by considering them simultaneously. In doing so, we adopt the fuzzy approach to clustering by relaxing the constraints that the elements of **U** are either 0 or 1. In order to handle the concept of partial truth, where the truth value may range between completely false and completely true, it implies that every observation unit can be assigned to a certain cluster with the so-called fuzzy membership degree ranging from 0 (non-membership, completely false) to 1 (full membership, completely true).

The new procedure, called Fuzzy Reduced and Factorial *K*-Means (FRF*K*M), is developed by considering a linear convex combination of the loss functions in f_{RKM} in (2) and f_{FKM} in (4). Thus, we have

$$f_{FRFKM} = (1-\alpha) f_{RKM} + \alpha f_{FKM}, \quad (5)$$

where α is a tuning parameter taking values in the interval [0, 1]. It can be shown that (5) can be rewritten as

$$f_{FRFKM} = \| \mathbf{X}[\mathbf{AA}' + (1-\alpha)^{1/2}(\mathbf{I}-\mathbf{AA}')] - \mathbf{UFA}' \|^2. \quad (6)$$

Therefore, the FRF*K*M model can be formulated as

$$\mathbf{X}[\mathbf{AA}' + (1-\alpha)^{1/2}(\mathbf{I}-\mathbf{AA}')] = \mathbf{UFA}' + \mathbf{E}. \quad (7)$$

The FRF*K*M procedure contains several clustering methods as special case. In fact, if **U** is a membership matrix, the F*K*M is obtained when $\alpha = 1$, whilst FRF*K*M reduces to R*K*M when $\alpha = 0$. Furthermore, if $Q = J$, $\mathbf{A} = \mathbf{I}_J$ and thus (7) is simplified as

$$\mathbf{X} = \mathbf{UF} + \mathbf{E}, \quad (8)$$

which resembles the well-known Fuzzy *K*-Means algorithm (Bezdek, 1981) in case **U** is a fuzzy membership degree matrix (i.e., $\mathbf{U} \geq \mathbf{0}_{I \times K}$, and $\mathbf{U1}_K = \mathbf{1}_I$).

Thus, the optimal parameter matrices **U**, **F** and **A** are found by minimizing (6) subject to the following constraints:

$$\mathbf{U} \geq \mathbf{0}_{I \times K}, \quad (9)$$

$$\mathbf{U1}_K = \mathbf{1}_I, \quad (10)$$

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_Q, \quad (11)$$

where $\mathbf{0}_{I \times K}$ is the matrix of zeroes of order $(I \times K)$, $\mathbf{1}_K$ and $\mathbf{1}_I$ are the vectors of ones of lengths K and I, respectively. The minimization of (6) under the constraints in (9)-(11) can be done by an ALS algorithm. As for F*K*M and R*K*M, the FRF*K*M solution is not unique due to rotational indeterminacy for the weights and the centroids and cluster label switching. This can be used to simplify the interpretation of the components by means of, e.g., the varimax procedure (Kaiser, 1958). Moreover, the fuzzy nature of **U** leads to an additional source of indeterminacy. In order to fix it and to further improve the interpretability of the solution, additional constraints on the centroids may be used following Suleman (2015). This is motivated by the fact that the optimal centroid matrix **F** can be determined by regression. This does not guarantee that the estimated centroids have a reasonable meaning. For this purpose, we can constrain the centroids to be convex combinations of the observation units according to the so-called archetypal analysis (see, for instance, Cutler & Breiman, 1994; Epifanio et al., 2018). In detail, Suleman (2015) proposes to estimate the centroids by means of a particular archetypal analysis where the weights of the convex combinations are equal to the fuzzy membership degrees. In the current framework, it may be convenient to estimate the centroids in a similar way by considering a weighted mean of the observation units projected on the subspace spanned by the columns of **A**.

2.4.1. Choice of *Q*, *K* and α:

The selection of the number of components Q, the number of clusters K and the weighting parameter α can be done subjectively or according to model selection heuristics by using different triplets of values for Q, K and α. Note that the selection of α can be done by considering a subset of values in

the interval [0, 1], e.g., {0, 0.05, …, 0.95, 1}. The optimal triplet can then be determined by means of, for instance, the Calinski Harabasz index (Caliński, & Harabasz, 1974) or the Silhouette one (Kaufman and Rousseeuw, 1990), widely used in the standard (non-fuzzy) clustering framework. Other indexes for fuzzy clustering, such as the Fuzzy Silhouette index (Campello & Hruschka, 2006) or the Xie & Beni one (Xie & Beni, 1991), can also be adopted. An alternative strategy can be based on cross-validation techniques.

3. Results

We analyzed the NBA data (Ferraro et al., 2018) referring to $I = 30$ NBA teams on which $J = 11$ statistics (see Table 1) for the regular season 2017/18 were collected. The data also contained two additional variables concerning the conference (Western or Eastern) and the playoff appearance (Yes or No).

By means of FRF*K*M we aimed at studying whether a partition of the NBA teams can be discovered and whether the team statistics can be summarized through a limited number of components. The data were standardized before running FRF*K*M. We decided to vary K and Q in the set {2, 3, 4, 5} and α in the set reported in Section 2.4.1.

Table 1. Variables

Acronym	Statistic
FGP	field goal percentage
3PP	3-point field goals percentage
FTP	free throw percentage
OREB	offensive rebounds
DREB	defensive rebounds
AST	assists
TOV	turnovers
STL	steals
BLK	blocks
BLKA	blocked field goal attempts
PTS	points

The optimal values of K, Q and α were found by maximizing the Fuzzy Silhouette index. The maximum value (equal to 0.82) was registered when $K = 2$, $Q = 2$ and $\alpha = 0.95$, hence, the FRF*K*M solution was mainly based on the F*K*M one. The obtained components were simplified by exploiting the rotational freedom of FRF*K*M. Varimax-rotated component weights reported in Table 2 were found.

Table 2. Component weights (weights higher than 0.30 in absolute value are in bold)

Variable	Component 1	Component 2
FGP	0.01	**0.86**
3PP	**0.33**	0.04
FTP	0.21	0.18
OREB	-0.18	-0.01

DREB	-0.04	**0.35**
AST	**0.42**	0.07
TOV	0.05	0.04
STL	0.17	0.04
BLK	**0.43**	-0.01
BLKA	-0.28	-0.20
PTS	**0.58**	-0.25

Component 1 was characterized by a large number of points (PTS) due to a high 3-point field goals percentage (3PP) and to a high number of assists (AST) and by a large number of blocks (BLK). Component 2 was related to a high field goal percentage (FGP) and a lot of defensive rebounds (DREB). These were variables playing the most relevant role in determining the clusters. Some variables such as TOV, STL and OREB do not seem to contribute to the description of the clusters.

The two cluster centroids distinguished the teams with respect to low (Cluster 1) or high (Cluster 2) levels of the two components. Therefore, the teams assigned to Cluster 1 had a worse performance in comparison with those belonging to Cluster 2. The membership degree matrix (not reported here) offered a better insight into the obtained clusters. We found that all the teams assigned to Cluster 2 reached the playoff stage, whilst Cluster 1 was composed by all the teams not qualified to the playoff stage and some other teams which played the next stage. Note that Cluster 2 contained the NBA champion (Golden State Warriors) and the runner-up of the finals (Cleveland Cavaliers) and teams ending the regular season within the first four positions. Hence, the obtained partition was able to identify the best NBA teams for the regular season 2017/18, i.e., those assigned to Cluster 2, which were extremely good if compared to the remaining teams, i.e., those assigned to Cluster 1.

4. Discussion and Conclusion

In this paper, a new clustering procedure for detecting a fuzzy partition of observation units in a reduced subspace is introduced. It is based on the R*k*M and F*k*M methods by considering a linear combination of their loss functions and by replacing the (hard) allocation matrix by the fuzzy membership degree matrix. The effectiveness of the proposal is shown by a real-life example.

References

1. Arabie, P. & Hubert, L. (1994). Cluster analysis in marketing research. In: Bagozzi, R.P. (Ed.), Handbook of marketing research. Blackwell, Oxford.
2. Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithm. New York. Plenum Press.
3. Caliński, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. Comm. Statist., 3, 1–27. R.J.G.B. Campello, R. J. G. B. & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets Syst., 157, 2858–2875.
4. Cutler, A. & Breiman, L. (1994) Archetypal analysis. Technometrics, 36, 338–347.
5. DeSarbo, W. S., Jedidi, K., Cool, K. & Schendel, D. (1990). Simultaneous multidimensional unfolding and cluster analysis: an investigation of strategic groups. Marketing Lett., 2, 129–146.
6. De Soete, G. & Carroll, J. D. (1994). *k*-means clustering in a low-dimensional Euclidean space. In: Diday, E., et al. (Eds.), New Approaches in Classification and Data Analysis. Springer, Heidelberg, pp.212–219.
7. Epifanio, I., Ibáñez, M. V. & Simó, A. (2018). Archetypal shapes based on landmarks and extension to handle missing data. Adv. Data Anal. Classif., 12, 705–735.
8. Ferraro, M. B., Giordani, P. & Serafini, A. (2018). fclust: an R package for fuzzy clustering. Submitted.
9. Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187–200.
10. Kaufman, L. & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York. Wiley.
11. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1, Univ. of Calif. Press, 281-297.
12. Suleman, A. (2015). A convex semi-nonnegative matrix factorisation approach to fuzzy *c*-means clustering. Fuzzy Sets Syst, 270, 90–110.
13. Timmerman, M. E., Ceulemans, E., Kiers, H. A. L. & Vichi, M. (2010). Factorial and reduced *K*-means reconsidered. Comput. Statist. Data Anal., 54, 1858–1871.
14. Vichi, M. & Kiers, H.A.L. (2001). Factorial *k*-means analysis for two-way data. Comput. Statist. Data Anal., 37, 49-64.
15. Xie, X. L. & Beni, G. (1991). A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell., 13, 841–847.
16. Zadeh, L. A. (1965). Fuzzy sets. Inf. Control, 8, 338–353.

Smart meters data as a source of household and farms statistics a case study from the United Arab Emirates

Zaid AlQadhi, Marwa F. Elkabbany, Hessa Al Shehhi, Noora Ali
Federal Competitiveness and Statistics Authority, United Arab Emirates

Abstract

Adoption of electricity and water meters data has proven to be an efficient source of household information and land-use distribution. In 2013, Federal Competitiveness and Statistics Authority (FCSA) completed a project that aimed to define, validate and acknowledge the methodology of utilizing meter-driven data in identifying household locations and densities. Such digital data adoption methods will minimize field survey and thus save governments' resource and save a lot of time. This paper represents a smart approach to extract farm and residential units' statistics through deriving data from active utility meters. The identification and classification of active residential/farm water and electricity meters was harnessed to detect the presence of a residential/farm unit. The number of meters was then aggregated at a district and sub-district levels to extract the count of families and farms within the geographical area and deduce family and farm density per square kilometre. This approach was used for 2013 and 2017 datasets to present the growth happening after 4 years.

Keywords

Government information sharing; digital enumeration; smart digital census; meters derived statistics; household statistics;

1. Introduction

FCSA continuously works with the United Arab Emirates government agencies and statistical centres to provide accurate and reliable statistics on social, economic, environmental and other conditions for decision makers, policy makers, the public, the media, the business community, researchers and the international community.

In order to achieve this, FCSA is developing the national statistical system in the country to ensure the safety of statistical methodologies and ensure that they conform to the best international standards and follow new methods and practices to benefit from various traditional and other sources of registration.

The preparation of population density maps supports the development of the national statistical system in order to provide accurate statistics for decision makers, policy makers as well as all domestic and international users.

While national censuses consumes governmental resources, FCSA conducts continuous researches to validate and practice new approaches that shall decrease the urge of conducting field surveys and control enumeration/census operation costs.

2. Methodology

The methodology behind the project presented in this paper is developed on the utilization of water and electricity meters data as a source for calculating household density and farm density which is considered as a different methodology from the traditional population density methods however has proven to be a feasible source to derive household information (Gajowniczek and T.Ząbkowski, 2015). Moreover, in 2013 FCSA combined water and electricity meters data with the output of the 2013 family framework project, to validate the use of smart meters dataset as a source of household information which may be considered as an alternative to the household census after applying the necessary data processing methodology.

The Federal Electricity and Water Authority (FEWA) has an established meter data management system which was the main data source for this project. The meter database included details on each meter type, location, activity, owner name, owner nationality, unit land use, serial, etc.

In collaboration with FEWA, the meters data for 2 years – 2013 and 2017- was derived from FEWA meter data management system and processed, and analysed in order to extract the required information.

The process performed on the meters dataset aimed to:

- Include only meters that are active / in use. (Table 2: Summary of number of meters found in each emirate before processing and filtering)
- For each household unit, only one utility meter shall be counted; either water meter or electricity meter) (Figure 2: Meters Density Distribution)
- Identify where the unit usage is residential or farm.
    - For residential units: Identify whether the owner of the unit is UAE national or an expat.

The project database and implementation leveraged geographic information systems (GIS) tools and techniques. The diagram below (Figure 1: Project methodology and implementation steps) briefly elaborates the methodology and implementation steps. They were set after iterations of process reviewing and results auditing to validate data assumptions and ensure the maintenance of accurate results.

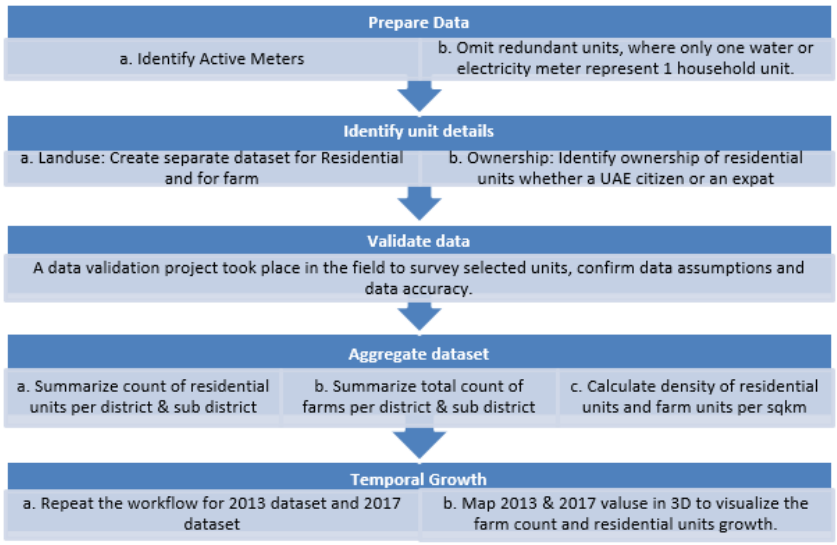


Figure 1: Project methodology and implementation steps

Numbers of Meters Records received for each emirate		
Emirate	Electricity	Water
Ajman	12,4189	83,286
Umm AlQuwain	19,645	11,774
Ras Al khaimah	66,264	49,189
Fujairah	44,790	30,023
Sharjah	12,991	6,252

Table 2: Summary of number of meters found in each emirate before processing and filtering

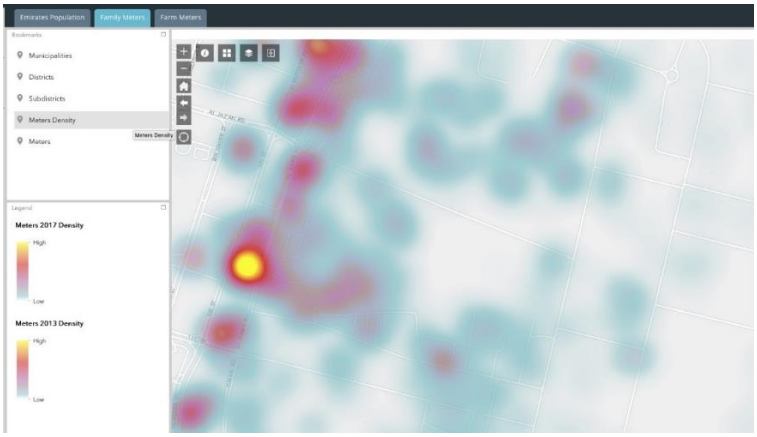


Figure 2: Meters Density Distribution

3. Results

The results of data extraction methods, applied methodology and field verifications deduced that each active meter represents one farm or one residential unit – based on the unit use stated in the meters data management system- and belongs administratively to the geo-location of the meter.

The outputs included two residential datasets for 2013/2017 and two farm datasets for 2013/2017 as well. The datasets were spatially joined and aggregated with different administrative boundaries *emirates, municipalities, district and sub districts.* Thus the digital enumeration and density of residential units as well as farm units were extracted (Figure 3: Families/ Residential Units Density Distribution (labelled with administrative unit name) and Figure 4: Farm Density 2017).

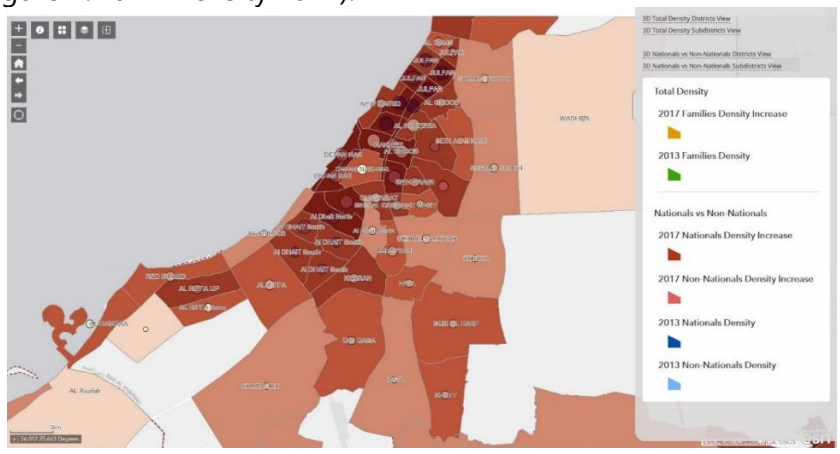


Figure 3: Families/ Residential Units Density Distribution (labelled with administrative unit name)

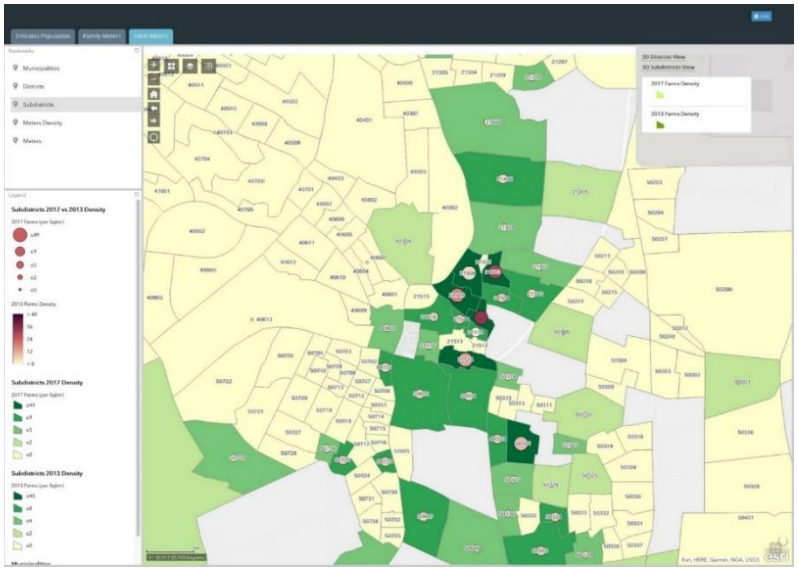


Figure 4: Farm Density 2017

Moreover, temporal growth between 2013 and 2017 was identified and mapped as shown in Figure 5: Families Density 2017 versus 2013 and Figure 6: Figure 3: Farms Density 2017 versus 2013.

The results visualisation harnessed various GIS tools and techniques were used including classified maps, 3D mapping, web app builder, story maps, etc.

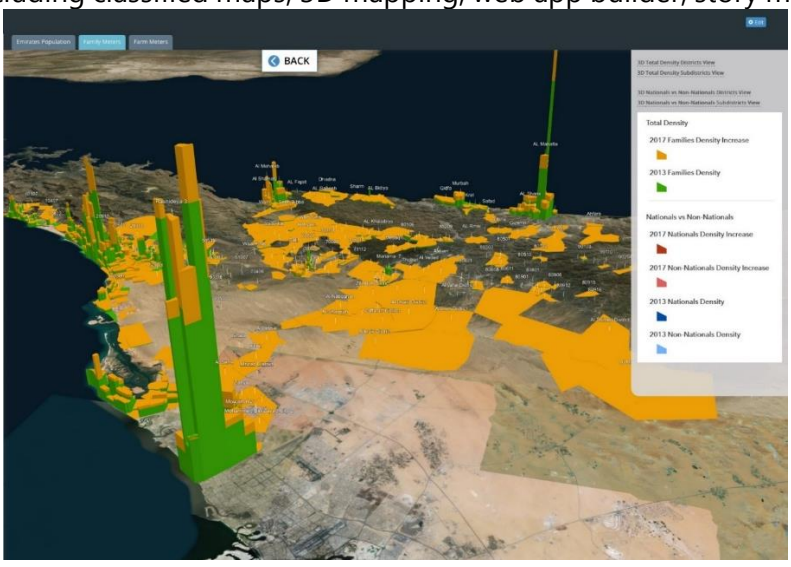


Figure 5: Families Density 2017 versus 2013

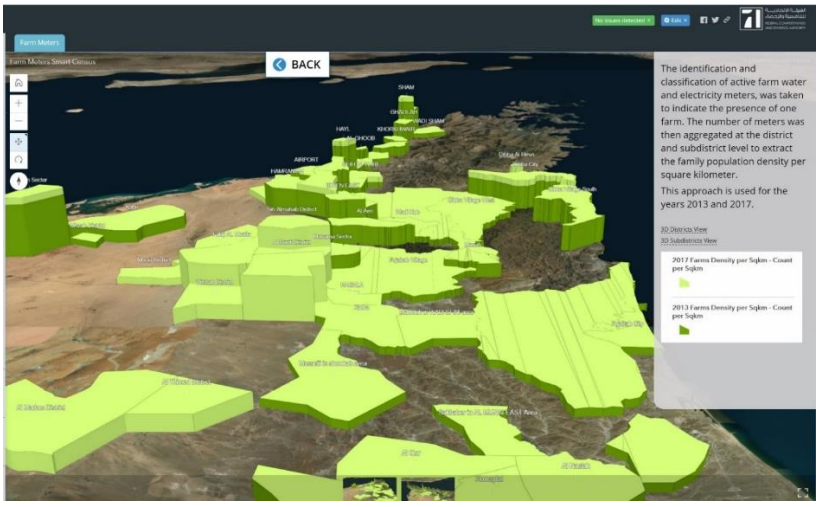


Figure 6: Figure 3: Farms Density 2017 versus 2013

4. Discussion and Conclusion

In the era of data revolution and digital transformation, the United Arab Emirates government strategies have incorporated initiatives to accommodate digital revolution and leverage technology capabilities.

On the other hand, Open Government Data and Government Information Sharing (GIS) initiatives are notably promoted and implemented within the UAE federal entities. The initiatives' objectives are to promote data sharing and

integration between governments, private institutions and citizens (Choi et al, 2013). The information sharing significance across governmental entities varies from sharing services, enhancing governmental services and end user experience, supporting a decision aid tool, planning assets allocation, to supporting emergency management systems.

As a pillar for the digital transformation journey, most federal entities in the United Arab Emirates have established data management systems for their assets and facilities, which in turn feed operational processes and support decision makers. Moreover, this project is a great example of digital transformation and Government Information Sharing initiatives within the UAE governmental entities.

FEWA as the federal authority for electricity and water have an established meters data management system which is quite resourceful in the amount of details availed per meters, like meter type, location, activity, owner name, owner nationality, unit land use, serial, etc. This meters data was part of Government Information Sharing initiative where FCSA diverted the utilisation of smart meters data to employ it in extracting household information and farms enumeration in a cost effective approach while addressing the below goals:

- Providing the distribution of household and farm density in the country in the form of electronic maps and geospatial location.
- Providing up-to-date data for decision makers and policymakers.
- Analysing temporal household and farm density data to visualise annual growth for decision makers.
- Providing interactive maps enriched with geospatial visualisation techniques to give a better end user experience for the stakeholders.
- Presenting a resource efficient practice where the least resources were acquired to implement a census project.

References

1. J. Choi, S.A. Chun, D.H. Kim, and A. Keromytis. (2013, June). SecureGov: secure data sharing for government services. In Proceedings of the 14th Annual International Conference on Digital Government Research 2013 Jun 17 (pp. 127-135)
2. K. Gajowniczek and T.Ząbkowski. (2015). Data Mining Techniques for Detecting Household Characteristics Based on Smart Meter Data. In: Energies 2015, 8, 7407-7427. Retrieved from; https://www.mdpi.com/journal/energies

Index System and Evaluation Research of New and Old Kinetic Energy Conversion in Qingdao

Hou Hongwen, Zhou Mianxian, Xu Guicai
Qingdao Statistical Institute, China

Abstract
It is an important opportunity for the city to realize the transformative, innovative and initiative development by implementing the old and new kinetic energy conversion project. In order to give full play to Qingdao's comprehensive advantages like enormous economic strength, enrichment of innovative resources, make Qingdao lead in making breakthrough with radiative driving force, create the main engine of new and old kinetic energy conversion, and accelerate the overall layout of old and new kinetic energy conversion of "Guidance by Three Core Cities, Multiple Breakthroughs, Integration and Interaction", this study explores and constructs the statistical index system of new and old kinetic energy conversion, and adopts the quantitative measurement method of composite index to reflect the old and new kinetic energy conversion in the city.

Keywords
New and Old Kinetic Energy Conversion; Statistical Index System; Evaluation Research

1. Introduction

(I) Design principle

I. Scientificity. The theme of new and old kinetic energy conversion is focused, and cultivation of new kinetic energy is highlighted. The new kinetic energy covers the primary, secondary and tertiary industries, with the emphasis on technological innovation as guidance, new technologies, new industries and new business models and new patterns as the core, new factors of production such as knowledge, technology, information and data as the support, reflecting the development trend of new productive forces. The new kinetic energy is a powerful driving force for the development and upgrading of real economy. Specifically, the chosen indexes are representative, which can fully reflect the old and new kinetic energy conversion in the city.

II. Universality. The statistical monitoring index system of old and new kinetic energy conversion of the city is established by relying on the framework of statistical index system, fully considering the evaluation needs of kinetic

energy conversion, in line with the principles of maximum reference, appropriate replacement and partial supplement.

III. Feasibility. All indexes are clear-cut and can be described and analyzed quantitatively, which can be obtained directly from the current statistical data or through relevant measurement, to ensure the smooth monitoring and evaluation.

IV. Openness. Moderate dynamics and openness should be kept for the establishment of index system. The new-old kinetic energy conversion is a continuously innovative and transformational systematic project, and the monitoring system needs to be modified, supplemented and improved according to the progress of major projects.

(II) Basic structure

The statistical index system of old and new kinetic energy conversion consists of two levels. The first level is divided into six sectors: knowledge capability, economic vitality, innovation drive, digital economy, transformation and upgrading, development achievement, reflecting the old and new kinetic energy conversion from different dimensions; The second level is specific indexes: each domain is elaborated and synthesized through 4-9 indexes, and there are 40 indexes in total. In terms of knowledge capability, it is mainly reflected by indexes such as quality of laborers, professional and technical personnel, the scale and intensity of R&D personnel investment, etc.; In regard to economic vitality, it is mainly reflected by commercial system reform, opening-up, cultivation and fostering of scientific and technological enterprises and vitality of capital market, etc.; In the innovation- driven sector, it is mainly reflected by the investment scale of R&D funds, construction of incubators, and commercialization of research findings, etc.; In the aspect of digital economy, it is mainly reflected by indexes like e-commerce, informatization degree and new business models, etc.; In the aspect of transformation and upgrading, it is mainly reflected by new industries such as high-tech, strategic emerging industries, urban commercial complex and new pattern index. In terms of development achievement, it is mainly reflected by output efficiency, quality efficiency, and sense of gain, etc. The above six sectors are interrelated and progressive, with integrity, coordination and balance.

2. Methodology

(I) Processing by the same measurement

Because the different measuring unit and order of magnitude among indexes in the statistical index system of the old and new kinetic energy conversion, the indexes can be converted to dimensionless indexes with

complete comparability during the comprehensive evaluation of new kinetic energy index of economic development.

The commonly used methods of the same measurement are: relativization processing, standardization processing and efficacy coefficient method, etc. Because standardization processing and efficacy coefficient mainly focus on the lateral correlation evaluation, but this study mainly reflects the vertical development trend, the relativization treatment method is selected to calculate the index value.

Single index (positive index) ki = Single index value in reporting period I1/ Single index value in base period I0

Single index (inverse index) ki = Single index value in base period I0/ Single index value in reporting period I1

(II) Weight definition

The common methods to define weight are Delphi method, Analytic Hierarchy Process, Coefficient of Variation, Multiple Correlation Coefficient method, Entropy method, among which Delphi method and AHP method will be affected by subjective factors, and Coefficient of Variation method, Multiple Correlation Coefficient method and Entropy method need a large number of samples to determine the accurate data model. In view of the limited data and the equal importance of all aspects of kinetic energy conversion, this study adopts the method of equal-weighted assignment.

(III) Composite index measurement

Commonly used methods of index measurement are: composite index method, comprehensive score of target value and principal component analysis method, etc. Because the target value is difficult to define for kinetic energy conversion, and the sample size required by the principal component analysis method is large, this study mainly reflects the old and new kinetic energy conversion through the development and change of the composite index. The principle of "index of mean" is adopted in the determination of composite index. The calculation procedure and process are as follows: the value of single index is calculated first, and then the weighted average of single indexes (or sub-index) is calculated. The formula is:

$$k = \frac{\sum k_i \cdot w_i}{\sum w_i}$$

Where k is composite index, I_0 is index value in base period, I_1 is index number in reporting period, and W is weight.

3. Results

(I) Multi-dimensional view on new and old kinetic energy conversion

1. Overall, the new kinetic energy index continues to rise.

The monitoring results show that the new kinetic energy index of economic development in 2016 is 126.6% and 2017 150.9%, with a year-on-year increase of 24.3 percentage points based on 2015 as the benchmark period, which shows that the incubation of new kinetic energy is further accelerated. It can be predicted that in the next few years, with the accelerated implementation of the major projects of new and old kinetic energy conversion, the new kinetic energy index of economic development in the city will reach a new level.

2. The digital economy has the highest index by sectors

In terms of different sectors, the digital economic index has the highest level and the largest extent of increase, reaching 262.7% in 2017, with a 55.7-percentage-point increase year-on-year, which has become the primary factor of the new kinetic energy index in lifting economic development; The second is economic vitality, reaching 155.9% in index, with a 33.7-percentage-point increase year-on-year, which has become a powerful pusher of the new kinetic energy index for economic development. The third is transformation and upgrading, reaching 145.5% in index, with a 28.1-percentage-point increase year-on-year. In contrast, the index level of innovation drive, development achievement and knowledge capability is low, which is 121.7%, 107.3% and 105.1%, respectively, and the increase range is relatively small, which affects the new kinetic energy index of economic development to some extent, and more efforts need to be put forth for future development.

New kinetic energy index in different sectors

Sector	2016 (%)	2017 (%)	Increase (%)
Knowledge capability	103.7	105.1	1.4
Economic vitality	122.2	155.9	33.7
Innovation drive	109.0	121.7	12.7
Digital economy	207.0	262.7	55.7
Transformation and upgrading	117.4	145.5	28.1
Development achievement	102.6	107.3	4.7

3. In terms of contribution, the digital economy index contributes the most.
In terms of contribution rate, in 2017, digital economy accounts for the largest contribution rate in all sectors, at 29.2%, with a 2-percentage-point increase year-on-year; The second is economic vitality at contribution rate of 17.4%, with a 1.4-percentage-point increase year-on-year. The third is transformation and upgrading at contribution rate of 16.2%, with a 0.8-percentage-point increase year-on-year. The contribution rates of innovation drive, development achievement and knowledge capability indexes are 13.5%, 12% and 11.7%, respectively, decreasing to different extents on year-on-year basis.

Contribution rate of indexes in all sectors in 2017

Sector	Contribution rate (%)	Increase and decrease (%)
Knowledge capability	11.7	-1.9
Economic vitality	17.4	1.4
Innovation drive	13.5	-0.8
Digital economy	29.2	2.0
Transformation and upgrading	16.2	0.8
Development achievement	12.0	-1.5

(II) Factor analysis of improving the new kinetic energy index

1. Effective improvement of knowledge capability
Knowledge capability is the source of innovation and basis of economic development. In recent years, the city has vigorously implemented the strategy of rejuvenating the city through science, education and talent development, and actively introduced famous colleges and institutes to provide strong support for economic development. In the perspective of the monitoring indexes, in 2017, the high-skilled talents in the city have reached 240,000, with a 7% increase year-on-year; The proportion of population with master's degree or above in the economic activity has reached 1.1%, which is higher than the national average.
2. Significant increase of economic vitality
Economic vitality reflects the greatest potential for economic development. From the view of monitoring indexes, in 2017, the newly registered market players in the city have increased by 34.2% year-on-year; Technology business incubators have increased by 20.8% year-on-year; The number of

enterprises in the National High-tech Development Zone has increased by 67.9% year-on-year; Listedc enterprises at home and abroad have increased by 13.1% year-on-year; The actual utilization of foreign capital has increased by 13.9% year-on-year; Courier business volume has increased by 24.7% year-on-year, which shows that economic vitality is significantly increased.

3. Transformation boosted by innovation drive
   In recent years, the city has accelerated the implementation of innovation-driven development strategy, which has effectively boosted the transformation and upgrading of economic development. From the monitoring indexes, in 2017, the number of invention patents granted per 10,000 R&D personnel in the city has increased by 2.6% year-on-year; The accumulative graduated enterprises in the technology business incubator have increased by 13.3% year-on-year. The turnover of technology market has increased by 21.6% year-on-year; The number of applications for international registration of Madrid trademarks has increased by 176.7% year-on-year, ranking first in the country.
4. Strong drive of digital economy
   Digital economy is the typical representative of new economy. Networking, intelligence and digitalization promote the rapid development of "Internet +", and the digital economy has become an important driving force for economic and social innovation and development. From the angle of monitoring indexes, in 2017, the fixed Internet broadband access users in the city have increased by 25.4% year-on-year; The total amount of telecommunication services has increased by 72% year-on-year; With e-commerce transactions doubled, the online retail sales of enterprises above designated size wholesale and retail have accounted for 17.8% of the total retail sales of enterprises above designated size, with a 2.6-percentage-point increase year-on-year; The agricultural informatization rate has reached 62%, with a 4-percentage-point increase year-on-year, which shows that digital economy plays a strong leading role in economic development.
5. Continued transformation and upgrading
   Transformation and upgrading is an indispensable and important aspect in promoting economic development. From the perspective of monitoring indexes, the economic transformation and upgrading of the city have been continued in 2017. The energy consumption per unit GDP has decreased by 3.99%, and the proportion of non-water renewable energy power generation has increased from 6.8% to 8.6%; New industries, new business models and new patterns have been developing vigorously, and the added value of strategic emerging industries has accounted for 10% in GDP; The added value of high-tech service industry has increased by 11.2% year-on-

year, with an increase of 2.6 percentage points; The proportion of the four types of enterprises above designated size selling goods or services through e-commerce trading platform has reached 12.1%, with a 3.2-percentage-point increase year-on-year. The number of merchants in the urban commercial complex has continued to increase, with a 24.3% increase year-on-year, which has increased by 16.6 percentage points; Exports of high-tech products has changed from decrease to increase, with a 11.6% increase year-on-year. Investment in technological upgrading has increased, with industrial technology upgrading accounting for 65.6% of industrial investment, which has a 15.5-percentage-point increase year-on-year, and traditional kinetic energy has released new vitality.

6. Hopeful future of development achievement

   Development achievement is the result of economic development and old and new kinetic energy conversion. From the monitoring indexes, in 2017, the total labor productivity of the city has increased by 10.4% year-on-year; New economic added value has accounted for more than a quarter of GDP. Marine economic added value has accounted for 26.4% of GDP, and cultural industry added value has accounted for nearly 6% of GDP. The new economy and characteristic economy has gained considerable development. The real economy has supported steady fiscal growth, with tax revenue accounting for 71.2% of the general public budget, increasing from the previous 69.2%; People's sense of gain has been further enhanced, and the annual per capita disposable income of urban and rural residents has increased by 8.6% year-on-year, which is significantly faster than the economic growth.

4. Discussion and Conclusion

(I) The market foundation needs to be further consolidated.

Overall, the number of market players in the city has been increasing in recent years. However, in recent years, especially with the start of commercial system reform, the newly registered market players has been in a downward trend, which needs to be paid attention to in the future development.

(II) Innovation drive need to be further highlighted

As a whole, the city's innovation-driven capacity has been improving, but attentions should be paid to the slowdown of increase of some indexes. The perspective of R&D personnel investment needs to be strengthened in the future development.

(III) Emerging industries need further development

The proportion of new industries represented by strategic emerging industries and high-tech industries is not high, which is lower than that of Beijing by 5.7 and 18.4 percentage points respectively.

(IV) Investment promotion needs further boosting

In 2017, the contractual foreign investment in the city has increased by 22.1% year-on-year; The actual utilization of foreign capital has increased by 13.9% year-on-year, and the disbursement rate of capital has dropped by 6.1 percentage points year-on-year. The actual utilization of domestic capital in the whole year has increased by 13.3% year-on-year. From the structure of actual utilization of foreign and domestic capitals, the proportion of manufacturing industry is not high, so the industrial investment should be increased to provide strong support for the development of real economy.

(V) The pressure of fiscal revenue increase needs to be further relieved

The proportion of tax revenue to fiscal revenue is one of the important indicators to measure the quality of economic operation. In recent years, although the city's fiscal revenue has maintained double-digit growth, but the growth rate of tax revenue has fallen back, and tax revenue is relatively low.

(VI) Some indexes need to be further improved

From the perspective of transformation and upgrading, the growth rate of strategic emerging industries and high-tech industries in the city is obviously lower than the average rate of the whole country and the whole province.

References

China Statistical Yearbook, *Shandong Statistical Yearbook*, *Qingdao Statistical Yearbook* and shared data among all provinces and municipalities.

Identifying preferred life insurance products using classification trees, multinomial logistic regression, and random forest

Jessa Luzelle S. Cuaresma, Francisco N. delos Reyes
University of the Philippines, Quezon City Philippines

Abstract

Targeting is one of the strategies implemented by marketers to achieve efficiency in the promotion of its products. In the life insurance industry, targeting can be useful to achieve higher penetration in the market. However, it is still a challenge to adopt this methodology in the industry given the limitation on the knowledge on how the available data can be utilized. This study aims to explore the application of several predictive modelling techniques in identifying client characteristics and behaviors that determine their preference on life insurance products. Models that are considered are Classification Trees (CART), Multinomial Logistic Regression (MLR), and Random Forest (RF). Results show that while the models are not capable of predicting minority life insurance products accurately, they are able to generate insights on the predictor relationships which can be used by marketers in crafting strategy for distribution, promotion, and product development. Such insights include the preference of Unit-Linked products for protection by an immediate family, and by the insured himself if he has high earnings.

Keywords

Life Insurance, Unit-Linked Insurance, Insurance Marketing, Predictive Modeling

1. Introduction

While sources of data increase and become more varied over time – from the traditional sources such as the policy management system to the more modern ones like health devices – insurance companies are challenged to come up with data-driven approach to carry out their initiatives for targeting. One method which increases its popularity in the industry is predictive modelling. This paper will be discussing about the application of several predictive modelling techniques in identifying client characteristics and behaviors that determine their preference on insurance products. The results will be used by the marketers and distributors alike to identify the products that can be best offered to clients based on their profiles which in return can increase their propensity to buy. In the process, predictive modelling methods that can handle multidimensional data sets with categorical output variables

will be proposed, the performance of the models above using actual insurance data set will be assessed and a scheme to match product type to client profile will be proposed.

2. Methodology

The data is comprised of 2,758 actual records of policies sold from a bancassurance company from 2014 to 2015. The data was up to time of extraction hence variables that changed since policy effectivity was not tracked. In actual setting it is possible for a client to purchase more than one plan. To ensure uniqueness of the observations, only the first policy purchase was kept in the data set. The first purchase also provides information on the priority of the client in purchasing an insurance plan.

There are seventeen (17) independent variables namely *Payment Mode, Life Insurance Coverage in PhP, Status Indicator, Age of Insured, Age of Policy Owner, Relationship Category, Owner Gender Indicator, Insured Gender Indicator, Owner Income, No. of Sick Family Member, Medical History, Insured Height, Insured Weight, X-ray Indicator, Indicators for Attached Accident, Critical Illness,* and *Hospitalization Riders.* The dependent variable is the *Type of Insurance (Unit Link, Endowment, and Other Traditional).* These variables are of several types - Nominal Categorical, Dichotomous, Continuous Numerical, and Discrete Numerical. These also indicate demographic, behavioral, and economic information about the insurance clients.

No variable considered contained missing data that is higher than 18% of the observations. For those with missing data, imputation method that is specific to each model was carried out. For CART, MLR, and RF, these are surrogate splits, Multivariate Imputation by Chained Equations (MICE), and Random Forest Imputation respectively. All of which and all evaluations moving forward were done in R.

In the modeling process, observations from 2014 was treated as in-time observations while those from 2015 was tagged as off-time. The 2014 data was further divided into training and validation sets. The former was used in the development of the model while the latter was used to fine tune the model to avoid overfitting. For all models, the dependent variables were balanced and distributed randomly between the training and validation data set.

To address potential overfitting, the 2014 data was grouped into the five splits: 80%-20%, 70%-30%, 60%-40%, and 50%-50%, which represent the splits between the training and validation data. Since CART performs cross-validation, the splits were translated to 2 to 5 folds and prediction errors were computed. The two other models were fitted on the training and validation data. The discrepancy between the training's and validation's residual deviance, and out-of-bag errors for MLR, and RF respectively were computed. The split which generated the least discrepancy of the computed values was

chosen for each method. These are 70%-30% or 4 folds for CART, 50%-50% for MLR, and 80%-20% for full model Random Forest and 50%-50% for Random Forest with CART variables which will be explained later.

CART is a supervised modelling technique that apply a recursive partitioning algorithm (Loh 2011). In the process, the data set is split into dichotomous groups such that the subsequent groupings would be more homogenous than the parent group (Gass et al., 2014). Let t be a parent node which contains the data set of interest. This will be further split to child nodes tl and tr which are the left and right nodes respectively such that each of the child nodes are more homogenous than the parent node. The split is based on the value of an independent predictor variable. The splitting for each of the child nodes and their resulting child nodes will continue. This will lead to the terminal node where the final classification is determined. (Gass et al., 2014)

CART is implemented in R using the *rpart* command and the default impurity function of which is the Gini Index (Therneau et al., 2018).

One of the risks encountered in modeling using CART is overfitting. Pruning is the process of cutting off branches that do not have a significant contribution to the predictive performance of the model (Gromping. 2009), One method used to prune a tree is cross-validation. This method aims to generate a tree that balances fit and complexity in an optimal manner. These are measured by the misclassification error cate R( ) and the complexity parameter α respectively (Timofeev, 2004), In mathematical terms, it is equivalent in choosing the optimal tree T such that $R\alpha(T) = R(T) + \alpha|T|$ is minimum where $|T|$=v is the number of terminal nodes and $R(T) = \sum_{i=1}^{v} P(T_i)\, R(T_i)$ is the overall misclassification rate of the tree computed from the sum of misclassifications from the v terminal nodes T_i, (Therneau et al., 2018). Computation of the complexity parameters is done via the *printcp* command in R while cross-validation is done using the *prune* command (Themeau et al., 2018). Aside from pruning. other parameters can also be manipulated such as the maximum number of observations per node, N_{stop}, and the application of the 1-standard error rule.

For MLR, the interpretation for the relationship is not straightforward. Instead, the likelihood of choosing the class of interest against a base is determined. For this study, *Unit Link* was defined as the base class. The odds ratio or the likelihood that the class of interest will be chosen over the base outcome is modelled as:

$$ln\left(\frac{\mathrm{P}(y = q' \mid x)}{\mathrm{P}(y = Unit\ Link \mid x)}\right) = B_{q',0} + x_{q',1} * B_{q',1} + \cdots + x_{q',n} * B_{q',n}$$

where y represents the dependent variable, *xi's* are the predictor variables where $i = 1, 2, \ldots, 17$ for the full model, q' is the class whose probability of choosing is evaluated, and *B's* are the regression coefficients. The function

shows how to compute for the log-likelihood that class *q'* will be chosen over *Unit Link* given the set of predictor variables *x*. (Hosmer and Lemeshow, 2000)

Fitting the model is equivalent to estimating the value of the vector B in the equation above such that the deviations of predicted values from the observed values are minimized. This was done by applying the method of maximum likelihood which is applied in R using the *multinom* function. To check how the results align with CART, MLR was also run to consider only the final predictors from CART. The predictors were transformed to factors. To assess the significance of a predictor the Wald's 2-Tailed Z-Test was conducted in R. The goodness-of-fit test was implemented using the Hosmer-Lemeshow Goodness-of-fit test through the *logitgof* function in R (Jay, 2017).

As an ensemble, Random Forest is a method that involves a collection of Classification Trees. Constructing the ensemble is an application of the bootstrap aggregation or bagging method. The first step is the repeated sampling in the observations via bootstrap such that the samples that will be generated will be used to construct the Classification Trees. To generate the overall prediction of the ensemble on the classification of an observation, a voting scheme is implemented to aggregate the classification outcomes of the Classification Trees. Each of the tree will cast out a vote based on its own classification and the votes will be tallied per class. The class with the highest number of votes will be considered the class for that specific observation. This process is done for all observations included in the analysis. Lastly, the Random Forest is assessed based on its capacity to predict the validation data. (Fawagreh et al., 2014; Gromping, 2009, Breiman, 2001, Breiman, 1996).

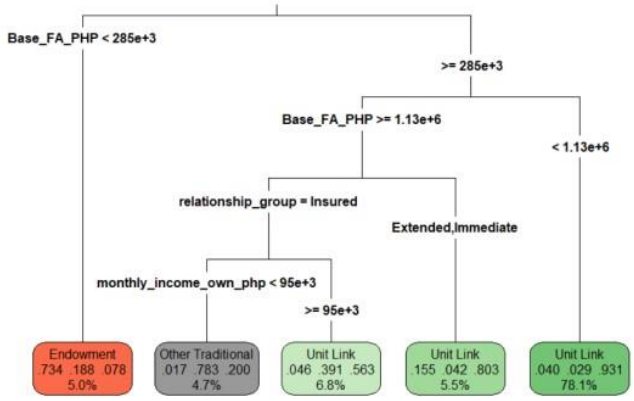
The development of the Random Forest in this study was done using the command *randomForest* in R. The number of Classification Trees *M* was set at *2,000* which is double the minimum value of *1,000* (Izenman, 2008). The number of choice predictors per node *mtry* is *2* which were computed as $0.5\sqrt{n}$ where *n* is the number of predictors (Izenman, 2008). Similar to MLR, RF was also run for the final CART predictors to assess if the predictor relationships will be aligned with the final CART model. In this case *mtry* is set at 2 which is the minimum.

3. Results

Illustration 1 shows the Final Classification Tree after applying $N_{stop} = 30$, the pruning method, and the 1-standard error rule. It was shown that Unit-linked products are often purchased with life insurance coverages that are higher than P285,000. From this result, it can be deduced that Unit-linked is the preferred class for clients who would like to have large protection. Also, for those plans with life insurance coverage that are higher than P1,130,000, it is only purchased by the client for himself if his income is more than P95,000, otherwise, the plan is purchased for him by his extended or immediate family.

It can be understood from this that higher protection is chosen if the loss is seen as more detrimental such as loss of a loved one or loss of significant income stream due to death.

Illustration 1: Final Classification Tree



For MLR, below summarizes coefficients of significant variables for the two models mentioned which are also large enough for useful interpretation.

Table 1: Coefficients for MLR (Full Model)

	3 Sick Family Members	NonStd Med History	with Hospital Rider	with Accident Rider
Endowment	0.81	-12.09	-14.76	20.76
Other Traditional	-17.27	-15.53	-18.95	21.56

Table 2: Model Coefficients for MLR (CART Final Predictors)

	Coverage => 500k but < 1M	Coverage below 500k	Insured is Owner	Owner Income below 50k	Owner Income => 50k but < 150K
Endowment	-1.30 (insig)	0.017 (insig)	-1.85	-0.38 (insig)	-0.65
Other Traditional	-3.61	-3.00	0.79 (insig)	1.61	-0.31

The goodness-of-fit test for the full and second model yielded a statistic $C = 22.25$ and 3.41, and a p-value of 0.13 and 0.99 respectively. This means that for both models, the null hypothesis will not be rejected and that there is no evidence that the model has a poor fit at 5% level of significance.

Calculating the misclassification for the models when applied to the validation data generated rates of 16.85% for the full model and 16.07% for

the second model. Both are within the threshold of 20%. However, since the latter has a smaller misclassification, this was considered the final model.

From the significant coefficients above, it can be interpreted that Unit-linked is preferred over Other Traditional if the amount of life insurance coverage shifts from between P1,500,000 and P2,000,000 to below P1,000,000. This indicates that aside from Unit-linked which was discussed earlier, Other Traditional is also preferred by clients who are interested in getting large protection coverage. However, it can also be seen that Other Traditional is preferred over Unit-linked if the income shifts from above P150,000 to below P50,000 which means that Other Traditional is the preference for the segment with lower income but would like to have a larger protection. This is reasonable since under Other Traditional falls Term Products which offer large insurance coverage at a lower cost than Unit-linked.

Another insight is that Unit-linked is preferred over Endowment if the income shifts to between P50,000 and P150,000 from above P150,000; and Unit-linked is preferred over Endowment if the insured is same as owner. These two when combined matches the insight from CART that clients that earn around P95,000 or higher prefers to purchase Unit-linked plans for himself.

For Random Forest, the estimated error rate for each of the combinations of *mtry* and *M* was computed for the validation data of the splits. The total misclassification rates are 12.45%, and 13.40% respectively. Since the full model has lower total misclassification, this was considered as the final model for Random Forest.

Unlike CART, Random Forest does not generate a structure that represents the relationship between the predictors. Instead, insights on the model are produced from the influence of the variable to the prediction. The importance of a variable is determined by the change in impurity in the prediction. This was performed using the *importance* function in R. Likewise, the partial influence of the most "important" variables were determined using Partial Dependence Plots. The results are shown below:

Illustration 2: Variable Importance

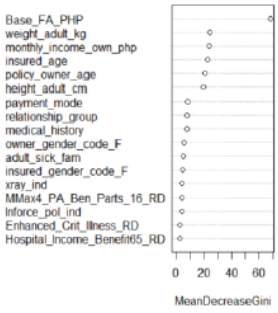
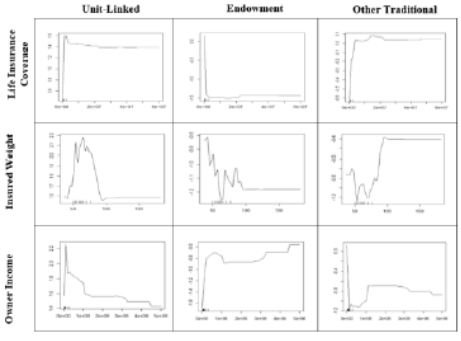


Illustration 3: Partial Dependence Plots



For the full model, it shows that more Unit-linked and Other Traditional policies are preferred if the amount of insurance is higher. Also, if the insured

has a higher weight, which may also indicate a decline in health, Other Traditional is more often bought. In terms of income, Unit-linked can also be preferred by low earning individuals.

To assess the accuracy of the models in predicting, the final models were used to predict the test data.

Table 3: Summary of Misclassification Rates for All Models (Test Data)

	Total Misclassification	High-risk Misclassification
Classification Trees	24.37%	12.83%
Multinomial Logistic Regression (CART Final Predictors)	17.52%	12.08%
Random Forest (Full Model)	18.81%	15.55%

Among the models, only the final models from Multinomial Logistic Regression and Random Forest passed the misclassification benchmark of 20% assuming total misclassification. However, if the high-risk misclassification was considered, which is the portion of Endowment and Other Traditional observations classified as Unit-Linked, Classification Trees can also be accepted at the threshold. The high-risk misclassification represents the conservative clients which may be offered with an aggressive investment insurance product without caution.

4. Discussion and Conclusion

All models are acceptable within 20% misclassification threshold if only the high-risk misclassification is considered. The models are also shown to be poor in predicting minority classes but good in classifying the dominating class which is Unit-linked. Given these results, the models cannot be used to profile the markets for Endowment and Other Traditional but were used to understand the market for Unit-linked better. The relationships that were generated from each of the models can be consolidated to come up with a targeting strategy to promote Unit-linked plans. Tapping the Endowment market as well to purchase Unit-linked is reasonable since the two products address the same need which is savings and that selling Unit-linked is more profitable for the company given the instability of the interest rate over the long-run.

Caution must be observed however in offering Unit-linked plans using the insights from the model since much observations from the minority classes are misclassified. Offering an aggressive financial tool to conservative clients might cause severe misselling. This risk can be controlled by full disclosure of the features of a Unit-linked product especially its non-guarantees.

There were a lot of insights that were derived from the models that can be used for the targeting strategy. For instance, Unit-linked plans can be offered

for clients looking for protection. They can be offered to the dependents of a breadwinner or to the owner himself especially if he is earning much and has a lot to lose. Other Traditional is also an option for protection push but it caters to the extreme segments in terms of income earnings. Hence, Unit-linked product as a protection solution can be used to target the middle class.

These results can be supplemented by qualitative studies that can validate the targeting strategy. For example, actual clients can be asked through a survey if they will be interested to purchase a Unit-linked product. Similarly, it can be checked if indeed the reason why higher life insurance cover is purchased is because of protection and not for higher premium which can generate higher investment. Either way, this quantitative study can be used as a basis and a starting point for a qualitative study.

References

1. Breiman, L. (October 2001). Random Forest. *Machine Learning, 45.* Retrieved from: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
2. Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24.* Retrieved from: https://link.springer.com/article/10.1023/A:1018054314350
3. Fawagreh, K., et al. (August 2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal, 2.* Retrieved from: https: https://www.tandfonline.com/doi/abs/10.1080/21642583.2014.956265
4. Gass, K., et al. (2014). Classification and regression trees for epidemiologic research: an air pollution example. *Environment Health Journal, 13.* Retrieved from: http://www.ehjournal.net/content/13/1/17
5. Gromping, U. (November 2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *American Statistical Association, 63.* Retrieved from: https://www.tandfonline.com/doi/abs/10.1198/tast.2009.08199
6. Hosmer, D., & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
7. Izenman, A. (2008). Modern Multivariate Statistical Techniques. New York: Springer Science+Business Media
8. Jay, M. (December 2017). *Goodness of Fit Tests for Logistic Regression Models.* Retrieved from: https://cran.r-project.org/web/packages/generalhoslem/generalhoslem.pdf
9. Loh, W. (January 2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery, 1.* Retrieved from: https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf
10. Therneau, T., et al. (January 2018). *An Introduction to Recursive Partitioning Using the RPART Routines.* Retrieved from: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf
11. Timofeev, R. (December 2004). *Classification and Regression Trees (CART) Theory and Applications.* Retrieved from: https://www.scribd.com/document/256433144/Classification-and-Regression-Trees-CART-Theory-and-Applications

Testing for independence on statistically matched categorical variables

Janna M. De Veyra
University of the Philippines Diliman

Abstract

In most instances, conducting a new survey is impossible due to time constraints and limited resources. Matching data sources has been used as a way to obtain a data set where all the intended variables are available. This paper proposes the use of the MCMC and the inclusion of random error in matching categorical variables as well as the application of bootstrap procedure in testing for their independence. A simulation study indicates that the test is most effective when the proposed procedures are all applied because combining all these procedures produces a correctly sized test that yields the highest power among all other proposed procedures combined.

Keywords

random error; mcmc; bootstrap; size; power

1. Introduction

Orazio, et.al (2006) defines statistical matching as a statistical procedure that aims to integrate two or more datasets characterized by the fact that the different datasets contain information on a set of common variables and variables that are not jointly observed and that the units observed in the data sets are different. The goal of this procedure is to derive a synthetic data and to estimate the joint distribution of the variables that are not jointly observed in a single data set. The need for this type of procedure increases when the chance of conducting a new survey is almost impossible in a given time frame and resources. This paper deals with matching procedures in the categorical data to test for the independence of the two variables that are not jointly observed. Seltman (2015) mentioned that the usual statistical test in the case of categorical outcome and a categorical explanatory variable is whether or not the two variables are independent. Matching procedures used were regression imputation, stochastic imputation, and an application of MCMC in those two imputations. A check for independence on the four imputation procedures will be made using the Chi-square statistics. An application of bootstrap method under the four imputation procedures will also be considered to identify if this will produce a more reliable result in the test for independence.

2. Methodology

In this paper two data sources were matched, one was labelled as source A composed of variables $x1$, $x2$ & y and the other one was labelled as source B composed of variables $x1$, $x2$, & z. The goal in this case is to test for the independence of y and z. The variables that are common in both sources which are $x1$ and $x2$ are essential in performing the test. The following steps are written to be able to provide a clearer illustration on how the test will be conducted:

Step 1: Z variable will be declared as missing in source A while y variable will be declared as missing in source B

Step 2: In source A, identify the relationship of y variable in $x1$ and $x2$ variables. The same goes with source B

Step 3: Impute the missing values based on the values of $x1$ and $x2$ and on the relationship of the missing data to $x1$ and $x2$ obtained from Step 2.

Step 4: Combine the two data sources so the total sample size will be n=$nA+nB$.

Step 5: Compute for the Chi-square statistics of y and z in the combined data source to test for their independence

2.1 Matching Methods

This part discusses the different matching procedures that were considered in deriving a synthetic data.

2.1.1 Regression Imputation

Logistic regression will be used in matching two data sources where the variables of interest are categorical. Imputed values will be based on the probability of success. In the case when the variable of interest is binary, the missing value will be imputed as 1 when the probability of success obtained from a logistic regression is greater than or equal to 0.5 and 0 otherwise. On the other hand, when the variable of interest contains more than two categories, the probability of obtaining each category that is based on log-odds model will be computed and the missing value will be imputed based on the highest probability.

2.1.2 Markov Chain Monte Carlo

This type of procedure was used in the application of parameter estimation in the logistic distribution. In this paper, the MCMC sample from the posterior distribution of a logistic regression model using a random walk Metropolis algorithm. This type of algorithm is an iterative algorithm and produces a Markov Chain and permits empirical estimation of posterior distributions.

2.1.3 Stochastic Imputation

This procedure uses the estimates from the regression imputation and the markov chain monte carlo procedure in imputing missing value by adding a random error in the model. Imputation equation can be written as

$$\text{Logit}[\boldsymbol{\Pi}(\text{X})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \boldsymbol{\epsilon}$$

where α, β_1, and β_2 are estimates obtained in the procedures used and $\boldsymbol{\epsilon}$ is the random error. In this paper, the random error is assumed as normally distributed with mean 0 and variance 3 in imputing z value in source A and normally distributed with mean 0 and variance 4 in imputing y value in source B. Same as with the regression imputation, missing values will be imputed based on the probability of success.

2.2 Bootstrap Procedure

Bootstrap in this paper could be done in two ways. It could be done by either resampling within the synthetic data or resampling across the synthetic data. Resampling within synthetic data refers to resampling with replacement from a combined data source where estimates on the missing data are already available while resampling across the synthetic data refers to resampling with replacement from a combined data source where estimates on the missing data are not yet available. Both procedures were considered in this paper to check for a possible difference in the result. Resampling would be done 200 times in both procedures where the number of samples in each pseudo sample is just the same as with the original sample

2.3 Evaluation of the Procedure

The performance of the proposed procedures in testing for the independence of statistically matched categorical variables was evaluated according to the following:

1. To evaluate if the proposed procedures correctly reject the null hypothesis, the power of the test will be computed by simulating data set 200 times wherein each replicate has a Chi-square statistic that is greater than the critical value at a 0.05 level of significance. The proposed procedures will then be applied and the Chi-square statistics in the applied procedures will be computed in each replicate. The computed power will be the total number of replicates with Chi-square statistics obtained from the proposed procedures that is greater than the critical value at a 0.05 level of significance divided by the total number of replicates.
2. To evaluate if the proposed procedures falsely reject the null hypothesis, the size of the test will be computed by simulating data set 200 times wherein each replicate has a Chi-square statistic that is less than the critical value at a 0.05 level of significance. The proposed procedures will then be applied and the Chi-square statistics in the applied procedures will be computed in each replicate. The computed size will be the total number of replicates with Chi-square statistics obtained from the

proposed procedures that is greater than the critical value at a 0.05 level of significance divided by the total number of replicates.

2.4 Simulation Design

A simulation study was conducted to assess the efficiency of the procedures mentioned in the previous part of the paper in matching categorical variables. The simulation will be done by generating a data source where all $x1$, $x2$, y, and z variables are available. The y and z variable will be generated from a logistic regression

$$\begin{cases} \text{y: logit}\left[\Pi\left(x_y\right)\right] = \alpha_y + \beta_{1y}x_1 + \beta_{2y}x_2 + \varepsilon_y \\ \text{z: logit}[\Pi(x_z)] = \alpha_z + \beta_{1z}x_1 + \beta_{2z}x_2 + \varepsilon_z \end{cases}$$

where x_1 and x_2 are the covariates that are either categorical or continuous depending on the scenario and are generated by either assigning a known probability in each category or from a normal distribution with mean μ and variance σ^2 and ε is a random residual that is generated from a normal distribution with mean 0 and variance σ^2 . However, in some scenarios where the covariates are continuous, x_2 as a function of x_1 was considered. When the variables of interest are binary, the assigned value for y and z will be based on this cut-off

$$\text{y or z} = \begin{cases} 1 \; if \; \Pi(x) \geq 0.5 \\ 0 \; if \; \Pi(x) < 0.5 \end{cases}$$

when the variables of interest have 3 categories, the assigned value for y and z will be based on this cut-off

$$\text{y or z} = \begin{cases} 1 \; if \; \Pi(x) in \; [0.67,1] \\ 2 \; if \; \Pi(x) in \; [0.34,0.67] \\ 3 \; elsewhere. \end{cases}$$

when the variables of interest have 5 categories, the assigned value for y and z will be based on this cut-off

$$\text{y or z} = \begin{cases} 1 \; if \; \Pi(x) in \; [0.80,1] \\ 2 \; if \; \Pi(x) in \; [0.60,0.80] \\ 3 \; if \; \Pi(x) in \; [0.40,0.60] \\ 4 \; if \; \Pi(x) in \; [0.20,0.40] \\ 5 \; elsewhere. \end{cases}$$

Simulation consists of cases when y and z are independent and when y and z are dependent according to the chi-square statistics. This is based on the values set on α_y, α_z, β_{1y}, β_{1z}, β_{2y}, β_{2z}, ε_y and ε_z. The simulated data will serve as a benchmark in assessing the performance of the procedures in the test for independence.

Sample size used in this study is 1000. Source A and source B will be generated by splitting the simulated data into two and deleting variable z in one data and variable y in another data. Source A will be the data that contains x_1, x_2, and y variable while source B will be the data that contains x_1, x_2, and z

variable. Ratio of source A to source B that will be considered are 50-50, 70-30, and 90-10.

3. Results

The performance of the procedures for testing independence will be evaluated by the size and power of the test. For the purpose of this study, a test will be considered as correctly sized when the computed size is at most 0.05. Discussions of the analysis will be divided into sections concerning the number of categories of the variables being matched.

3.1 Dichotomous y and z

When the variables of interest to be tested are both binary, the test will only be correctly sized when random error was included in the imputation and when the bootstrap procedure was applied in the test. From the two resampling methods tested on this study, resampling within synthetic data has a higher power than resampling across synthetic data. The power of the test would also slightly increase when the MCMC procedure was applied in the imputation.

Table 1. Average Size and Power of the Test when y and z are dichotomous

	Regression Imputation			MCMC Regression Imputation			Stochastic Imputation			MCMC Stochastic Imputation		
Evaluation	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across
Power	0.61	0.51	0.33	0.61	0.51	0.33	0.92	0.58	0.43	0.93	0.60	0.44
Size	0.71	0.52	0.11	0.70	0.52	0.11	0.10	0.00	0.00	0.10	0.00	0.00

3.2 Both y and z have 3 categories

Similar as in the case when both y and z are binary, the size of the test in this case would only be correctly sized when random error was included in the imputation and when the bootstrap procedure was applied in the test. In addition, resampling within synthetic data has a higher power than resampling across synthetic data. This would then slightly increase when MCMC was applied in the imputation.

Table 2. Average Size and Power of the Test when both y and z have 3 categories

	Regression Imputation			MCMC Regression Imputation			Stochastic Imputation			MCMC Stochastic Imputation		
Evaluation	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across
Power	0.64	0.58	0.46	0.65	0.58	0.47	0.87	0.56	0.27	0.88	0.57	0.28
Size	0.85	0.52	0.11	0.70	0.52	0.11	0.10	0.00	0.00	0.10	0.00	0.00

3.3 Both y and z have 5 categories

Unlike in the first two sections, the size of the test when both y and z have 5 categories will be correctly sized even without the application of the bootstrap procedure in the test when random error was included in the imputation. Though the application of bootstrap procedure would still produce a correctly sized test upon the inclusion of random error in the imputation, the test has a higher power when bootstrap procedure was not applied. Similar as in the first two sections, power in this case slightly increases when MCMC was applied in the imputation.

Table 3. Average Size and Power of the Test when both y and z have 5 categories

	Regression Imputation			MCMC Regression Imputation			Stochastic Imputation			MCMC Stochastic Imputation		
Evaluation	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across	w/o Bootstrap	Bootstrap within	Bootstrap across
Power	0.61	0.53	0.43	0.61	0.54	0.43	0.80	0.61	0.38	0.82	0.61	0.38
Size	0.55	0.39	0.16	0.54	0.39	0.16	0.03	0.00	0.00	0.03	0.00	0.00

4. Discussion and Conclusion

This paper analyzed the effect of matching categorical variables to test for their independence using different simulation scenario on these matching techniques (1) Logistic Regression, (2) MCMC on Logistic Regression, (3) Logistic Regression with the inclusion of random error, and (4) MCMC on Logistic Regression with the inclusion of random error. The test here is to be obtained by (1) not resampling from the synthetic data (without bootstrap), (2) resampling with replacement within the synthetic data (bootstrap within), and (3) resampling with replacement across the synthetic data (bootstrap across). These methods were evaluated by computing for the size and power of the test.

Simulation shows that the use of MCMC slightly increases the power of the test. The increase in power can evidently be seen when random error was included in the imputation model. Bootstrap, on the other hand, produces a correctly sized test when applied in an imputation procedure that has a random error in the model. Among the two bootstrap approaches that were considered, bootstrap within yields a higher power than bootstrap across. However, when the variables of interest both have 5 categories, the test is already correctly sized upon the inclusion random error in the model even without applying the bootstrap procedure. Hence, it is safe to say that the bootstrap procedure is more useful when the number of categories for both y and z is small.

References

1. D'Orazio, M., Scanu, M., Zio, M. (2006). Statistical Matching Theory and Practice. England:Wiley
2. Seltman, H. (2015). Experimental Design and Analysis. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

A research on indicator system and evaluation of high-quality development in Qingdao

Yuan Zhenqiang, Leng Jiaqi, Wei Keai
Qingdao Statistical Institute, China

Abstract

According to the current research situation at home and abroad, there are more qualitative, local and individual descriptions of high-quality development, while quantitative, comprehensive and complete descriptions are still under discussion. This thesis attempts to give a quantitative comprehensive evaluation of high-quality development based on the new development concept and the development quality and efficiency.

Keywords

High-quality Development; Indicator System; Comprehensive Evaluation

1. Introduction

The key for promoting high-quality development is to follow the requirements of new development concept, take supply-side structural reform as the main task, and promote the reform of the quality, efficiency and driving force of economic development. To build a modern economic system and promote high-quality economy development, it is urgent to accelerate the formation of indicator system, policy system, standard system, statistics system, performance evaluation and performance appraisal system to promote high-quality economic development and create and improve the institutional environment. Given all that, we should give full play to the advantages of familiar standards and indicators, take the initiative to mine data resources, and quickly build a high-quality statistics indicator system and evaluation methods, in order to better serve the construction of modern economic system and high-quality economic development.

2. Methodology

1.Design Principle

1.1 Orientation. We should take new development concept as the guidance and work out to achieve the targets and tasks of supply-side structural reform in a bid to reveal the current situation of quality reform, efficiency reform and driving force reform in the economic development.

1.2 Comprehensiveness. The indicators cover a vast field of economic and social development, taking into accounts both economic development and people's livelihood, as well as the national level, enterprises and residents.

1.3 Comparability. Most of indicators we choose are relative. We should give up the thought of just laying stress on the differences of absolute amounts like the total number of economy, population scale and geographic area. By doing so, we aim to make benchmarking analysis and emphasize the quality and efficiency differences caused by unbalanced and insufficient development.

1.4 Operability. We attempt to improve the availability of indicators and focus on the publicity and authority of data sources. Data should be open and continuous as far as possible to ensure the objectivity of evaluation results.

1.5 Openness. The indicator system includes objective and subjective indicators to reflect the dynamic and open nature of the evaluation. It not only considers the common indicators in the fields of economic, social and people's livelihood, but gives consideration to the reflection of Qingdao's economic and social characteristics. We should base on current situation, focus on long-term development and timely make appropriate adjustment, supplement and improvement.

2. Basic Framework

The statistics indicator system of Qingdao high-quality development is built based on the requirements of high-quality development and the new development concept, in which macro and micro development quality and efficiency are taken as the priorities. The indicator system is divided into three levels. The first level is composed of 6 primary indicators, mainly focusing on innovative development, coordinated development, green development, open development, shared development and development quality and efficiency. The second level is composed of 50 secondary indicators, mainly reflecting specific matters of the primary indicators. Meanwhile, by taking consideration of the requirements of nation, provinces and targeted analysis, it sets up 6-12 secondary indicators under the primary indicators to make further detailed explanation to them. The basic framework is as follows:

2.1 Indicator group of innovative development. It mainly reveals the deep integration of science and technology with economy, the ability of science and technology to create a new economy, the ability of fostering new growth drivers to speed up traditional industry upgrading, the situation of the nationwide business startup and innovation drive. Besides, we need to release new demands and create new supply. The group consists of seven secondary indicators, including R&D funds, PCT international patent application, investment of technological reform, strategic emerging industry and added value of high-technology industry.

2.2 Indicator group of coordinated development. It mainly focuses on the cooperativity, integrity, inclusiveness and openness of regional economic development. We should also take efficient measures to improve the integrated mechanism of urban-rural development, promote the synchronized development of new industrialization, informationization, urbanization and agricultural modernization, and uplift the responsiveness of enterprises to market changes and demands and make timely adjustment. The group consists of seven secondary indicators, including urbanization rate, resident consumption rate, non-performing loan rate and the proportion of marine economic added value.

2.3 Indicator group of green development. It is mainly reflected in adhering to the basic state policy of conserving resources ad protecting the environment, adhering to sustainable development, firmly following the path of civilized development featuring production development, affluent life and sound ecology, accelerating the construction of a resource-conserving and environment-friendly society, and forming a new pattern of modernization featuring harmonious development between man and nature. It has eight secondary indicators, including the reduction rate of major pollutant emissions, the reduction rate of energy consumption per unit of GDP, the death rate of safety accidents, and new energy power generation.

2.4 Indicator group of open development. It mainly reflects the trend of deep integration of the economy into the world economy, the development of a higher level of open economy and the construction of a broad community of interests. It has six secondary indicators, including the proportion of total international trade, the proportion of export of high-tech products, the utilization of foreign capital and international air routes.

2.5 Indicator group of shared development. It mainly reflects the increase in the supply of public services, the implementation of poverty alleviation projects, the establishment of a fairer and more sustainable social security system, and the people's demand for a better life, sense of gain, happiness and other aspects. It has 10 secondary indicators, including the unemployment rate, basic living allowances, the proportion of people covered by old-age insurance and medical insurance, housing per capita, life expectancy and business environment.

2.6 Indicator group of development efficiency. It is mainly reflected in meeting people's growing needs for a better life, improving the quality, efficiency, fairness and sustainability of the economy, and reducing the cost of the real economy. It consists of 12 secondary indicators, including GDP per capita, GDP per mu, labor productivity, the proportion of added value of the "new economy", the proportion of tax revenue, enterprise asset-liability ratio and profit rate.

The third level consists of 120 tertiary indicators, which mainly decompose and interpret the primary and secondary indicators from the perspectives of industry, region, urban and rural areas and categories, and give full consideration to the reflection of Qingdao's characteristics, so as to facilitate further detailed analysis.

3. Evaluation Method

3.1 Target value evaluation method. Set up target value for every indicator and compare it with actual value in report period. And then calculate the realization degree of individual and graded indicators and finally get corresponding evaluation value.

3.2 Efficacy coefficient evaluation method. Define threshold value for every indicator (including satisfaction value and not-allowable value) taking satisfaction value as the upper limit and not-allowable value as the lower limit, and then calculate the realization degree of satisfaction value of every indicator and accordingly determine their grades. Finally, the weighted average is used for comprehensive evaluation and ranking.

3.3 Principal component analysis method. Use dimensionality reduction method to transfer multiple indicators into a few of comprehensive ones (i.e. principal component) which contain mutually non-repetitive information and could reveal most of information of original variables. While introducing various variables, the complex factors are reduced to several principal components to simplify the problem and obtain more scientific and effective results.

3.4 Comprehensive index evaluation method. According to the principle of "average index method", the index of single indicator is calculated by relative processing method first, and then the weighted average of single index (or sectorial index) is formed to form comprehensive index.

3.5 Expert rating evaluation method. Invite several experts to score evaluation objects and make comprehensive evaluation through back-to-back method, and make an expected judgement on the future development trend.

4. Method Selection

This research mainly selects the comprehensive index method and records the changes of index to reflect high-quality development status. Besides, we innovatively introduce contribution rate to measure the contribution share of every sector, during which we collect relative data in recent three years and individually calculate individual index, comprehensive index and contribution rate taking 2015 as the base period. Measurement procedures are as follows:

4.1 Treat as a measure of the same factors

There are different units of measurement and orders of magnitude of every statistical indicator in statistical index system of high-quality development. Thus, every indicator needs to be nondimensionalized when doing comprehensive measurement with index, so as to make every indicator turn into dimensionless value with complete comparability. Common treatment as a measure of the same factors includes relative treatment, standard treatment and efficacy coefficient method. Of these, standard treatment and efficacy coefficient method mainly focus on crosswise comparison evaluation, yet, this study primarily reflect lengthwise development trend, so we choose relative treatment method to calculate index value.

$$\text{Single index (positive indicator)} k_i = \frac{\text{single index value in report period } I_1}{\text{single index value in base period } I_0}$$

$$\text{Single index (negative indicator)} k_i = \frac{\text{single index value in base period } I_0}{\text{single index value in report period } I_1}$$

4.2 Weight Definition

Frequently-used weight definition methods include Delphi Method, Analytic Hierarchy Process, Coefficient of Variation Method, Multi-correlation Coefficient Method and The Entropy Method, among which the first two may be affected by subjective factors, while the latter three need plenty of samples to define accurate data model. Five aspects of the new development concept play the same significant role in promoting economic and social development, so they need to be weighted in a balanced way. Since the final efficiency of high-quality development is reflected by development achievements, the weight of development achievement indicator should be higher than other first-tier indicators. Six primary indicator weight are set up by taking consideration of these factors and following the principles of focusing on clear priorities over development achievements and balancing five development sectors. Further, we will apply equal-weighted assignment method towards secondary indicators under the primary indicators in order to reflect the same significance of several high-quality development indicators.

4.3 Comprehensive Index Measurement

The principle of "average index method" is applied to define comprehensive index. The calculating procedures are as follows: firstly, calculate indexes of single indicator; then, treat these indexes (or sectorial indexes) in a way of weighted average. The calculation formula is:

$$k = \frac{\sum k_i \cdot w_i}{\sum w_i}$$

'K'refers to comprehensive index, I_0 refers to index value in base period, I_1 refers to index value in report period and 'w'refers to weight.

3. Result

1. Evaluation Result

1.1 Evaluation Result Obtained through Comprehensive Index

As shown by the calculation result obtained in the way of comprehensive index, we can see that the high-quality development index in Qingdao in 2016 is 105.9% compared with that in 2015, while in 2017 the index reaches 106.03%, rising 0.13 percentage point on a year-on-year basis, which reflects a steady increase.

In light of category, the open development index reveals the highest level and the biggest rising amplitude, which becomes the primary factor for increasing the high-quality development index. Other indexes in 2017 ranked from the highest to the lowest are green development index, innovative development index, coordinated development index, shared development index and development achievement index.

In regard of contribution rate, the highest is open development index which also has the biggest increasing magnitude. Other indexes ranked from the highest to the lowest by the contribution rate are green development index, innovative development index, coordinated development index, shared development index and development achievement index.

1.2 Evaluation Result Obtained Through Expert Rating

We invited experts from the Statistical Expert Consultation Committee to additionally check the evaluation result obtained through comprehensive index. These experts are from colleges, scientific research institutions and large corporate groups of Qingdao and own high reputation in economic and statistic research fields, so their evaluation results feature some degree of authority and representativeness.

We carried out a questionnaire survey on 'what level do you think the high-quality development of Qingdao stands'. From the survey results, we found that people choosing 70-80 points, 80-90 points and 90-100 points respectively accounted for 61.1%, 38.9% and 0, and the comprehensive evaluation point were 78.9. This means that high-quality development stands in the lower overall-good level. The result from another survey about 'how did the high-quality development in Qingdao in recent three years change' showed that people choosing increasing and impartial trend respectively accounted for 22.2% and 77.8%. While collecting the survey result of 'the expectation over the future status of high-quality development in Qingdao', we found that people selecting increasing and impartial trend respectively

accounted for 72.2% and 27.8%, which reflected that they take a positive attitude toward the future status of high-quality development in Qingdao.

2. Factor Analysis

The fluctuation of comprehensive index is caused by the collaborative effect of various concrete indexes, in which positive factors play a leading role. The specific analysis procedures are listed as follows:

2.1 Opening-up fields have been increasingly expanded.

In recent years, with the commitment to the Belt and Road Initiative and the strategy of going abroad and bringing home, Qingdao has expanded open fields of the service sector, reduced negative lists, uplifted international trade level and increased international routes. Thus, our open-up level has significantly improved and economic development has gained new momentum. In 2017, our city increased the proportion of total amount of international trade to GDP by 3.9 percentage points on the year-on-year basis, and raised the marketing amount of overseas investment trade by 12.9% from a year earlier. The proportion of export amount of high-tech products to that of goods witnessed a growth of 0.5 percentage points on the year-on-year basis. The amount of actually-used foreign capital increased 2.6 percentage points from a year earlier. Besides, our city has also expanded international routes, greatly uplifting our open-up level and international reputation.

2.2 New momentum for economic development has been accumulated.

In recent years, our strategic emerging industries, high-tech industries and technological reform investment have steadily increased and new economic development momentum has accumulated increasingly. In 2017, high-tech industries in Qingdao raised added value by 10.9%, while strategic emerging industries increased added value by 10.9%. New industries release new momentum, strengthening the internal driving force of economic development.

2.3 The coordination of industrial development has been strengthened.

The coordination of industrial development is mainly evidenced by the structural optimization of three industries and the industry diversity. In recent years, the industrial structure of Qingdao has increasingly optimized, the proportion of added value of the service sector to GDP has increasingly strengthened, and the proportion of added value of modern service sector to GDP has also expanded. Meanwhile, the industrial structure of the service sector has also upgraded continuously.

2.4 Green development promotion measures have been strengthened.

The green development has obtained significant achievements after pollution control and emission reduction, ecological protection and supervision of environmental protection were carried. In 2017, the emission of major pollutants declined consistently. For example, the reduction rate of SO_2 reached to 30%, and the proportion of generation amount of new energy to the total generation amount accounted for 7.45%, rising 2 percentage points on the year-on-year basis. The coverage rate of forests was kept above 40% and the success rate of water quality in collective drink water resources successively remained 100%. All these achievements play a vital role in promoting green development.

4. Discussion and Conclusion

There are some shortcomings in high-quality development, especially compared with national advanced cities.

1. R&D investments are insufficient and the overall soft strength for high-quality development is limited. In light of R&D investment and a macro background of innovation-driven development and technology-driven industry, the proportion of R&D investments to the GDP of Qingdao is insufficient. The number of invention and patent applications in recent three years has decreased year by year, which reflects that the philosophy, capability and funds in technological innovation have not yet matured.

2. Industries with high-pollution, high energy-consumption and high emission take higher proportion in real economy. The strategic orientation and policy docking of high-quality development are currently inaccurate. There are sufficient plans, strategies and policies in Qingdao's industrial development, yet departments and sectors usually go with their own pace and way, causing large dispersion of plans, strategies and policies.

3. The fairness of private economy and market competition are yet to be perfected, and the market vitality for high-quality development still needs to be improved. The market vitality in Qingdao is insufficient and private economy is not strong enough. There are still some phenomena hindering market unification and fair competition in some industries and fields.

4. There is a significant gap between urban and rural development. The degree of coordination of high-quality urban development still needs to be uplifted. Qingdao's main difference between urban and rural development reflects on the development achievement of non-urban areas, namely the imbalance between urban and rural areas. Therefore, the utilization rate of

city resources is limited, the average productivity on every mu of vast hinterland is low and the industry layout is not ideal.

References

China Statistical Yearbook, Shandong Statistical Yearbook, Qingdao Statistical Yearbook and exchange materials of other provinces and cities.

On the mixture of two power function distributions

Epimaco A. Cabanlit, Jr., Mycah Shaene R. Nailon
Mindanao State University, General Santos City, Philippines

Abstract

The mixture of distributions can serve as a model to some realities where the population consists of heterogeneous components. The mixture of two power function distributions provides a mathematical-based approach to the statistical modelling of data. This paper presents some of the important summaries of the mixture of two power function distributions such as the mean, variance, and *rth* moment about the origin.

Keywords

power function distributions; Mixture of distributions; Summaries of Distribution

1. Introduction

A suitable generalized lifetime model is often of interest in the analysis of survival data, as it can provide insight into characteristics of failure times and hazard functions that may not be available with classical models. Four distributions, Exponential, Pareto, Power and Weibull, are of interest and very attractive in lifetime literature due to their simplicity, easiness and flexible features to model various types of data in different fields (Cordeiro, et al., 2012). Exponential distribution is a distribution of the time to an event when the probability of the event occurring in the next small time interval does not vary through time. Pareto distribution is often described as the basis of the 80/20 rule. Weibull distribution can represent decreasing, constant, or increasing failure rates (Forbes, C., et al., 2011). Power distribution can be used to fit the distribution of certain likelihood ratios in statistical tests and it can be used to compare two tests which have the same significance level (Cordeiro, et al., 2012).

Meniconi and Barry (1995) discussed the application of the power function distribution (PFD) along with other lifetime models, and concluded that the PFD is better than the Weibull, log-normal and exponential models to measure the reliability of any electrical component. The use of exponential, Weibull, and log-normal, which are frequently preferred over mathematically more complex distribution, suggests that most engineers favour the application of simpler models to obtain failure rates and reliability figures quickly. It is therefore proposed that the power function distribution be considered as a simple

alternative which exhibit a better fit for failure data and provides more appropriate information about reliability and hazard rates (Meniconi, M. and Parry, D.M. ,1996).

Several authors have reported characterization of the PFD based on order statistics and records. One of these authors was Rider (1964) who first derived the distribution of the product and ratio of the order statistics from a power function distribution(Rider, P.R.,1966). Another, Ahsanullah (1973) defined necessary and sufficient conditions based on PFD order statistics. Also, Kabir and Ahsanullah (1975) discussed the estimation of the location and scale parameters of a power function distribution. And Moothathu (1884) gave characterizations of the PFD through Lorenz curve.

In probability theory and statistics, the power function distribution is a continuous probability distribution. It is a flexible lifetime model which can be obtained from the Pareto model and it is also a special case of the beta distribution (Dallas, A.C.,1978).

The probability density function is defined as

$$f_X(x) = \begin{cases} \frac{cx^{c-1}}{b^c}, & 0 \le x \le b \\ 0, & otherwise \end{cases}$$

with shape parameter c, and scale parameter b > 0 [6].

A mixture distribution, a multivariate distribution, is the probability of a random variable (may be random real numbers or they may be random vectors, each having the same dimension) that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized. Mixture models based on probability density function have been used successfully on a number of applications ranging from speaker recognition to bioinformatics (Dinampo, W.,2016). The formula for the mixture of two power function distributions is defined by

$$f_X(x) = \phi_1 f_1(x) + \phi_2 f_2(x)$$

where $\phi_1 + \phi_2 = 1$, and

$$f_1(x) = \begin{cases} \frac{c_1 x^{c_1-1}}{{b_1}^{c_1}}, & 0 \le x \le b_1 \\ 0, & otherwise \end{cases}$$

and

$$f_2(x) = \begin{cases} \frac{c_2 x^{c_2-1}}{{b_1}^{c_2}}, & 0 \le x \le b_2 \\ 0, & otherwise. \end{cases}$$

2. Methodology

The paper is a pure research. The results are obtained based on well-defined definitions and theorems.

3. Result

a. The Probability Density Function of Two Power Function Distributions

Theorem 1. *If the probability density function fX (x) of the mixture of two power function distributions, with shape parameter ci and scale parameter bi > 0 where i = 1, 2, is*

$$f_X(x) = \phi_1 f_1(x) + \phi_2 f_2(x)$$

where $\phi_1 + \phi_2 = 1$, and

$$f_1(x) = \begin{cases} \dfrac{c_1 x^{c_1-1}}{{b_1}^{c_1}}, & 0 \le x \le b_1 \\ 0, & otherwise \end{cases}$$

and

$$f_2(x) = \begin{cases} \dfrac{c_2 x^{c_2-1}}{{b_2}^{c_2}}, & 0 \le x \le b_2 \\ 0, & otherwise, \end{cases}$$

then

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1.$$

b. The Mean and Variance of the Mixture of Two Power Function Distributions

Theorem 2. *If X is a random variable of the mixture of two power function distributions, with shape parameter ci and scale parameter bi > 0 where i = 1, 2, then the mean of X, denoted by E(x), is*

$$E(x) = \phi_1 \frac{b_1 c_1}{c_1 + 1} + \phi_2 \frac{b_2 c_2}{c_2 + 1}.$$

Theorem 3. *If X is a random variable of the mixture of two power function distributions, with shape parameter ci and scale parameter bi > 0 where i = 1, 2, then*

$$Var(x) = \frac{\phi_1 b_1^2 c_1 c_2 + 2\phi_1 b_1^2 c_1 + \phi_2 b_2^2 c_1 c_2 + 2\phi_2 b_2^2 c_2}{(c_1 + 2)(c_2 + 2)} - \frac{\phi_1^2 b_1^2 c_1^2 (c_2 + 1)^2 + 2\phi_1 \phi_2 b_1 b_2 c_1 c_2 (c_1 + 1)(c_2 + 1) + \phi_2^2 b_2^2 c_2^2 (c_1 + 1)^2}{(c_1 + 1)^2 (c_2 + 1)^2}$$

c. The rth Moment about the Origin of the Mixture of Two Power Function Distributions

Theorem 4. *If X is a random variable of the mixture of two power function distributions, with shape parameter ci and scale parameter bi > 0 where i = 1, 2, then*

$$\mu_r' = \phi_1 \frac{b_1^r c_1}{c_1 + r} + \phi_2 \frac{b_2^r c_2}{c_2 + r}.$$

4. Discussion and Conclusion

The following are the results obtained in this study:

1. The probability density function of the mixture of two power function distributions satisfies the property

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1.$$

2. The mean of the mixture of two power function distributions is

$$E(x) = \phi_1 \frac{b_1 c_1}{c_1 + 1} + \phi_2 \frac{b_2 c_2}{c_2 + 1}.$$

3. The variance of the mixture of two power function distributions is

$$Var(x) = \frac{\phi_1 b_1^2 c_1 c_2 + 2\phi_1 b_1^2 c_1 + \phi_2 b_2^2 c_1 c_2 + 2\phi_2 b_2^2 c_2}{(c_1 + 2)(c_2 + 2)} -$$

$$\frac{\phi_1^2 b_1^2 c_1^2 (c_2 + 1)^2 + 2\phi_1 \phi_2 b_1 b_2 c_1 c_2 (c_1 + 1)(c_2 + 1) + \phi_2^2 b_2^2 c_2^2 (c_1 + 1)^2}{(c_1 + 1)^2 (c_2 + 1)^2}.$$

4. The rth moment about the origin of the mixture of two power function distributions is

$$\mu_r' = \phi_1 \frac{b_1^r c_1}{c_1 + r} + \phi_2 \frac{b_2^r c_2}{c_2 + r}.$$

The following are recommended for further studies:

1. Study of mixture of more than two power function distributions;
2. Study of other mixtures of distribution

References

1. Ahsanulla, M. (1989). "Estimation of the parameters of a power function distribution by Recordvalues", Pakistan Journal of Statistics Vol. 5, pp.189-194.
2. Cordeiro, Gauss, Mansoor, M., Morad, Alizadeh and Zubair, M. (2012) ."The Weibull-power Function distribution with Applications", Brazalian Journal of Probability and Statistics Vol. 26, pp.88112.
3. Dallas, A.C. (1978). "Characterizations of the power function distribution", Annal of Mathematical Statistics Vol. 28, pp.491-497.
4. Dinampo, W. (2016)."On the Mixture of Two Logistic Distributions.", Undergraduate Thesis, Mindanao State University, General Santos City.
5. Dudewicz, E. and Mishra, S. (1988)"Modern Mathematical Statistics", John Wiley Sons, Inc., USA.
6. Forbes, Catherine, Evans, Merran, Hstings, Nicholas and Peacock, Brian (2011). "Statistical Distributions.", 4th edition, A John Wiley Sons, INC. publication, New York.
7. Gordon, I. (2013)."Continuous Probability Distributions - A guide for teachers (Year 11- 12)", Education Services Austrilia.
8. Kabir, A.B.M.G. and Ahsanulla, M. (1974) "Estimation of location and scale parameters of a power function distribution by linear functions of order Statistics", Communication in Statistics Vol. 3, pp.463-467.
9. Meniconi, M. and Parry, D.M. (1996). "The power function distribution: A useful and simple distribution to assess electrical component reliability", Microelectronics Reliability Vol. 36, pp.1207-1212.
10. Moothathu, T.S.K.A. (1986) "A Characterization of power function distribution through a property of the Lorenz curve", Sankhya 1348, pp.262-265.
11. Rider, P.R. (1966) "Distribution of product and quotient of maximum values in samples from a power function population", Journal of the American Statistical association Vol. 59, pp.887-880.

Couple's time allocation to housework and childcare: Moroccan evidence

Asmae Mhmmoudi
The High Commission for Planning, Rabat, Morocco

Abstract

The purpose of this paper is to study the allocation of unpaid work time in Moroccan couples using data from the National Time Use Survey (2012). By including non-participants, censored regression model is used in order to estimate the impact of socio economic background on housework and childcare behavior's. In addition, unpaid work is divided into ordinary housework and childcare in order to investigate whether they behave differently with housework and childcare. In general, the results show a negative effect of urban area and a positive impact of children inversely related to their ages though their impact is important within women than men. Furthermore, the results of the estimation of housework are particularly the same. However, when childcare is estimated separately, the presence of children affects positively men childcare time allocation.

Keywords

housework and childcare; couple's unpaid work; time use

1. Introduction

In terms of paid and unpaid work, Moroccan women and men work approximately the same amount of hours. According to the National Time Use Survey (2012), women and men spend 6 hours on paid and unpaid work. But while men do most paid work, women are generally assigned to the home as their occupational area.

To gain more insight into activities sharing within the household and investigate gender specialization, attention is paid particularly to couples. Specifically, the study focuses on time spent by couples on unpaid work which combines housework and childcare, housework and childcare, and therefore analyzing the factors affecting the three allocations.

Traditional analyses of unpaid work defined as time spent on housework and childcare. Besides examining couples allocations to unpaid work, unpaid work is spilt into housework and childcare in order to investigate whether a structural difference exists between time spent doing housework and time spent caring for children. The rationale behind this split is that the utility associated with childcare is different from the utility generated by ordinary housework.

2. Recording housework time in the National Time Use Survey

The National Time Use Survey carried out in 2012 is the second survey of this type, after a first one in 1997 covering only women. In that second survey, women, men and children are all involved. Data were collected during a whole year and each participant was asked to cite their activities for a day. Activities were subsequently coded according to a list with more than 550 activity codes. The following activities are categorized as housework: Food preparation and serving, cleaning and upkeep of dwelling and surroundings, doing laundry, wood chopping and water fetching, maintenance and small repairs, purchases for the households, gardening activities and household management. Care provided within children aged less than 15 years constitutes, however, a far more central part of the daily activities of household members. Hence, it would be appropriate to include childcare into housework participation's measurement.

The main focus here is to study couple's time allocation to unpaid work. However, the National Time Use Survey (2012) collected diaries of women and men of any marital status. As such, the primary exercise is to construct a data base of couples that contain all the information required to investigate couple's time allocation without any influence of other factors. So that, the analysis is restricted to nuclear households, i.e. couples living with or without children.

3. Estimation method

In the econometric analysis, three allocations of time are estimated: the total time spent on unpaid work, housework and childcare. However, while each individual completed the time use diary for a single day, it is possible that in that particular day the person doesn't accomplish any activity of housework or childcare. This is particularly the case of Moroccan men for whom housework and childcare activities are rarely performed. Consequently, these variables contain many zeros. Hence, distributions are censored and in this case OLS estimation is not relevant. In such case of censored distribution, limited dependant variable models are mostly used. Specifically, in this paper a Tobit model is used for the estimation.

The econometric framework in this study focused on time spent by couples on unpaid work. In the first part of the analysis, the dependent variable is defined as the total time spent on housework and childcare jointly and analyzed according to place of residence for both women and men. In the second part of the analysis, the time spent caring for children and the time spent doing housework are analyzed separately. For all estimations R.3.4.1 is used.

To investigate the factors affecting couple's allocation of time to housework and childcare, two types of variables are included: individual-specific and couple-specific variables:

- Average age of the couple: introduced to capture susceptible changes of social norms. A traditional role sharing is therefore is expected for old couples. (Anxo & Carlin(2004)).
- Age difference: as the difference between husband's age and wife's age. (Beblo (1999)) argues that the larger the age differential, the more unequal is the gender division of work.
  "Wife is older than husband" is the omitted reference category in the estimation.
- Area of residence: Rural area is chosen as the reference
- Age of children: the presence of children and their age affect the time use of their parents. Tree variables corresponding to children's age are included. i) Couple has at least a child less than 2 years old ii) aged 3 to 5 years iii) aged between 6 and 15 years.
- Educational level: educational attainment potentially affects the spouses' allocation of time between market time and home production through two channels. i) Education directly affects earning opportunities. Consequently, this influences bargaining power within couples. ii) Highly educated households would tend to have a more equal distribution of housework time by gender. The educational variable utilized is educational level.
  Four dummies are constructed and those with no education are chosen as the reference
- Number of rooms: treated as a continuous variable and reflected the surface of housing.
- Employment status. This variable is differentiating between three employments status.
  Category with the "employed ones" is chosen as the reference
- Size of the household: number of persons living in the household may also influence couple's allocation to housework: the presence of additional person is expected to increase total time devoted to home production. This variable is treated as a continuous variable.

4. Results – unpaid work

The results discussed in this section are from the estimation of time spent on unpaid work. The next section analyses the results of the estimations of time spent on housework and childcare separately. In order to see how explanatory variables affect couple's behaviors according the area of residence, unpaid work is estimated in this section at national level, urban area and at the rural area.

The national level estimation of unpaid work shows a significant effect of the place of residence on time spent by Moroccan couples on unpaid work. Living in urban area decreases couple's time devoted to unpaid work compared to couple living in rural area. Nevertheless, its effect is more important among women than men. This result is expected if we take into account the greater availability of electricity, substitutable market goods and services in urban areas.

Furthermore, the time devoted to unpaid work vary with the number of children differently according to sex. Mother's time spent on housework increases significantly when children are in the household and inversely related to their age. Regardless of the place of residence, the increase in mother's unpaid work is rather for mothers with children less than 2 years. However, there is no significant effect of the presence of children on fathers' share of unpaid work. A similar result was founded by Deding and Mette Lauste (2006) for Danish couples.

The variables introduced initially to capture susceptible changes of social norms have fewer effects on couple's time devoted to unpaid work. Although, the cohort effect is significant in each area of residence, coefficients are quite small, indicating the persistence of traditional roles sharing in Moroccan society.

Contrary to rural area, educational attainment of husband living in cities matters significantly their time spent on housework and childcare. Indeed, husbands with high education perform a higher share of housework and childcare. This finding lead to the conclusion that a modern vision characterized by a redistribution of roles more equally tends to be developed for higher levels of educational attainment in urban areas However, education of wives has no significant effect on their unpaid work time. Moreover, high education level of wives doesn't matter significantly time spent on unpaid work of their husbands.

Inactivity and unemployment of the wife reduce unpaid work time of her husband. Nevertheless this effect is only significant in urban areas where the daily allocation of time spent in unpaid work decreases by about 11 minutes. However, the performance of unemployed and inactive women living in cities is more important compared to employed women. Moreover, employment status of the husband doesn't matter the unpaid work time of his wife only when he is unemployed. The time spent doing unpaid work by wife fall by 56 minutes if her husband is unemployed. In turn, this reduction is offset by an increase of his participation to unpaid work. Specifically, his time spent on unpaid work increase by 42 minutes compared to employed men. A possible explanation of this result is that the reduced market work of unemployed and inactive individuals is made up by additional housework and childcare. This is in line with Kitterød and Lyngstad (2005) who found that fathers who work

long hours in the labor market spend less time on housework than non-employed.

5. Result: housework and childcare

In the estimation of the four equations of this section, the housing variables are excluded from the childcare equations, while the dummy variables of number of children are excluded from the housework equations.

Overall, the findings of the estimation of men's housework equation are similar to the first section. All the variables that have previously significant effects on men's time spent doing unpaid work remain significant and keep the same signs but the effects are mostly smaller in the case of housework's time estimation. However, the results are slightly different for the estimation of fathers' time spent caring for children. The marginal effects of the variable of the presence of children have become significant. This significant effect suggests that time spent with children is considered as investment for fathers. This is somewhat similar to the result of Mette Deding and Mette Lauste (2006). Furthermore, the reduced effect of explanatory variables on time spent by men in housework could now be explained by more performance in childcare activities. Moreover, fathers living cities devoted more time caring for children compare to those living rural areas.

6. Conclusion

This paper has examined the influence of socio demographic background on time spent by Moroccan couples unpaid work. Beyond analyzing the time devoted to unpaid work, a separate estimation of unpaid work into housework and childcare has done also in order to see the structural differences that could exist between housework time allocation and childcare time allocation.

The results of the first part of estimation show a significant effect of the place of residence on couple's time allocation to unpaid work., in the sense that living in urban area reduce their share of unpaid work. Moreover, husbands living in cities with high education devote more time to unpaid work. A similar effect is also founded when they are inactive or unemployed, which could be explained by the fact that the reduced market work is offset by more activity in the household. However, contrary to mothers, the presence of children is without effect on time devoted to unpaid work of fathers.

One of the main results of the second part of estimations is that the number of children in the household matters significantly the time spent caring of children for both mothers and fathers. Nevertheless, this effect is more important among women compare to men. Furthermore, women living in large housing or with many persons perform a higher share of housework.

References

1. Agnès & Maurisson (2012), L'évolution des rôles masculin et féminin au sein de la famille. Les Cahiers français: documents d'actualité. La Documentation Française, 6 numéro spécial des Cahiers Français: Comment va la famille? (371), pp.22-29.
2. Alain Jacquot, Les modèles économétriques Logit, Probit, Tobit, Dossier d'étude N°6.
3. AniKatchova (2013), Limited Dependent Variable Models, Econometric Academy.
4. Beblo, M. (1999), Intrafamily time allocation: A panel econometric analysis, in: Merz, J. and M. Ehling (eds.), Time Use - Research, Data and Policy: Contributions from the International Conference on Time Use (ICTU), FFB-Schriftenreihe Band 10, Baden-Baden, Nomos, 473-489.
5. Christophe Hurlin, Modèles à Variable Dépendante LimitéeModèles Tobit Simples et Tobit Généralisés,Polycopié de Cours.
6. Clara Champagne, Ariane Pailhé et Anne Solaz (2015), Le temps domestique et parental des hommes et des femmes: quels facteurs d'évolutions en 25 ans? ÉCONOMIE ET STATISTIQUE N° 478-479-480, 2015.
7. Dominique Anxo and Paul Carlin (2004), Intra-family time allocation to housework -French evidence, Vol. 1, No 1, 14-36.
8. Haut commissariat au Plan (2014), Présentation des premiers résultats de l'Enquête Nationale sur l'Emploi du Temps.
9. Hugo Harari-Kermadec (2009), Econométrie, chapitre 4 le modèle Tobit.
10. Jens Bonke and Bent Jensen (2012), Paid and unpaid work in Denmark – Towards gender equity- in Electronic International Journal of Time Use Research, Vol. 9, No. 1, 108-119.
11. Kunzler, Walter, Reichart and Pfister (2001), Gender division of labour in unified Germany. European Network on Policies and the division of unpaid and paid work, WORC Report 00.00.000/0.
12. Kitterød&Lyngstad (2005), Diary versus questionnaire information on time spent on housework – The case of Norway, in Electronic International Journal of Time Use Research, Vol. 02, 13-32.
13. Lauk, Martina; Meyer, Susanne (2005), Women, men and housework time Allocation: theory and empirical results, Darmstadt Discussion Papers in Economics, No. 143.
14. MetteDeding and MetteLauste (2006), Choosing between his time and her time? Paid and unpaid work of Danish couples, in Electronic International Journal of Time Use Research, Vol. 3, No. 1, 28-48.
15. Régnier-Loilier&Hiron (2010), Évolution de la répartition des tâches domestiques après l'arrivée d'un enfant, Politique Sociales et Familiales Famille-Travail, No.99

16. Sourabh(2007), The culture of women's housework a case study of Bihar, India, academic dissertation, Helsinki University Printing House.
17. Vandeschrick&Sanderson (2013), Partage du travail domestique : des évolutions, peut-être; des résistances, sûrement, GGP Belgium Policy Breif 6.
18. Xavier d'Haultfoeuille, Censure et selection, Cours d´´économétrie 2 Première partie : variables d´dépendantes limitées.

A mixed models approach to extrapolation of clinical data

Daniel Bonzo, Evelyn Wang, Jillian Prescod
Global Biometry, LFB, 175 Crossing Blvd, Framingham, MA 01702, USA

Abstract

Extrapolation of clinical trials data is being accepted increasingly by regulatory agencies as a means of generating data in diverse situations during drug development process. We consider this problem of extrapolation using the concept of estimand [Akacha, M., et. al. (2017)] under a mixed models setting. The concept of estimand captures population, endpoint, and a measure of effect – in general, one can think about extrapolation of historical data from one estimand to another closely related estimand. A likelihood procedure is presented for estimating the parameters of interest under a generalized linear models setting. Allowing the possibility of censored/grouped data transforms the likelihood expression into a likelihood involving counts of interval data by utilizing the latent variable concept. A relatively simple estimation and testing construction is obtained when one assumes that the underlying distribution comes from the family of exponential distribution. Using large sample approximation, we show an approach for goodness-of fit testing and estimation of parameters of interest. Finally, we demonstrate the utility of this construction in a setting were we evaluate efficacy in a subgroup of a clinical trial population using a marker of efficacy. In conclusion, the concept of estimand allows an extrapolation approach that can cover a broad array of applications and settings, including the case when censoring is allowed. Useful expressions of estimators and tests are given for application purposes, though they require sufficiently large sample to be efficient. These expressions have an intrinsic weighting mechanism for the different sources of data.

Keywords

Estimand; latent variable; censored; likelihood; goodness-of fit

1. Introduction

In this paper we present a procedure for extrapolation that can be applied in a general setting where the underlying distribution comes from the family of exponential distributions. This procedure can be used to treat a variety of problems in drug development that call for extrapolation of results, e.g., from adult to pediatric population, from one or several geographic regions to another such as in bridging studies from one indication to a related indication,

and from one biologic product to a biosimilar product (or one formulation to a bioequivalent formulation).

The regulatory guideline for extrapolation calls for the extension of information and conclusions available from studies in one or more subgroups of the patient population (source population(s)), or in related conditions or with related medicinal products, to make inferences for another subgroup of the population (target population), or condition or product. This definition was proposed in a European Medicines Agency (EMA) concept paper on extrapolation of efficacy and safety in medicine development. This then reduces the need to generate additional information (types of studies, design modifications, number of patients required) to reach conclusions for the target population, or condition or medicinal product.

The procedure for extrapolation is facilitated by the notion of estimand as introduced in the revision of International Conference on Harmonization (ICH) E9(R1). The proposed framework states that an estimand reflects what is to be estimated to address the scientific question of interest posed by a clinical trial. The choice involves the population of interest, endpoint of interest, and measure of intervention effect. Ultimately one can think about the problem of extrapolation as weighting of available data based on defined estimands [Akacha, M., et. al. (2017)] for the scientific question of interest.

We assume that observations of interest arise from a parametric distribution function whose parameters can be estimated based on an estimating equation such as the derivative of the log-likelihood function. Allowing the possibility of censored/grouped data transforms the likelihood expression into a likelihood involving counts of interval data by utilizing the latent variable concept [Lazarsfeld, P. F. & Henry, N. W. (1968)]. We also assume that the observable outcomes are intervals of the form (a, b], a < b, where b is the maximum measurement level at which the subject experiences a defined response, e.g. treatment success, and a is the lower bound for such a response. In this construction the measurement level defining the response is latent.

The paper is organized as follows. Section 2 presents the maximum likelihood procedure for estimating distribution parameters under a generalized linear models setting. Section 3 provides the extrapolation construction for a normal latent measurement variable and a homogeneous Poisson latent count of events. Then using large sample approximation, we show an approach for goodness-of fit testing that can be extended to accommodate an extrapolation setting. Utility of the estimation and goodness-of-fit testing approach is shown in a setting where we evaluate efficacy in a subgroup of a clinical trial population using a marker of efficacy. Finally, Section 4 gives some concluding remarks.

2. Methodology

Assume that a random sample of size n from a fixed distribution F is observed, where the observations are confined to intervals given by pre-defined fixed points within the support of F. In this section we consider estimation of parameters using interval information.

Formally, let $(Y_1, \dots, Y_n)$ be independently and identically distributed (IID) latent random variables with distribution F and consider fixed points $a_1, \dots, a_k$ in the support of F such that $a_{i-1} < a_i, i = 1, \dots, k$, where $a_0 = -\infty$ and $a_k = \infty$. Let X_{0i} represent the count or frequency of the (unobserved) $Y_i's$ falling in the ith interval $(a_{i-1}, a_i]$, i = 1, . . . , k. In order to account for information from subjects who prematurely discontinue from the trial, we also consider right-censored observations, i.e., intervals of the form (a, ∞) where a is in the support of F. Such intervals will overlap the fixed intervals $(a_{i-1}, a_i]$, i = 1, . .k. Denote the count for the interval (a_s, ∞) by m_s, s = 1, . . . , k-1 and the total count for the non-overlapping intervals by n_0. Thus, the data consists of two sets of frequencies, (X_{01}, … , $X_{0k)}$ for the non-overlapping intervals and (m_1, … , m_{k-1}) for the overlapping intervals. From these let $n_0 = \sum_{i=1}^{k-1} X_{0i}$ and $n_1 = \sum_{s=1}^{k-1} m_s$ so that the total number of observations is $n = n_0 + n_1$.

In the presence of right-censored observations, the log-likelihood is given by $l(\theta) = \sum_{i=1}^{k} X_{0i} \log[F(a_i; \theta) - F(a_{i-1}; \theta)] + \sum_{i=1}^{k-1} m_s \log[1 - F(a_s; \theta)]$.
Note that $l(\theta) = l_0(\theta) + l_i(\theta)$ is just the respective sum of the log-likelihood for uncensored intervals and for right-censored intervals. Similarly, the score function $S(\theta)$ Hessian matrix $H(\theta)$ and information matrix $I(\theta)$ admit the same decomposition. Thus, $S(\theta) = S_0(\theta) + S_1(\theta), H(\theta) = H_0(\theta) + H_1(\theta)$ and $I(\theta) = I_0(\theta) + I_1(\theta)$, where S_0, H_0 and I_0 are the quantities corresponding to the non-censored intervals and S_1, H_1 and I_1 are the quantities corresponding to the right-censored intervals.

Estimation of the unknown parameter θ can be done by maximizing the log-likelihood $l(\theta)$. In general, maximization of the log-likelihood will not admit a closed solution for the estimating equation $\frac{\partial}{\partial \theta} l(\theta) = 0$. Hence, the maximum likelihood estimator (MLE) $\hat{\theta}$, must be obtained by using an iterative approach such as a Newton-type method, a modified Fisher scoring algorithm, or a modified EM-type algorithm. Using the Newton-Raphson method, the estimate of θ at the $(m+1)^{th}$ iteration is given by

$$\theta^{(m+1)} = \theta^m - H^{-1}(\theta) S(\theta)|_{\theta=\theta^m} \quad (1)$$

Where $S(\theta) \frac{\partial}{\partial \theta} l(\theta)$ is the Fisher score function and $H(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} l(\theta)$ is the Hessian matrix. Iteration is terminated when $\left|\theta^{(m+1)} - \theta^m\right| \leq \varepsilon$, for some pre-specified amount ε. Under reasonable assumptions on $F(y; \theta)$ and sufficiently accurate initial value, the sequence of estimates $\theta^{(m)}$ enjoys local quadratic convergence to a solution $\hat{\theta}$. This solution is in fact the MLE of θ if $l(\theta)$ is concave and unimodal.

If $\hat{\theta}$ is the MLE, its limiting distribution can be used to draw inferences on $\theta_{i'}, i' = 1, \dots, d$. Using the fact that $\hat{\theta} \rightarrow_L N\left(\theta, I^{-1}(\theta)\right)$, where $I(\theta) = -E\left(H(\theta)\right)$ The standard error of the estimator $\widehat{\theta}_{i'}$ of $\theta_{i'}$ is given by $I_{i'}^{-1/2}(\theta)$ where $I_{i'}(\theta)$ is the i'th diagonal element of $I(\theta)$ [see Serfling, R.J. (1980)]. Thus, an approximate 95% confidence interval for $\theta_{i'}$ is given by $\widehat{\theta}_{i'} \pm 1.96 I_{i'}^{-\frac{1}{2}}\left(\widehat{\theta}_{i'}\right)$. A test for the general hypothesis $H: C\theta = c_0$ can be conducted based on a Wald-type statistic $\chi_W^2 = \left(C\hat{\theta} - c_0\right)'(CI^{-1}(\theta)C')'\left(C\hat{\theta} - c_0\right)$. Under H, $\chi_W^2 \rightarrow_L \chi_r^2$, where $r = rank(C)$.

To simplify model construction, we assume that F belongs to the family of exponential distributions, i.e., the probability density (mass) function is given by $f(y;\theta) = h(y)exp\{\sum_{i'=1}^{r} \eta_{i'}(\theta)\, T_{i'}(y) - A(\theta)\}$. The exponential family is said to be in canonical form if $\eta_{i'}(\theta) = \theta_{i'}, i' = 1, \dots, d$ and it said to be curved if $\dim(\Theta) < \dim\left(\eta(\Theta)\right)$. In the simplest case, i.e., canonical representation with r = 1, we assume that $\theta(\mu) = z'\beta$, where $\mu = E(Y|z,\beta)$, z is the vector of covariates (fixed and random), β is the vector of structural parameters and $\theta(.)$ is the link function. This defines a generalized linear model representation for Y.

3. Result

We now provide some constructions for extrapolation under a generalized linear mixed model setting using interval information. These constructions can certainly be extended to allow for more involved extrapolation cases.

a. Extrapolation with Normal Latent Variables

We consider the simple case where $\theta\left(\mu_{ij}\right) = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2j} + \beta_3 z_{3ij}$. In this representation, z_{1ij}, for all $j = 1, \dots, p$, is the fixed subject covariate where $i = 1, \dots, n_j$ represents the subjects associated with p different estimands. Furthermore, z_{2j} is the fixed estimand covariate where $j = 1, \dots, p$. Lastly, z_{3ij} represents the treatment covariate. Also, we assume that the latent variable Y is normally distributed and observed in terms of the ordinal variable $\tilde{Y}$ through the intervals $\left(a_{i-1,j}, a_{i,j}\right], i = 1, \dots, k$ as defined in Section 2. Note that in applications, the fixed covariates for subject and estimand are measured through proxy variables to facilitate the estimation of the fixed subject effect β_1 and fixed estimand effect β_2.

Assuming that $\theta' = \left(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2\right)$ as the parameter to be estimated with $\Theta = \{(\beta', \sigma^2): \beta_i \in \mathbb{R}, i = 0, \dots, 3; \sigma^2 > 0\}$ as the parameter space then (1) is implemented using the log-likelihood

$$\begin{aligned} l(\theta) &= \sum_{j=1}^{p} \sum_{i=1}^{k} X_{0ij}\, log\left[\Phi\left(\frac{a_{ij} - z_{ij}'\beta}{\sigma}\right) - \Phi\left(\frac{a_{i-1j} - z_{i-1j}'\beta}{\sigma}\right)\right] + \\ &\quad \sum_{j=1}^{p} \sum_{s=1}^{k-1} m_{sj}\, log\left[1 - \Phi\left(\frac{a_{sj} - z_{sj}'\beta}{\sigma}\right)\right] \\ &= l_0(\theta) + l_1(\theta) \end{aligned}$$

Where $(m_{1j}, \dots, m_{k-1j})$ represent the right-censored counts and $z'_{ij} = (1, z_{1ij}, z_{2j}, z_{3ij})$ are the covariates. Estimation of θ and testing of $H{:}C\theta = c_0$ will proceed as described in Section 2.

b. Extrapolation for Poisson Counts

Here we introduce a construction for modelling of event counts. We assume that the underlying count process is a homogeneous Poisson process with the conditional mean given by $\theta(\mu_{ij}) = \log(\mu_{ij}) = \beta_0 + \log z_{1ij} + \log z_{2j} + \beta_1 z_{3ij}$. In this representation, Z_1 is the random subject effect, whereas Z_2 is the random estimand effect. In a variety of applications, it is reasonable to assume that Z_1=$Z_{11}Z_{12}$ where Z_{11} represents the subject's time at risk and Z_{12} denotes frailty. For simplicity, one can assume that Z_{11},Z_{12} and Z_2 are independent and as in the previous construction we let z_{3ij} denote the treatment covariate. Also, we assume that the latent homogeneous count process Y is Poisson distributed and observed in terms of the ordinal variable $\tilde{Y}$ through the integer intervals $(a_{i-1,j}, a_{i,j}]$, i = 1, … , k as defined in Section 2.

Conditional on $Z_1 = z_{1ij}$ and $Z_2 = z_{2j}$ and assuming that $\theta' = (\beta_0, \beta_1, v')$ as the parameter to be estimated with $\Theta = \{(\beta', v'): \beta_i \in \mathbb{R}, i = 0,1; v \in \mathbb{R}^q\}$ as the parameter space and v denoting the vector of nuisance parameters associated with time at risk, frailty, and estimand effect, then (1) can be implemented using the conditional log-likelihood

$$l(\theta) = \sum_{j=1}^{p} \sum_{i=1}^{k} X_{0ij}\, log \left[\sum_{r=a_{i-1j}+1}^{a_{ij}} \frac{(\beta_0 + log z_{1ij} + log z_{2j} + \beta_1 z_{3ij})^r}{r!\, z_{1ij} z_{2j} e^{\beta_0 + \beta_1 z_{3ij}}} \right] + \sum_{j=1}^{p} \sum_{s=1}^{k-1} m_{sj}\, log \left[1 - \sum_{r=0}^{a_{sj}} \frac{(\beta_0 + log z_{1sj} + log z_{2j} + \beta_1 z_{3sj})^r}{r!\, z_{1sj} z_{2j} e^{\beta_0 + \beta_1 z_{3sj}}} \right] = l_0(\theta) + l_1(\theta)$$

where $(m_{1j}, \dots, m_{k-1j})$ represent the right-censored counts and $z'_{ij} = (1, z_{1ij}, z_{2j}, z_{3ij})$ are the covariates. The random covariates for subject and estimand are measured through proxy variables for estimation purposes. Estimation of θ and testing of $H{:}\, C\theta = c_0$ will proceed as described in Section 2.

c. Goodness-of-Fit Testing

In the following section we demonstrate how interval information can be used to conduct test of hypotheses on latent variable distribution assumptions. The method here can be extended to construct a goodness-of-fit test under an extrapolation setting, i.e., in the presence of subject and estimand effects.

We construct the test procedure based on the classical distribution fitting problem. Here we make use of count sets $(X_{s1}, \dots, X_{sk}), s = 0,1, \dots, k-1$, corresponding to the random sample $(Y_1, \dots, Y_n)$ from some distribution F as discussed in Section 2. Utilizing the notations defined in Section 2 we have at s=0 the counts $(X_{01}, \dots X_{sk})$ corresponding to the non-overlapping intervals. For $s > 0$ define $X_{sj} = m_s I\{j = s\} + X_{0j} I\{j > s\}$ as where $(m_1, \dots, m_{k-1})$

represent the counts for the right-censored intervals. Thus, the total count of the observations falling in a tail interval (a_s, ∞) is given by

$$M_s = \sum_{j=s+1}^{k}(X_{sj} + m_{j-1}) = \sum_{j=s+1}^{k}(X_{0j} + m_{j-1}).$$

For the problem of testing that the underlying distribution is equal to F_0, we consider alternative distributions F which are asymptotically close. For these class of distributions, an asymptotically uniformly most powerful (UMP) invariant test for H: $F=F_0$ vs. K: $F\neq F_0$ is given by the test: Reject H if

$$T_n = n_0 \sum_{i=1}^{k} \frac{1}{\pi_{0i}}\left(\frac{X_{0i}}{n_0} - \pi_{0i}\right)^2 + \sum_{j=1}^{k-1} \frac{1}{p_{0j}}\left(W_j - p_{0j}\right)^2 I\{m_j > 0\} \qquad (2)$$

is sufficiently large, where X_{0i} is the number of observations in $(a_{i-1}, a_i]$ and $\pi_{0i} = F_0(a_i) - F_0(a_{i-1}), i = 1, \ldots, k;\ p_{0j} = \sum_{s=j+1}^{k} \pi_{0s}$ represents the cumulative proportion of observations in (a_j, ∞) and $W_j = \frac{1}{n_0 + \sum_{s=j+1}^{k} m_{s-1}} \sum_{s=j+1}^{k} (X_{0s} + m_{s-1})$ the corresponding estimator, $j = 1, \ldots, k-1$; and $I\{m_j > 0\}$ is an indicator function signifying that the corresponding term will be included only if the tail interval count is non-zero. The critical value C is determined by $\int_C^{\infty} \chi^2(z)\, dz$ where $\chi^2(.)$ is the chi-square probability density function with k-1 degrees of freedom. The asymptotic power of this test is given by $\beta = \int_C^{\infty} \chi^2_{k-1,v}(z) dz$, where the non-centrality parameter v is equal to $n_0 \sum_{i=1}^{k} \frac{\left(F(a_i) - F(a_{i-1})\right)^2}{F_0(a_i) - F_0(a_{i-1})}$.

The asymptotic distribution of T_n under H follows from the fact that $n_0 \sum_{i=1}^{k} \frac{1}{\pi_{0i}}\left(\frac{X_{0i}}{n_0} - \pi_{0i}\right)^2 \rightarrow_L \chi^2_{k-1}$. Also, the second term in (2) can be shown to satisfy $\sum_{j=1}^{k-1} \frac{1}{p_{0j}}\left(W_j - p_{0j}\right)^2 I\{m_j > 0\} \rightarrow_P 0$. Thus the distributional result is obtained via the direct application of Slutsky's theorem.

The asymptotic optimality of the test follows directly from the property of rejection regions constructed on the basis of quadratic forms with underlying multivariate normal distribution. That is, if $(X_1, \ldots, X_k)' \sim \text{MVN}(\eta, \sum)$ then there exists a uniformly most powerful invariant (under a suitable group G of linear transformations) with rejection region of the form $\sum_{i,j} \sigma_{ij}^{-1}\left(\hat{\eta}_i - \hat{\hat{\eta}}_i\right)\left(\hat{\eta}_j - \hat{\hat{\eta}}_j\right) > C$ where $\hat{\eta}_i$ minimizes the quadratic form under a linear space K, $\hat{\hat{\eta}}_i$ minimizes the quadratic form for H $\subset$ K and $\Sigma^{-1} = \left(\sigma_{ij}^{-1}\right)$[see Lehmann, E.L. (1986)].

Implementation of the test for parametric families $F(y;\theta)$ proceeds as follows. If $\hat{\theta}$ is the MLE obtained by maximizing an appropriate log-likelihood as given in Section 2, then substituting $\pi_{0i}^{(n)} = F_0\left(a_i; \hat{\theta}\right) - F_0\left(a_{i-1}; \hat{\theta}\right)$ in place of π_{0i} and $p_{0j}^{(n)}| = \sum_{s=j+1}^{k} \pi_{0s}^{(n)}$ in place of p_{0j} in equation (2) yields an asymptotically UMP invariant test for H: $F = F_0$ vs. K: $F \neq F_0$. Asymptotic invariance of the test follows directly from the $\sqrt{n}$ −consistency of MLE's.

d. An Application

Table 1 shows the interval counts for both the study drug and active comparator where the intervals represent regions of efficacious response based on some biomarker in a subgroup of a clinical trial population considered as having a severe disease status.

Using (2) the goodness-of-fit test results for normal distribution are given in Table 2. For the active comparator (standard drug) the approximate chi-square statistic had a p-value equal to 0.9544. This means that the observed proportions do not significantly differ from the null proportions of a normal distribution. In the case of the study drug (test drug) the approximate chi-square statistic yielded a p-value equal to 0.4256. This also shows that the observed proportions do not significantly differ from the null proportions of a normal distribution. Thus, the efficacy response for the study and active drugs can be assumed to have come from the normal distribution.

4. Discussion and Conclusion

The preceding sections provide a generalized linear mixed model method for estimation and testing of parameters in an extrapolation setting when observed information comes in the form of interval data. A likelihood procedure is obtained by assuming that the underlying distribution comes from the family of exponential distributions. Application of the approach was shown in an extrapolation construction for a normal latent measurement variable and a homogeneous Poisson latent count of events. Then using large sample approximation, an approach for goodness-of fit testing that can be extended to accommodate an extrapolation setting was shown. Utility of the construction was then shown in a setting where efficacy is evaluated in a subgroup of a clinical trial population using a marker of efficacy.

In conclusion, the concept of estimand allows an extrapolation approach that can cover a broad array of applications and settings, including the case when censoring is allowed. Useful expressions of estimators and tests are given for application purposes, though they require sufficiently large sample to be efficient. These expressions have an intrinsic weighting mechanism for the different sources of data. The approach presented can be extended to allow utilization of prior information expressed in terms of a power or commensurate power model [Gamalo-Siebers et. al., (2017)] under a hierarchical Bayesian model setting. Irrespective of the approach taken, these models can be useful tools for extrapolation allowing one to model the uncertainty as between-estimand variance, evaluate different scenarios through simulation and calculate sample sizes.

Table 1. Interval Information for Efficacious Response by Treatment Group (Intent-to-Treat Population)

Type of Subject	Totals	Active Comparator		Study Drug	
		Interval	Count	Interval	Count
Completer		(-∞, 20]	12	(-∞, 12.5]	11
		(20, 40]	5	(12.5, 25]	7
		(40, 60]	5	(25, 50]	5
		(60, ∞)	7	(50, ∞)	6
	n_0		29		29
Non-Completer	m_1 (M_1)	(20, ∞)	0 (17)	(12.5, ∞)	2 (20)
(Right-Censored)	m_2 (M_2)	(40, ∞)	3 (15)	(25, ∞)	2 (13)
	m_3 (M_3)	(60, ∞)	0 (7)	(50, ∞)	0 (6)

Table 2. Asymptotic Goodness-of-Fit Test and Estimation Results for Normal Distribution

Active Comparator			Study Drug		
Interval	Observed Proportions	Null Proportions	Interval	Observed Proportions	Null Proportions
(-∞, 20]	0.4138	0.3707	(-∞, 12.5]	0.3793	0.3505
(20, 40]	0.1724	0.1783	(12.5, 25]	0.2414	0.1544
(40, 60]	0.1724	0.1687	(25, 50]	0.1724	0.2846
(60, ∞)	0.2413	0.2822	(50, ∞)	0.2069	0.2105
(20, ∞)	-----	-----	(12.5, ∞)	0.7878	0.6495
(40, ∞)	0.4687	0.4509	(25, ∞)	0.4193	0.4951
(60, ∞)	-----	-----	(50, ∞)	-----	-----
χ^2 Value (df)	0.3244 (3)		χ^2 Value	2.7871 (3)	
p-Value	0.9554		p-Value	0.4256	
Mean	34.5569		Mean	24.6160	
Variance	1948.8692		Variance	995.2267	
SE(Mean)	8.7568		SE(Mean)	6.2186	
Log-Likelihood	-40.6699		Log-Likelihood	-42.5122	

References

1. Akacha, M., Bretz F., & Ruberg, S. Estimands in clinical trials – broadening the perspective. Statistics in Medicine, 36(1):5–19, 2017. ISSN 1097-0258.
2. EMA. Reflection paper on extrapolation of efficacy and safety in paediatric medicine development.
3. Lazarsfeld, P. F. & Henry, N. W. (1968). Latent structure analysis. Houghton Mill, Boston, MA.
4. Serfling, R. J. (1980). Approximation theorems of mathematical statistics. John Wiley & Sons, New York.
5. Lehmann, E.L. (1986). Testing statistical hypotheses. John Wiley & Sons, New York.
6. Gamalo-Siebers, M., Savic, J., Basu, C., Zhao, X., Gopalakrishnan, M., Gao, A., Song, G., Baygani, S., Thompson, L., Xia, H. A., Price, K., Tiwari, R. and Carlin, B. P. Statistical modelling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. Pharmaceutical Statistics, 16(4):232–249, 2017.

Statistics training in a developing country: prospects and challenges experienced in the last two decades at the School of Statistics and Planning, Makerere University, Uganda

Agnes M.N. Ssekiboobo
School of Statistics and Planning, Makerere University, Kampala, Uganda

Abstract

Some of the experiences and challenges being met today in statistics training at the School of Statistics and Planning (SSP) are not new but a confirmation of what has remained relevant while some strategies that have been put in place to improve statistics training need to be refocused or new ones developed. It should however be noted that there are new or emerging issues like climate change, management of natural disasters, rural statistics and others. Capacity for collection and handling of statistics in a number of developing countries in terms of institutional arrangement, trained personnel and methodological development is rather limited. As a result the national statistics system is also limited in providing statistical products and services to meet user data needs. The gap between data requirements and data availability and quality is widening in many developing countries and thus the need to invest in statistics training and data management so as to bridge this gap and allow cross-cutting analysis. There is also need in Uganda to carry out a training need assessment to find out the basic skills required. Given the unprecedented challenges as well as opportunities for statistical development in Africa which stem from the new focus on managing results and evaluations of a number of continental action plans, and concerns raised thereof, SSP had to reposition itself to play a greater role in current statistical reforms and development in Africa by realigning its training programmes appropriately. The paper will therefore review the journey that SSP has embarked on in the last two decades, the available training opportunities, the challenges that have been encountered and prospects of dealing with them. The main objective of all this will be to see how SSP can contribute better to the efforts of developing countries becoming self-reliant in the provision of trained statistical personnel at different levels so as to build a robust, self-sufficient and sustainable national capacity for statistics production.

Keywords

review, development, programmes, research, opportunities.

1. Introduction

The School of Statistics and Planning (SSP), formerly known as the Institute of Statistics and Applied Economics (ISAE) was established as an autonomous

body within the legal framework of Makerere University in July 1969. Under the United Nations Statistical Training Programme for Africa (STPA) 1978-93 and the plan of the Coordinating Committee for Statistical Development (CASD), ISAE then was designated as one of the regional statistical training centres in the African region and was supported to spearhead high level professional training of personnel in statistics and applied economics and population science urgently needed for social and economic planning and development to meet the human resource needs of Uganda and other English speaking African countries. SSP is therefore both a regional statistical training centre and a mainstream Makerere University unit (ISAE, 2007). Since 1969, SSP has trained over 10,000 statisticians, planners, applied economists, actuarial scientists and demographers from 22 African countries. Most ISAE/SSP alumni hold key professionally posts in National Statistical Services, regional and international organizations. Makerere University has since 2011 turned collegiate and ISAE became SSP and together with the School of Business and School of Economics formed the College of Business and Management Sciences (CoBAMS).

2. Methodology

The review involved the analysis of:

i. Annual reports of Makerere University and the then ISAE and now the SSP;
ii. The strategic plans for ISAE for 2000/1 – 2004/5 and 2006/7 – 2010/11;
iii. Literature review of research activities and empirical studies; and
iv. National and international development frameworks.

3. Results

The SSP has grown both in the scope and depth of its programmes and services since it was founded five decades ago. In order to consolidate its past successes in training statisticians at professional level and to adequately develop the post-graduate and specialized programmes, the University Council approved three departments for the SSP, namely:

- Department of Statistical Methods and Actuarial Sciences
- Department of Planning and Applied Statistics; and the
- Department of Population Studies

SSP currently with a population of about 1850 students up from 60 in 1969, runs five undergraduate, ten post-graduate training programmes and short courses/workshops. All undergraduate programmes have duration of three years. These include: Bachelor of Statistics, Bachelor of Science in Actuarial Science, Bachelor of Science in Quantitative Economics, Bachelor of Science in Population Studies and Bachelor of Science in Business Statistics. The post-

graduate diploma programmes which are run for one year include postgraduate Diploma in Demography and postgraduate Diploma in Statistics; the postgraduate Diploma in Population and Reproductive Health is due to start next academic year.

The Master's degree programmes are run for two years and include Master of Statistics, Master of Science in Quantitative Economics, Master of Science in Population and Reproductive Health, Master of Arts in Demography, Master of Arts in Population and Development, and Master of Science in Population Studies. The last three Masters programmes are going to be phased out and replaced with the Master of Demography and Population Studies. All Ph.D. programmes take at least three and half years and a maximum of five years. They include Ph.D. in Population Studies and Ph.D. in Statistics.

Short courses and workshops are conducted by SSP from time to time to meet specific skills needs of user countries. In addition to these programmes, SSP services statistics courses in different schools and departments. In turn, SSP is serviced by the Department of Mathematics and the School of Economics.

3.1. The changing environment

The last two decades have been a time of unprecedented challenges as well as opportunities for statistical development in Africa both of which stem from the new focus on managing for results. It has also led to unprecedented increase in demand for data (both a challenge and an opportunity) and exposed the weaknesses of National Statistical Systems in most African countries to meet this demand. In order to mitigate or eliminate weaknesses in their statistical systems, countries have undertaken statistical reforms, reengineered the statistical systems and designed National Strategies for the Development of Statistics (NSDSs). All this has posed a new challenge for the SSP and other statistical training centres to produce more relevant professionals and to reposition themselves to play a greater role in shaping the development of statistics in Africa (ISAE, 2007).

The evaluation of the implementation of the Addis Ababa Plan of Action which was undertaken in 2000 also raised a number of concerns about statistical capacity building in Africa. These included the fact that: training especially at Universities tends to be theoretical; training is done from the "supply side" and generally crowds out "demand issues"; and there was insufficient collaboration between training centres and National Statistics Offices (NSOs) (United Nations Economic Commission for Africa (UNECA), 2000).

3.2. The journey

SSP brought on board these concerns and other issues raised and repositioned itself to play a greater role in current statistical reforms and development in Africa by undertaking the following initiatives:

a. Realigning its training programmes, research agenda and consultancy work with the new national and international focus on managing for results where demand is for development statistics and not traditional statistics. As an organization, SSP has carried out a number of consultancies and research activities independently or in collaboration with other organizations. The research undertaken is in line with the University-wide research agenda which is linked to the national development framework, principally the National Development Plan (NDP) and international development agenda; formerly the Millennium Development Goals (MDGs) and now the Sustainable Development Goals (SDGs). A number of members of staff also do consultancies for governments, national or international organizations, as individuals or in collaboration with other schools or consultants. This has helped deepen the professional credentials of SSP and its staff. In line with the University-wide research agenda, SSP staff have been encouraged to spend more time doing research and publishing; sourcing funds for research and, in general, promoting research work and building partnerships for research between SSP and collaborators; closing research capacity gaps by expanding disciplines on which research is done; and SSP putting the Centre for Population and Applied Statistics (CPAS) in place fully fledged to undertake research.
b. Periodically reviewing and updating the curriculum to make the programmes offered at SSP more relevant to the information needs under different arrangements, decentralization being one of them. This enhances the relevance of the curriculum in a dynamic policy and development environment. It also helps the School to keep pace with changing user demand for data. More emphasis was also put on integrative and consistency frameworks or data quality frameworks.
c. Teaching statistics from the demand side rather than the supply side - realigning training programmes to meeting the needs of data users who are becoming more diversified and more critical and addressing emerging needs.
d. Improving training in statistical analysis, reporting, presentation and communication.
e. Conducting short training courses on policy and decision-making using statistics.
f. Harnessing IT to improve statistical training and to transform statistical operations.

g. Using new technologies (GIS, GPS, etc.), software and developing and teaching new/appropriate methodologies and systems for data collection and management.
h. Advocating for more intensive use of websites that provide training materials like for Food and Agriculture Organisation of the United Nations (FAO), UNECA and the World Bank. This helps SSP to have a point of reference especially in as far implementing recommendations that may be relevant to them are concerned. New methodologies that have been developed as well as useful information for improving teaching materials are available on these websites. Efforts are also being made to have connectivity with the library resource centre of the Uganda Bureau of Statistics.
i. Introducing internship/field attachment for all the students in addition to the research project that all students have to undertake as a partial requirement for the award of their degrees.
j. Improving the quality of teaching, learning and research environment (in terms of physical infrastructure, staffing, information technology, etc.).
k. Using of holistic approaches (statistical consistency and integration frameworks as teaching frameworks - national accounting, satellite accounts, integrated information systems, etc.).
l. Improving the quality of research and community services including generating greater awareness about role of statistics to society.
m. Developing/strengthening partnerships for statistical training and research. It is well recognized that most institutions and systems in Africa need to be nurtured and further strengthened. One way of doing this is to build partnerships for statistical development including twinning arrangements. Partnerships have been enhanced between the SSP and other units within the University, between the SSP and user countries, between academic statisticians at universities and training centres on the one hand and official statisticians working in NSOs on the other, and between training centres [Eastern Africa Statistical Training Centre (EASTC), Dar-es-Salaam, Tanzania; Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSEA), Abidjan and of recent Institut Sous-regional de Statistique et d'Economie Appliquée ISSEA-Yaoundé and in the near future, Institut National de Statistique et d'Economie Appliquée (INSEA) –Rabat]. The twinning arrangements between the SSP and other STCs have been limited but of recent have been enhanced by the UNECA/African Development Bank funded project on Improving Statistics for Food Security, Sustainable Agriculture and Rural Development in Africa: Support to the Global Strategy that the SSP has benefited from. Experiences obtained from other training institutions by SSP in developing statistical capacity including best practices and

knowledge of what has and has not worked in similar circumstances have been studied. Partnership between academic statisticians at universities and training centres on the one hand and official statisticians working at NSOs is very important and has been very useful to SSP as it augurs well for the teaching of official statistics. Scaling up partnerships and interactions between academic staff at SSP and official statisticians at the Uganda Bureau of Statistics (UBOS) has enhanced the relevance of statistical training at SSP. Some of the practical and applied courses are taught by senior and experienced staff from UBOS and other data producing agencies on a part-time basis so that the students are exposed to practical and applied aspects of these subjects. The Plans for National Statistical Development (PNSDs) encourage this relationship. The African Group on Statistical Training and Human Resources (AGROST) has indeed played its role in fulfilling its mandate by coordinating ongoing and future statistical training activities in Africa. Together with UNECA and other UN agencies, they will propel the continent towards achieving its targets in the area of human resource development and also provide the required advisory services.

n. Investing in staff development to improve performance of SSP cannot be over-emphasized. Efforts on staff development are being implemented through training, mentoring and motivation. To improve the quality of service to students, the University has embarked on rigorous training of staff in the last decade. A number of SSP staff have registered for further studies at Makerere University or outside Uganda over the years and there has been a steady increase in the number of PhD holders. However, because of the policy of holding a PhD to be appointed a Lecturer, the staff development funds at Makerere University have been overstretched, though development partners like the Swedish International Development Cooperation Agency (SIDA) have come in to support this initiative.

3.3 Main Challenges at SSP and in general of Improving Statistics Training:

These include:

a. Loss of monopoly in providing training in statistics and related subjects in the country and in the African region has been a challenge. New Universities in Uganda have started statistical training programmes and also a number of old Universities in the African region are now teaching statistics. This has created a challenge for the SSP to respond to this loss of monopoly.

b. SSP needs a home much as the university in general has adequate central teaching facilities. The number of students continues to grow and

therefore this limits the effectiveness of SSP in their teaching and research programmes.

c. The training of middle level staff has remained a challenge. Like professional level training, the course content should be periodically reviewed and changes made to reflect current priorities and latest data collection and processing methods. The question is whether for further training of professional staff, priority should be given to practical courses of short duration (not more than 9 months) or training up to Masters level. Does the work of NSOs or the National Statistical Systems (NSSs) in general deem training beyond Masters degree necessary or desirable? Are the NSS or the NSOs sufficiently developed to undertake in-depth substantive or methodological research which may require post-masters training? It should however be noted that these offices should undertake some analytical work and countries have to assess the need for post-masters training from time to time. The question would then be whether the training in analytical techniques should be one that leads to the award of a degree.
d. Limited funding from government.

4. Discussion, Conclusions and Recommendations

The last two decades have been a period of promise and SSP has done well in addressing the main challenges namely the dynamic user needs in African countries and loss of monopoly. The School is now an active centre for both basic and applied research and through research has made significant partnerships and contributions to different areas of national development and also to the international arena. It is important to note though that the expansion of study programmes and student numbers has constrained existing SSP and other university infrastructure. For SSP to continue functioning effectively, the following aspects must be looked into:

a) The staff of the SSP should be encouraged to spend their Sabbatical Leave at various NSOs in the region to gain practical experience in the subjects they teach and also to contribute directly to the improvement of statistical work at the NSOs.
b) Modularize Courses. Modularization of teaching programmes should also being considered to enable short-term competency-based training of varying durations providing students with on the job and off the job training.
c) Introduce e-learning. There is need for SSP to introduce e-learning as a means to vary and strengthen teaching programmes. E-learning will not only make it possible for the SSP to provide training for staff that cannot be away for a long period of time but also more people can be trained using this training method.

d) Training in management. Managers of the SSP need to use modern management principles and practices in their work especially in planning, work programming and budgeting, proposals and report writing. All staff with management responsibilities will be given training in these principles.
e) Rigorous marketing of SSP. The Dean of the School should be seen attending key sub-regional, regional and international meetings, symposia and conferences as a way of marketing the SSP and foster networking arrangements.
f) It is necessary that scholarships are made available at postgraduate level in order to build capacity in the different areas of specialization. This is because there are no government scholarships for graduate students.
g) A foundational course on official statistics should be put in place for all students both undergraduate and postgraduate to cover national and international development frameworks, integration frameworks, data sources and data quality frameworks, sources of data for indicators and their development, statistical ethics and legislation. In addition other areas to put emphasis on include qualitative assessments, gender responsive statistics, measuring progress, etc.
h) The need to develop statistical capacity at different levels of local government must be taken up by SSP so as to include in the school's training agenda a sub-national orientation that caters for the training demands of statistical personnel in local government.
i) The trainers especially for short courses should recognize the experience of the trainees and build on it so that they connect prior experience to learning and build on it by relating it to current job responsibilities. This will ensure that the staff of statistical agencies and units have the knowledge, skills and technical competencies they need and that these competencies are kept up to date.
j) A Needs Assessment should be undertaken among users of statistics as well as employers of the products of STCs and universities, to ensure that what is demanded is actually going to be used. Demand and supply information is important to training providers and statistical agencies. The training plan so developed would then translate the requirements of the Statistical Systems into specific qualification profiles of staff or other people for whom training is then called for. This should then avoid problems such as:
    i. Having the right training for the wrong people.
    ii. Having the right training for the right people only to go back and they cannot use their new skills, and
    iii. People wanting different training from what they need.
k) It is important to also strengthen the field attachment programme for the

students which had worked very well when for example the SSP was supported by UBOS to place the students in the districts and sub-counties. This has proved very beneficial to the students who get to know how local governments at different levels operate as far as data collection, entry, processing and dissemination are concerned. It has also helped the SSP to appreciate the demands of local governments and other stakeholders and the quality of graduate they want in order to meet these demands.

l) Finally, more financial support is needed to enable the above and in addition for:
   i. Fellowships for both trainers and trainees.
   ii. Research and development of appropriate methodologies.
   iii. Development of guide syllabuses and production of relevant teaching materials as well as periodic review of syllabuses.
   iv. Support for the holding of seminars, workshops and short courses in priority areas of applied statistics.
   v. Equipment and accessories.
   vi. Harmonising courses standards. This may require comparative analysis of the syllabuses of the STCs in terms of courses and hours covered per course, number of applied courses, the reading lists, etc. This also enables students to move from one training centre to another without any problem.

References

1. African Development Bank, PARIS21, UN Economic Commission for Africa (2005). Communique: National Strategy for the Development of Statistics. Briefing Session, Addis Ababa, Ethiopia, 8-11 August 2005.
2. Institute of Statistics and Applied Economics (2007). Strategic Plan 2006/7-2010/11.
3. Lufumpa Charles and Michel Mouyelo-Katoula (2005). Strengthening Statistical Capacity in African Countries under the Framework of the International Comparison Program for Africa (ICP-Africa). The African Statistical Journal, Vol. 1.
4. Makerere University, Planning and Development Department (2000 – 2017). Annual Reports. Kampala, Uganda
5. Organisation for Economic Co-operation and Development et al (2004). Action Plan on Managing for Development Results. Second International Roundtable, Marrakech, Morocco.
6. United Nations Economic Commission for Africa (1990). Addis Ababa Plan of Action for Statistical Development in Africa in the 1990s. Addis Ababa, Ethiopia.
7. United Nations Economic Commission for Africa (1990). Creating and Strengthening a Statistics Teaching Group. Addis Ababa, Ethiopia.
8. United Nations Economic Commission for Africa (1990). Guide Syllabus for Middle-Level Personnel in Statistics (English-speaking Countries. Addis Ababa, Ethiopia.
9. United Nations Economic Commission for Arica (1993). Statistical Newsletter, No. 81.
10. United Nations Economic Commission for Africa (2000). Overall Report of the Evaluation of the Addis Ababa Plan of Action for Statistical Development in Africa in 1990s. Addis Ababa, Ethiopia.
11. United Nations Economic Commission for Africa et al (2006). Reference Regional Strategic Framework for Statistical Capacity Building in Africa: Better Statistics for improved Development Outcomes. Addis Ababa, Ethiopia.

Richness estimation with species identity error

Jai-Hua Yen, Chun-Huo Chiu
Division of Biometry, Department of Agronomy, National Taiwan University, Taiwan

Abstract

Richness estimation of an interesting area is always a challenge statistical work due to small sample size or species identity error. In the literatures, most richness estimators were only proposed to tackle the underestimation of the size-limited sample. However, species identity error almost occurs in each species survey and seriously reduces the accuracy of observed, singleton, and doubleton richness in turns to influence the behavior of richness estimator. Therefore, to estimate the true richness, the biased collected data due to species identity error should be modified before processing the richness estimation work. In the manuscript, we propose a new approach to correct the bias of richness estimation due to species identity error. First, a species list inventory from a subplot obtained by the investigator was used to estimate the species identity error rate. Then, we can correct the biased observed, singleton, and doubleton richness of the raw sampling data from the interesting area. Finally, the rich-ness estimators proposed in the literatures could be supplied to get the more correct estimates based on adjusted observed data. To investigate the behavior of the proposed method, we performed simulations by generating data sets from various species models with different species identity error rates. For the purpose of illustration, the real data was supplied to demonstrate our proposed approach. A presence/absence weeds species was surveyed in the organic farmland located at Soft Bridge County in the North of Taiwan.

Keywords

Biodiversity; Singleton; Doubleton; Sampling error

1. Introduction

Long-term biodiversity monitoring is the basis for ecological research and promotion of organic agriculture. In recent years, more and more non-professional citizen scientists have participated in the projects of monitoring diversity, so the possibility of species identity errors may increase dramatically in the collected data. Therefore, correcting the impact of species identification errors becomes an important statistical issue.

Species richness is the most intuitive and widely used as biodiversity index due to its ecological intuitive concept and simplest form. However, due to the

sampling limitation of time or other re-sources, completely species inventories in the wild field are almost unattainable goals. Therefore, the observed richness in the sample always underestimates the true species richness in the assemblage. In the literatures, among the discussed estimation approaches of species richness, the nonparametric methods are widely used in practical application, which include first order Jackknife approach, second order Jackknife approach by Burnham and Overton (1978) and Chao1 (or Chao2) lower bound estimator by Chao (1984). They all use the observed rare species in the sample (i.e. singletons and dou-bletons) to estimate the unseen richness in the sample. However, species identity error almost occurred in each survey especially in vegetation sampling was ignored before and recently discussed in the literatures by Vittoz and Guisan (2007), Burg et al. (2015), and Morrison (2015). This identity error may seriously make observed richness biased and in turn the estimation of true richness will be seriously biased. Therefore, without error adjustment, the species richness estimation will be inaccurate based on original sampling data. In this manuscript, we have proposed a modify approach to revise the biased sampling data caused by species identity error. From the results of simulation study in secession 3 show that our adjusting approach can be nearly unbiased to revise the biased observed richness, singleton and doubleton richness. Also, the richness estimators based on the revised data effectively correct the bias caused by the species identity error.

2. Methodology

In this article, we choose Chao2 lower bound estimator for incidence data as our species richness estimator. Since we assume that species identity error exists in the process of sampling, adjustment of richness estimator should be considered.

First, we need to estimate the mean species identity error rate of observer or investigator. Plant inventories from subplot of the area which the survey is conducted. We assume that the number of species (S_{sub}) and the categories of species in the subplot are known by the experiment designer but unknown by the observer who goes conducting inventories. After conducting inventories, we have the information that the number of observed species belongs to the subplot $(S_{sub,e})$ and the number of observed species does not exist in the subplot $(f_{sub,0})$. X_i represents the record status of the survey of species i. When $X_i = 1$, species i has been recorded. When $X_i = 0$, species i has not been recorded. We assume the species identity error (e) is a random variable follows the distribution of $F(e)$ with mean $\bar{e}$. r denotes the mean probability that a species is misidentified into another species which belongs to the sampling plot. $f_{sub,0}$ equals to the number of species which is misidentified and recorded as species do not exist in the subplot. Also, if plant inventories of the subplot are correct, then $S_{sub,e}$ should be equal to S_{sub}

species. However, when species identity error occurs, $S_{sub,e}$ may not be equal to S_{sub} species. When the i-th species is misidentified and other species are not misidentified to the i- th species, i-th species is not recorded. After that, we have the equations:

$$E(f_{sub,0}) = \int S_{sub} \times e \times (1-r) dF(e) \approx S_{sub} \times \bar{e} \times (1-r), \quad (1)$$

and

$$E(S_{sub,e}) = S_{sub} - \sum_{i=1}^{S_{sub}} E[I(X_{i=}0)] \approx S_{sub} - S_{sub} \int e \times \left(1 - \frac{e}{\frac{S_{sub}}{r} - 1}\right)^{S_{sub}-1} dF(e)$$

$$\approx S_{sub} - S_{sub} \times \bar{e} \times \left(1 - \frac{\bar{e} \times r}{S_{sub} - r}\right)^{S_{sub}-1}. \quad (2)$$

By solving those two equations, we have the estimate of $\bar{e}$ and r which are denoted by $\hat{\bar{e}}$ and $\hat{r}$.

Second, the sampled observed, singleton, and doubleton richness should be adjusted after sampling in the plot. The true observed, singleton, and doubleton richness are denoted by S_{obs}, Q_1, and Q_2, respectively. The sampled observed, singleton, and doubleton richness without adjustment are denoted by $S_{obs,e}$, Q_{1e} , and Q_{2e} , respectively. When species identity error occurs, the sampled observed richness is formed by the observed species which do not misidentified and observed species which misidentified as species do not exist in the plot. Thus, we have the expected sampled observed richness:

$$E(S_{obs,e}) \approx E\{S_{obs}[(1-e) + e \times (1-r)]\}.$$

Next, we have the expected observed richness adjustment:

$$S_{obs,a} = \frac{S_{obs,e}}{1 - \hat{\bar{e}} \times \hat{r}} \quad (3)$$

When species identity error occurs, the possibilities of sampled singleton species are as follows: (1) singleton species which do not misidentified, and other species would not be misidentified as the singleton species at the same time, and (2) singleton species which misidentified as species do not exist in the plot, and other species would not be misidentified as the singleton species at the same time. Thus, we have the expected sampled singleton richness:

$$E(Q_{1e}) \approx E\left\{Q_1[(1-e) + e \times (1-r)] \times \left(1 - \frac{e}{\frac{S_{obs}}{r} - 1}\right)^{S_{obs}-1}\right\}$$

$$\approx E\{Q_1[(1-e) + e \times (1-r)] \times exp(-e \times r)\}.$$

Similarly, when species identity error occurs, the possibilities of sampled doubleton species are as follows: (1) doubleton species which do not misidentified, and other species would not be misidentified as the singleton species at the same time, (2) doubleton species which misidentified as species do not exist in the plot, and other species would not be misidentified as the singleton species at the same time, and (3) when a singleton species misidentified to a singleton species, the doubleton richness increases by one unit, and other species would not be misidentified as the doubleton species

which is formed by singleton species at the same time. Accordingly, we have the expected sampled doubleton richness:

$$E(Q_{2e}) \approx E\{Q_2[(1-e)+e\times(1-r)]\times exp(-e\times r)\} + E\left\{Q_1\times e\times r\times\left(1-\frac{1}{T}\right)\times\frac{Q_1}{S_{obs,a}}\times exp(-e\times r)\right\}.$$

where T denotes the number of sampling unit. By solving the two equations above, we have the singleton and doubleton richness adjustment:

$$Q_{1a} = \frac{Q_{1e}}{(1-\hat{e}\times\hat{r})exp(-\hat{e}\times\hat{r})}, \quad (4)$$

and

$$Q_{2a} = \frac{Q_{2e}-Q_{1a}\times\hat{e}\times\hat{r}\times\left(1-\frac{1}{T}\right)\times\frac{Q_{1a}}{S_{obs,a}}\times exp(-\hat{e}\times\hat{r})}{(1-\hat{e}\times\hat{r})\times exp(-\hat{e}\times\hat{r})}. \quad (5)$$

However, the estimation of traditional Chao2 estimator will be inaccurate even though Q_{1a} and Q_{2a} are asymptoticly unbiased. It causes the value of $\frac{Q_{1a}^2}{2Q_{2a}}$ overestimated. Hence, we choose first-order Jackknife and Chao2 richness estimator as the theoretical foundation of deriving the adjusted richness estimator. We propose an adjusted richness estimator by Taylor series expansion of $E\left(\frac{Q_1^2}{2Q_2}\right)$ by the mean Q_1 and Q_2. Then we get the difference between $\frac{[E(Q_2)]^2}{E(2Q_2)}$ and $E\left(\frac{Q_1^2}{2Q_2}\right)$ to have the adjust term:

$$E\left(\frac{Q_1^2}{2Q_2}\right) \approx \frac{[E(Q_1)]^2}{E(2Q_2)} + \frac{V\hat{a}r(Q_1)}{2E(Q_2)} - \frac{E(Q_1)C\hat{o}v(Q_1,Q_2)}{[E(Q_2)]^2} + \frac{[E(Q_1)]^2V\hat{a}r(Q_2)}{2[E(Q_2)]^3},$$

where $C\hat{o}v(Q_1,Q_2) = -\frac{Q_1Q_2}{\hat{S}}$, $V\hat{a}r(Q_i) = Q_i\left(1-\frac{Q_i}{\hat{S}}\right)$. Therefore, we have the adjusted richness estimator:

$$\hat{S}_{adj} = S_{obs,a} + \frac{T-1}{T}max\left\{\left(\frac{Q_{1a}^2}{2Q_{2a}} - \frac{Q_{1a}}{2Q_{2a}} - \frac{Q_{1a}^2}{2Q_{2a}^2}\right), 0\right\}. \quad (6)$$

When $0 \le Q_{2a} \le 1$, by simulation studies, the adjusted richness estimator will be:

$$\hat{S}_{adj} = S_{obs,a} + \frac{T-1}{T}Q_{1a}. \quad (7)$$

3. Result

a. Simulation Results

To test the performance of the adjusted richness estimator, we presented the simulation results by several species detection models and different settings of number of sampling units. We fixed $S_{sub} = 40$ and $S = 100$. 500 simulation data sets were generated and 200 bootstrapping trials were conducted by each simulation data. The bootstrapping method is regenerating $S_{obs,a}, Q_{1a}$, and Q_{2a} by binomial distribution independently in order to increase the estimated standard error while the traditional bootstrapping method usually underestimates the standard error in this case. In true method, the estimation of species richness used the traditional Chao2

estimator by the data without species identity error. In observed method, the estimation of species richness used the traditional Chao2 estimator by the data with species identity error. In adjusted method, the estimation of species richness used the adjusted richness estimator by the data with species identity error.

When species identity error occurs, the estimate of species richness by observed method will be underestimated, which causes larger bias. The large bias still exists even though the increase of the number of sampling units. Since adjusted method slightly overestimated species richness when the species identity error rate is large, it reduces a great quantity of bias. The variation of observed method is lower, and it remains the same by different species identity error rate. The adjusted method has a higher variation. When species identity error rate is larger, the variation of adjusted method is larger. By evaluating both bias and variation, the observed method has a larger RMSE (Root Mean Square Error) due to its larger bias. The adjusted method has about half RMSE of the observed method when the number of sampling unit is large.

b. Extrapolation for Poisson Counts

The data set was collected of weed species from organic farmland located at Soft Bridge county in the North of Taiwan. There are 12 transect lines with length 20m each were conducted. Only the incidence (detection or non-detection) of species in each transect line was recorded. Before richness estimation, a subplot occupied by 40 known weed species was treated as the testing of the degree of investigator's skill. Compare these 40 weed species list with the inventories of the investigator, we have $S_{sub} = 40$, $S_{sub,e} = 35$, and $f_{sub,0} = 1$. Therefore, we have the estimate of $\hat{e}$= 0.14 and $\hat{r}$= 0.82 based on equations (1) and (2). Many of the misidentified species were misidentified as species which did not exist in the plot. The summary of the frequency counts of weed species is in Table 5. The result using our adjusted estimator is in Table 6. By simulation studies, the error rate is high in this case. Hence, the estimate of species richness using row data directly underestimates and the adjusted estimator should be applied to get the accurate estimate of species richness.

Table 1.

Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under random uniform (0, 1) model, with $\bar{p}$= 0.51, CV = 0.53, S = 100, S_{sub} = 40, T = 5, and r = 0.91.

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\widehat{S}$	Bias	Sample s.e	Estimated s.e	Sample RMSE
0	0	True	85.2	15.3	17.3	91.37	-8.63	4.82	4.19	9.89
0.053	0.058	Observed	81.5	13.9	15.8	87.22	-12.78	5.46	4.06	13.9
		Adjusted	86.3	15.6	17.5	92.05	-7.95*	7.17	8.33	10.71†
0.097	0.098	Observed	78.3	13.2	14.8	83.72	-16.28	5.29	3.95	17.12

		Adjusted	86.3	15.9	17.5	92.2	-7.8*	7.92	9.4	11.12†
0.15	0.157	Observed	74	11.7	13.4	78.86	-21.14	5.24	3.75	21.78
		Adjusted	86.8	16	17.6	92.89	-7.11*	10.33	10.2	12.54†
0.199	0.209	Observed	70.7	10.3	12.7	74.71	-25.29	5.01	3.34	25.78
		Adjusted	88.3	15.8	18.5	94.34	-5.66*	14.05	11.12	15.15†

*Denotes the smaller bias. †Denotes the smaller RMSE.

Table 2.

Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under random uniform (0, 1) model, with $\bar{p}$ = 0.51, CV = 0.53, S = 100, S_{sub} = 40, T = 20, and r = 0.91.

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\widehat{S}$	Bias	Sample s.e	Estimated s.e	Sample RMSE
0	0	True	95.3	4.1	3.9	98.8	-1.2	4.9	4.25	5.06
0.053	0.055	Observed	91.2	3.9	3.6	94.8	-5.2	5.46	4.45	7.53
		Adjusted	96.1	4.3	4	97.85	-2.15*	5.26	5.39	5.68†
0.097	0.095	Observed	87.3	3.3	3.5	90.1	-9.9	5.15	3.76	11.15
		Adjusted	95.8	4	4.1	97.1	-2.9*	6.52	5.72	7.14†
0.15	0.151	Observed	82.9	3.1	2.9	85.61	-14.39	5.21	3.79	15.31
		Adjusted	96.7	4.1	3.9	97.94	-2.06*	8.94	6.23	9.17†
0.199	0.21	Observed	79.2	2.9	2.7	81.79	-18.21	5.25	3.66	18.95
		Adjusted	98.8	4.4	4	100.5	0.46*	11.52	7.04	11.53†

*Denotes the smaller bias. †Denotes the smaller RMSE

Table 3.

Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under a mixture model (0.8 × $Uniform$(0.1,0.3) + 0.2 × $Uniform$(0.4,1)), with $\bar{p}$ = 0.29, CV = 0.7, S = 100, S_{sub} = 40, T = 5, and r = 0.67

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\widehat{S}$	Bias	Sample s.e	Estimated s.e	Sample RMSE
0	0	True	72	32.4	19.8	94.98	-5.02	11.38	10.76	12.44
0.053	0.056	Observed	69.7	30.4	19	90.91	-9.09	11.13	10.22	14.37
		Adjusted	72.5	32.9	19.9	94.7	-5.3*	12.56	12.98	13.63†
0.097	0.1	Observed	67.3	28.8	18.3	87.32	-12.68	11.12	9.91	16.87†
		Adjusted	72.3	33.1	19.8	95.78	-4.22*	21.14	15.12	21.56
0.15	0.155	Observed	64.7	26.4	17.7	82.27	-17.73	11.77	9.06	21.28†
		Adjusted	72.7	33.1	20.1	96.26	-3.74*	21.81	17.28	22.13
0.199	0.203	Observed	63.1	24.9	17.2	78.81	-21.19	9.08	8.36	23.06
		Adjusted	73.9	33.9	20.3	98.02	-1.98*	22.62	19.58	22.71†

*Denotes the smaller bias. †Denotes the smaller RMSE.

Table 4.

Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under a mixture model (0.8 × $Uniform(0.1,0.3)$ + 0.2 × $Uniform(0.4,1)$), with $\bar{p}$ = 0.29, CV = 0.7, S = 100, S_{sub} = 40, T = 20, and r = 0.67.

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\widehat{S}$	Bias	Sample s.e	Estimated s.e	Sample RMSE
0	0	True	97.8	7	11.9	100.25	0.25	2.56	2.43	2.57
0.053	0.056	Observed	94.7	6.6	11.1	97.08	-2.92	2.98	2.37	4.17†
		Adjusted	98.5	7.1	12	100.62	0.62*	4.55	5.8	4.59
0.097	0.102	Observed	91.5	6.2	10.4	93.78	-6.22	3.72	2.34	7.25
		Adjusted	98.6	7.2	12	100.76	0.76*	6.24	6.97	6.29†
0.15	0.151	Observed	88.2	5.8	9.8	90.42	-9.58	3.69	2.31	10.27
		Adjusted	98.5	7.2	12.1	100.62	0.62*	7.5	7.5	7.53†
0.199	0.204	Observed	85.4	5.4	9.1	87.45	-12.55	4.2	2.3	13.24
		Adjusted	99.9	7.3	12.2	102.08	2.08*	9.64	7.98	9.86†

*Denotes the smaller bias. †Denotes the smaller RMSE.

Table 5.

Summary of the data set of weed species frequency counts at Soft Bridge county in the North of Taiwan, with T = 12.

Frequency	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}
Counts	18	9	12	8	6	4	1	4	3	3	2	3

Table 6.

Species richness adjustment for data set of weed species from Soft Bridge county in the North of Taiwan in farmland, with T = 12, $\hat{r}$ = 0.82, and $\hat{e}$ = 0.14.

Method	Sobs	Q_1	Q_2	$\widehat{S}$	Estimated s.e.
Observed	74.0	19.0	9.0	92.4	11.27
Adjusted	83.6	24.1	10.6	105.4	18.68

4. Discussion and Conclusion

Species richness is the simplest and most popular measure of biodiversity. The approach of estimating species richness is widely discussed due to its application in many ecological or agricultural issues mentioned by Carvalheiro *et al.* (2011) and Garibaldi *et al.* (2013). In the manuscript, we demonstrated the effect of species identity error while sampling in estimating species richness. When the mean probability that a species is misidentified into another species which belongs to the sampling plot is high, the observed richness and singleton richness will be seriously negative biased which implying most richness estimators' serious underestimation even though increasing sampling units. Our simulations show that the adjusted richness estimator removes a large proportion of the negative bias under different settings of sampling units, species identity error, and species detection model. We suggest that the adjusted richness estimator for incidence data should be

applied to estimate species richness of the target region since species identity error occurs almost in every investigation of species.

Acknowledgements
The research was supported by the Taiwan National Science Council under Project 107-2118-M-002001-MY2 and Council of Agriculture under Project 107AS-1.2.7-ST-a6.

References

1. Burg, S., Rixen, C., Stöckli, V. & Wipf, S. (2015). Observation bias and its causes in botanical surveys on high-alpine summits. *Journal of Vegetation Science* **26**, 191–200.
2. Burnham, K. P., & Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65(3)**, 625-633.
3. Carvalheiro, L. G., Veldtman, R., Shenkute, A. G., Tesfay, G. B., Pirk, C. W. W., Donaldson, J. S., & Nicolson, S. W. (2011). Natural and within-farmland biodiversity enhances crop productivity. *Ecology letters* **14(3)**, 251-259.
4. Chao, A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 265-270.
5. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783-791.
6. Garibaldi, L. A., Steffan-Dewenter, I., Winfree, R., Aizen, M. A., Bommarco, R., Cunningham, S. A., ... & Bartomeus, I. (2013). Wild pollinators enhance fruit set of crops regardless of honey bee abundance. *Science* **339(6127)**, 1608-1611.
7. Morrison, LW. (2015). Observer error in vegetation surveys: a review. *Journal of Plant Ecology* **9**, 367–379.
8. Vittoz, P. & Guisan, A. (2007). How reliable is the monitoring of permanent vegetation plots? A test with multiple observers. *Journal of Vegetation Science* **18**, 413–422

Fitting statistical models to daily rainfall data using gamma and weibull distributions

Syafrina Abdul Halim, Aqilah Halit

Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor

Abstract

Rainfall modelling is valuable to predict the rainfall pattern and assist to prepare for awareness of the extreme weather. In this study, two probability distributions namely Gamma and Weibull distributions were fitted to daily rainfall data for a period of 27 years (1990-2017) at Penang International Airport. The parameters for each distribution were estimated by using Maximum Likelihood Estimation (MLE) method. The model that best described the data was chosen by using Goodness of Fit (GOF) test, known as Cramer-Von Mises (CvM) and Kolmogorov-Smirnov (K-S). Based on the result of the GOF test, Weibull distribution was the best model.

Keywords

Cramer-Von Mises; Gamma; Kolmogorov-Smirnov; Maximum Likelihood Estimation; Weibull

1. Introduction

Malaysia is one of the Southeast Asia's country that has tropical climate seasons. Malaysia is located at north of the equator which has two non-contiguous regions namely West Malaysia (Peninsular Malaysia) and East Malaysia. By lying on the equator gives the climate being hot and humid throughout the year with the average temperatures between 22°C and 32°C. The humidity commonly affected by four seasonal monsoons, specifically named as Southwest monsoon (May to August), Northeast monsoon (November to February) and two inter-monsoon seasons (March to April and September to October) (Syafrina et al., 2015). Minimal rain was received at northern region in Peninsular Malaysia such as Perlis, Kedah, Penang and Perak since Titiwangsa Range blocks the region from getting rain by the north easterly winds (Syafrina et al., 2015). However, in many decades ago, Penang recorded a history due to the widespread flooding on 24 November 1932 when 15 inches (about 380mm) of rainfall was recorded over a 14-hour spell (Netto, 2017). As a result, there are various natural disasters such as flash floods, fallen trees, landslides and home roof damage. According to Suhaila and Abdul (2007), mathematical modelling could be used to get better information on rainfall pattern and its characteristics. In Southwestern region of Saudi Arabia, Abdullah and Al-Mazroui (1998) analysed annual rainfall and

found out that the main season for rainfall was spring, followed by summer. Gamma distribution was found to be the best model to describe the rainfall analysis in Saudi Arabia. In Ibadan metropolis, Oseni and Ayoola (2012) compared the statistical distributions for rainfall data by using two types of test from goodness of fit model, which were Chi Square (CS) and Kolmogorov-Smirnov (K-S) tests. Based from the study, Exponential distribution was the best fitting distribution followed by Poisson, Normal and Gamma distributions. In Malaysia, Suhaila and Abdul (2007) observed that Mixed Exponential was the best model for estimating the daily rainfall amount in Peninsular Malaysia with four different types of wet day. The selection was performed by testing the median of absolute difference (MAD) between the empirical and hypothesized distributions, the traditional Empirical Distribution Function (EDF) Statistics which were K-S statistic, Anderson Darling (AD) statistic and Cramer-von Mises (CvM) statistics.

In 2016, Poisson-Gamma distribution was found to be the most appropriate model for the two components which are the amount (rainfall totals on wet days) and occurrence (dry/wet days) of rainfall simultaneously (Rossita et al., 2016). The study was based on the four selected stations in Peninsular Malaysia namely Bayan Lepas, Jelebu, Kuala Terengganu and Mersing. The most common parametric distribution models that usually been used by the researchers were Gamma, Weibull, Exponential, Gumbel and Kappa. Based from the previous studies, Gamma distribution is widely used to evaluate the rainfall data. One of the reasons is Gamma distribution is well known and easy to understand with the parameters are used to analyse the characteristics of the rainfall regimes (Husak et al., 2006). Gamma distribution with two parameters gave considerable weight for the wet-day amounts (Thom, 1951). For another distribution, Weibull family of distributions has various sensible characteristics, which are invertibility, integrability and closed form expressions for all moments (Duan et al., 1998). Having characteristic of heavy tail distribution, Weibull was widely used in analysing the rainfall data (Syafrina et al., 2017).

2. Methodology

The daily rainfall data at Penang International Airport (Station ID: 486010) from the year 1990 to 2017 was provided by the Global Summary of the Day (GSOD) and National Atmospheric and Oceanic Administration (NOAA). The latitude and longitude of this study are 5.2961°N and 100.2752 °E respectively. Penang International Airport was hit by flood caused by the rainstorm on July 2016 and about 14 flights have been put on hold due to heavy rain and strong wind (Predeep, 2016). All these problems had affected to the economy, people and showed bad image to Penang since it should be a city of international standard. Therefore, this study will be performed to handle these problems.

Statistical Models

Two models for daily rainfall amount are tested with their probability density functions (pdf) as follows. Daily rainfall amount is represented by x.i) Gamma distribution with two parameters α and β represent the shape and scale parameters.

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}(x)^{(\alpha-1)}e^{-\frac{x}{\beta}} \; ; \quad x, \alpha, \beta > 0 \tag{2.1}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx \tag{2.2}$$

Weibull distribution with two parameters α and β represent the shape and scale parameters.

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \; \alpha > 0, \; \beta > 0, \; x > 0 \tag{2.3}$$

Maximum Likelihood Estimation (MLE)

Estimating the parameters is useful in order to fit the distributions to rainfall data. The shape and scale parameters will be treated as unknown values and it is assumed to be independent and identically distributed (i.i.d) for the joint density for all observations of the dataset (Nielson, 2011). MLE starts with the mathematical expressions which is called likelihood function of the sample data. These values of the parameter that maximize the sample likelihood are known as the maximum likelihood estimates. It provides a consistent approach to parameter estimation problems. Suppose that random variables $X_1, \dots, X_n$ have a joint density or frequency function $f(x_1, x_2, \dots, x_n|\theta)$. Given observed values $X_i = x_i$, where $i = 1, \dots, n$, the likelihood of θ as a function of $x_1, x_2, \dots, x_n$ is defined as

$$lik(\theta) = f(x_1, x_2, \dots, x_n|\theta) \tag{2.4}$$

The joint density can be considered as a function of θ rather than as a function of the x_i. The MLE of θ is the value of θ that maximizes the likelihood makes the observed data most probable. If the X_i are assumed to be i.i.d., their joint density is the product of the marginal densities, and the likelihood is

$$lik(\theta) = \prod_{i=1}^{n} f(X_i|\theta) \tag{2.5}$$

it is usually easier to maximize its natural logarithm (which is equivalent since the logarithm is a monotonic function). For an i.i.d. sample, the log likelihood is

$$l(\theta) = \sum_{i=1}^{n} \log[f(X_i|\theta)] \tag{2.6}$$

Goodness of Fit (GOF) tests

The null and alternative hypotheses of the goodness of fit tests are the data follow a specified distribution and the data do not follow the specified distribution, respectively.

i. Cramer-von Mises (CvM) test

This test is an alternative to K-S test for testing the hypothesis that a set of data comes from a specified continuous distribution. This test was introduced by Cramér in 1928 and von Mises in 1931. The CvM statistic is used in testing the goodness of fit of a probability distribution compared to a given empirical distribution function. This test statistic is based on the squared summation of the difference between the Empirical Distribution Function being tested. According to Deidda and Puliga (2007), the CvM test statistic can be computed as:

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_i) - \frac{2i-1}{2n}\right]^2 \tag{2.7}$$

where F is the cumulative distribution function of the specified or hypothesized distribution while x_i is the ordered data in ascending order and n is the number of sample size.

ii. Kolmogorov-Smirnov (K-S) test

Kolmogorov-Smirnov is one of the GOF test which is usually used to decide if a sample comes from a population with a specific distribution. This test is based on the empirical distribution function. Given n is the sample size and xx_{ii} is the data points in ascending order, $xx_1, xx_2, \ldots, xx_{nn}$. The K-S statistic is given by,

$$D = \max_{1 \le i \le n}\left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i)\right) \tag{2.8}$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (Sharma & Singh, 2010).

3. Results

A summary of daily rainfall amount at Penang International Airport is provided in Table 3.1. Based on Table 3.1, the results showed that the standard deviation for the station is large due to several large values in the dataset which possibly be affected by the extreme values as shown in Figure 3.1. This explains the shape of the rainfall distribution for the station that is skewed to the right as shown in Figure 3.2. In addition, the mean rainfall amount is near to zero value, which is 0.249. This shows that Penang International Airport received low mean rainfall amount. The irregularity of the daily rainfall data of the station can be represented by the coefficient of variation (CV) which is evident in all cases that the 100% is clearly exceeded with amount of variability of rainfall is 250.20%. As shown in the Figure 3.3, the highest amount of rainfall is recorded during the month of October, followed by September and

November. These results could be explained by the influences of the monsoon. Intermonsoon occurred starting from September to October. The result shows that Penang International Airport experienced maximal rainfall during inter-monsoon.

Table 3.1. Summary of daily rainfall at Penang International Airport from the year 1990 until 2017

Minimum (mm)	Median (mm)	Mean (mm)	Standard Deviation (mm)	Maximum (mm)	Coefficient of Variation (100%)	95% Confidence Interval (mm)
0.000	0.000	0.249	0.623	14.250	250.20	(0.237,0.261)

Gamma distribution is denoted by $Gamma \sim \Gamma(\alpha, \beta)$, while Weibull distribution is denoted by $Weibull \sim Weib(\alpha, \beta)$, which is α and β are for shape and scale respectively. Based on Table 3.1, the value of shape parameter for Gamma distribution is 2.79707801 while the value of scale parameter is 0.48540250. Meanwhile, shape parameter of Weibull distribution gives a value of 1.51149743 while the scale parameter gives a value of 1.52373851. The shape parameter describes the curve of the graph where it is positively skewed. In order to test the performance of the fitted model, the GOF test was performed for daily rainfall data in Penang International Airport. The statistical tests used in this study are CvM (ω^2) and K-S (D) tests. The highest total score of the test identify the best statistical model between Gamma distribution and Weibull distribution.

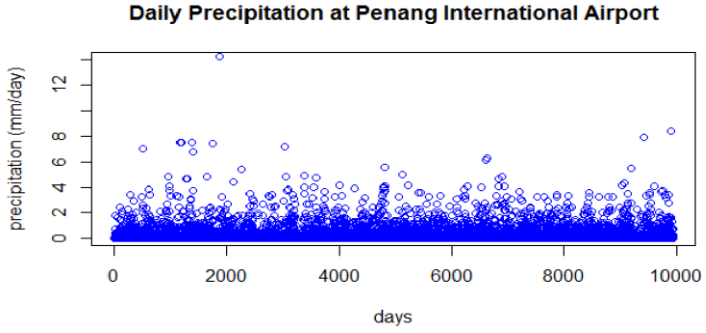


Figure 3.1. Scatter plot of daily rainfall data

Next, each test will be ranked from zero to two (0-2) as there are only two distributions that used in this research. Generally, the statistical test will be ranked from the least to the large value of statistical tests in descending order based on the value of the test statistic. The highest value of total rank score will be identified as the best fit model for daily rainfall data at Penang International Airport. The overall statistical and p-values for each test are shown at Table 3.3.

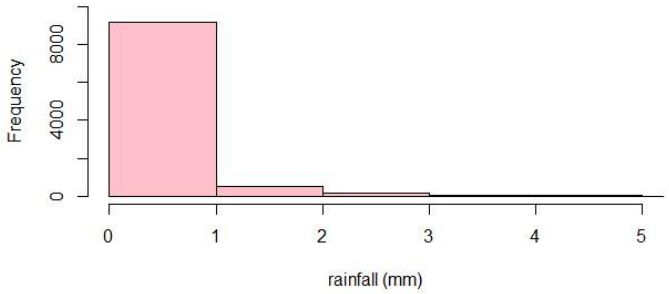


Figure 3.2. Histogram of daily rainfall data at Penang International Airport

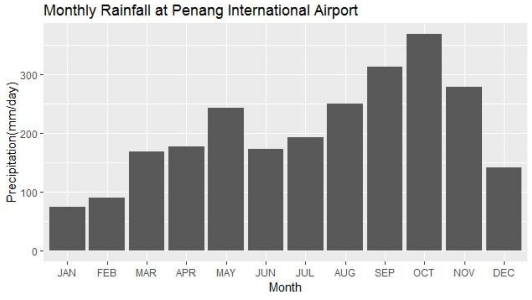


Figure 3.3. Histogram of monthly rainfall at Penang International Airport

Table 3.2. The parameter estimates for Gamma and Weibull distributions by using Maximum Likelihood Estimation (MLE)

Statistical Models	Parameters	
	α	β
Gamma distribution	2.79	0.48
Weibull distribution	1.51	1.52

Table 3.3. p-values for goodness of fit tests for Gamma and Weibull distributions

Goodness of fit tests / Statistical models	Cramer-von Mises	Kolmogorov-Smirnov
Gamma distribution	0.3288	< 2.2e-16
Weibull distribution	0.3206	< 2.2e-16

The p-value of the test is found to be greater than the significance level, α=0.05, where it is fail to reject the null hypothesis, which the data follow a specified distribution (Gamma and Weibull distributions). Based on the Table 3.3, the data are consistently follow the Gamma and Weibull distributions by using Cramer-von Mises test since the p-value is greater than 5% significance level that have been considered which are 0.3288 and 0.3206, respectively. However, the p-values of K-S test for both distributions show that the data do not follow the specified distributions.

Table 3.4. The value of test statistics using Goodness of Fit test for Gamma and Weibull distributions

Goodness of fit tests / Statistical models	Cramer-von Mises (ω^2)	Kolmogorov-Smirnov (D)
Gamma distribution	2362.2	0.76804
Weibull distribution	2224.9	0.7196

Table 3.5. Summary of the statistical test score results for Gamma and Weibull distributions

Goodness of fit tests / Statistical models	Rank		Score
	Cramer-von Mises (CvM)	Kolmogorov-Smirnov (K-S)	
Gamma distribution	1	0	(1+0)=1
Weibull distribution	2	0	(2+0)=2

Table 3.5 shows the summary for the statistical test score results for both distributions. As mentioned earlier, the test statistics value will be applied based on the value of test statistics of CvM ($\omega 2$) and K-S (D) for Gamma and Weibull distributions as provided in Table 3.4. Based on Table 3.5, both distributions are ranked 0 for K-S test since the distributions are not fitted to the data as shown in Table 3.3. In particular, rank 2 is for Weibull distribution since the test statistic for CvM is lower compared to Gamma distribution. Meanwhile, rank 1 is for Gamma distribution. The sum of the rank for each distribution, Gamma and Weibull are shown in the Table 3.5. The result shows that the Weibull distribution recorded highest total rank. Smaller value of the test statistic implies that the estimation value is closer to the data. Hence, the highest rank indicates that the data is almost perfectly fitted by the model. Therefore, Weibull distribution is the best fitted model to daily rainfall data at Penang International Airport compared to Gamma distribution.

4. Discussion and Conclusion

Rainfall modelling on the daily rainfall data is very useful in helping to understand more about the precipitation pattern especially in Malaysia due to the tropical region. By performing this study, various step of precautions can be prepared for any natural disasters that might be happened. In summary, the main study is to identify the best fitting statistical model for rainfall data based on the selected station in the state of Penang, which is Penang International Airport. The data from the year 1990 to 2017 which provided by GSOD-NOAA was used in this study. In order to determine the rainfall pattern in Penang, the suitable probability density function should be selected to give a better prediction. In order to select the best fitted distribution, two statistical

models are used namely Gamma and Weibull distributions. The shape and scale parameters for both distributions were estimated by using MLE. Different criteria of GOF had been used to test whether the model suits for the data by referring to the hypothesis testing. CvM and K-S tests as the methods of GOF with the hypothesized distributions, Gamma and Weibull distributions.

Based on the result of K-S test, both distributions are not suitable for the data with both p-values are less than 0. For another test, which is CvM test, the result shows that Gamma and Weibull distributions are fitted to the rainfall data with the p-values 0.3288 and 0.3206, higher than 5% significance level. In summarise, Weibull distribution is the best model to be fitted to the daily rainfall data at Penang International Airport based on the minimum of GOF tests (the highest statistical test score). For further study, must be focused on the performance comparison of modified distributions such as Mixed Gamma and Mixed Weibull distributions since daily rainfall amount for the majority states in Malaysia is very well represented using two components. Moreover, monsoon period also can be considered in the future study since rainfall amount in Malaysia is affected by monsoons season. This can be done by separating the time period by month according to the monsoon period. This indicated that the rainfall distribution in Malaysia is consists of minimum and maximum rainfalls. Other potential work includes of using variety of GOF test such as Anderson Darling and Chi Square tests. This may help to predict the most suitable model much precise based on the majority of the minimum GOF tests. Moreover, more stations from each state in Malaysia should be used to perform this study in future since different station of rainfall amount in Malaysia is affected by different monsoon season. This is due to the geographical of the location.

References

1. Abdullah, M. A., & Al-Mazroui, M. A. (1998). Climatological study of the southwestern region of Saudi Arabia. I. Rainfall analysis. *Climate Research*, *9*(3), 213-223.
2. Deidda, R., & Puliga, M. (2006). Sensitivity of goodness-of-fit statistics to rainfall data rounding off. *Physics and Chemistry of the Earth, Parts A/B/C*, *31*(18), 1240-1251.
3. Duan, J., Selker, J., & Grant, G. E. (1998). Evaluation of probability density functions in precipitation models for the Pacific Northwest. *Journal of the American Water Resources Association*, *34*(3), 617627.
4. Husak, G. J., Michaelsen, J., & Funk, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(7), 935-944.

5. Nilsen MA (2011). Parameter estimation for the two-parameter Weibull distribution. Msc. Thesis. Department of Statistics, Brigham Young University.
6. Oseni, B. A., & Ayoola, F. J. (2013). Fitting the Statistical Distribution for Daily Rainfall in Ibadan, Based on Chi-Square and Kolmogorov-Smirnov Goodness-Of-Fit Tests. *West African Journal of Industrial and Academic Research*, *7*(1), 93-100.
7. Predeep, N. (2016, July 19). Flooding at Penang airport due to 'unprecedented' rainfall. *Free Malaysia Today.* Retrieved from http://www.freemalaysiatoday.com/category/nation/2016/07/19/flooding-atpenang-airport-due-to-unprecedented-rainfall/
8. Rossita, M. Y., Masud, M. H., Nuradhiathy, A. R., Yong, Z. Z., & Dunn, P. K. (2017). Modelling daily rainfall with climatological predictors: Poisson-gamma generalized linear modelling approach. *International Journal of Climatology*, *37*(3), 1391-1399.
9. Sharma, M. A., & Singh, J. B. (2010). Use of probability distribution in rainfall analysis. *New York Science Journal*, *3*(9), 40-49
10. Suhaila, J., & Abdul, A. J. (2007). Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of applied sciences research*, *3*(10), 1027-1036.
11. Syafrina, A. H., Zalina, M. D., & Juneng, L. (2015). Historical trend of hourly extreme rainfall in Peninsular Malaysia. *Theoretical and Applied Climatology*, *120*(1-2), 259-285.
12. Syafrina, A. H., Norzaida, A., & Noor S. O. (2017). Rainfall analysis in the northern region of Peninsular Malaysia. *International Journal of Advanced and Applied Sciences*, *4*(11), 11-16.
13. Thom, H. C. (1951). A frequency distribution for precipitation. *Bulletin of the American Meteorological Society*, *32*(10), 397.

Index

Index

ISIWSC2019

Organised by :

DEPARTMENT OF STATISTICS MALAYSIA
MINISTRY OF ECONOMIC AFFAIRS

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

Supported by:

MINISTRY OF TOURISM, ARTS & CULTURE MALAYSIA

Malaysia Convention & Exhibition Bureau

Visit Truly Asia Malaysia 2020

ISBN 978-967-2000-72-3

#ISIWSC2019