62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 4

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 4)

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 4, 2019. 451 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 4

Nonnegative matrix factorization: A semi-parametric statistical view and selection model

Bertail Patrice[1], Cl´emen¸con St´ephane[2], Zetlaoui M´elanie[1]
[1]Paris Nanterre University, Nanterre, France,
[2]T´el´ecom ParisTech, Paris, France

Abstract

The goal of *Nonnegative Matrix Factorization* (NMF) consists in finding a convex cone in the positive orthant," representing accurately" a cloud of multivariate nonnegative data. The dimension of the convex cone is assumed to be smaller than the dimension of the data space. Whereas the majority of the literature dedicated to NMF focused on algorithmic issues related to the computation of representations maximizing some goodness-of-fit criterion, statistical grounds for such *M*-estimation techniques have not been exhibited yet. Here, we investigate the semiparametric framework: through the specification of a variety of probabilistic generative models and under statistical identifiability assumptions and we can construct a Z-estimator with estimated nuisance parameters based on the efficient score. Under appropriate assumptions, this Z-estimator yields asymptotically normal estimates of C's rays. In this context, model selection issues related to the dimension of the underlying cone C are considered through the AIC and BIC approaches. We show, under regularity assumptions, that we can recover the optimal number of C's rays.

Keywords

Nonnegative matrix factorization; latent variable model; semiparametric estimation; identifiability; model selection; efficient scores.

1. Introduction

In a wide variety of applications, data are nonnegative by nature: pixel intensities, amplitude spectra, occurrence counts, food consumption, user scores, stock market values, *etc.* Nonnegative matrix factorization (NMF) precisely aims at finding (linearly independent) *latent vectors* with nonnegative coordinates, of which observations can be viewed as convex linear combinations. Originally proposed by [7] in the context of facial images analysis, NMF has recently received a good deal of attention in the fields of machine learning and signal/image processing and has been applied to a variety of applications in different fields.

Whereas the design of NMF computational techniques has been the subject of intense research these last few years in the signal processing and

machine-learning communities (see [5] for instance), no rigorous asymptotic framework for statistical recovery of the NMF, even in a simple parametric setup, has been given yet in the statistical learning literature. It is the goal of this paper to formulate NMF as an identifiable statistical problem, for which *M*-estimation techniques in a semiparametric context, yield consistent estimates. We will compute the efficient scores (see the terminology in [3]for instance), the efficiency bound and propose new efficient semiparametric estimation methods based on a estimated version of efficient score. We will see that the NMF model has some strong links with the dimension reduction method considered in single index models so that the recent paper by [8] is also of interest for our work.

It is next shown how to use popular model selection methods in order to choose the number of latent vectors involved in the NMF representation. Consistency of the maximum-penalized-likelihood estimator is proved in this context, when the penalty term is the Bayesian Information Criterion. Finally, these approaches are illustrated by preliminary simulation results.

2. Background theory and concepts

In the following, for any $(p,q) \in \mathbb{N}^{*2}$, we denote by $\mathcal{M}_{pq}(\mathbb{R}_+)$ the space of $p \times q$ matrices with nonnegative entries. $\det(M)$ is the determinant of any square matrix M with real entries, A^t denotes the transpose of any rectangular matrix A . $\|.\|$ is the euclidian norm on $\mathbb{R}^F$. The indicator function of any event E is denoted by $\mathbb{I}\{E\}$. Finally, we use Φ_F for the characteristic function of any probability distribution F on $\mathbb{R}^F$ and by "$\Rightarrow$" the convergence in distribution. If a rectangular matrix A is full rank, we denote by A^{-1} the Moore-Penrose generalized pseudo-inverse of A, refer to [1]. Recall that we have $A^{-1} = (A^tA)^{-1}A^t$, denoting by M^{-1} the standard inverse of a square matrix M and by Q^t the transpose of any matrix Q, see [4] for instance.

Let $F \geq 1$ be the dimension of the space where the observations lie. The NMF task can be formulated as follows. One observes (column) vectors $v_i = (v_{1i}, ..., v_{Fi})$, $1 \leq i \leq n$, with nonnegative coefficients: $\forall (f,i) \in \{1, ..., F\} \times \{1, ..., n\}$, $v_{fi} \geq 0$. It is believed that these data can be 'well described' by a *conical hull* generated by $K \leq F$ linearly independent vectors $W_{.1}, ..., W_{.K}$ lying in the positive orthant $\mathbb{R}_+^F$ that is

$$C_w = \left\{ \sum_{k=1}^{K} h_k W_k : h_k \geq 0 \right\} \tag{1}$$

With $W_{fk} \geq 0$ for all $(f, k) \in \{1,...,F\} \times \{1,...,K\}$.

Assume that the observed data are i.i.d. copies of the random vector:

$$v = Wh \tag{2}$$

where $W \in \mathcal{M}_{FK}(\mathbb{R}_+)$ and h is a random column vector of length K with distribution $G(dh)$ supported by the positive orthant $\mathbb{R}_+^K$. In the following we

will assume that G has a density with respect to the Lebesgue measure, denoted by belonging to some regular space G (we will precise later the regulatity assumption needed on G ensuring convergence of our estimators) . The distribution of the r.v. v is entirely determined by the law of the pair (W,G): we denote it by $P_{W,g}$ and $p_{W,g}$ its density . The semiparametric model is thus entirely described by the set

$$\mathbb{P}_{\mathcal{W}_+,\mathcal{G}} = \left\{ p_{W,g} = \frac{dP_{W,g}}{d\lambda},\ W \in \mathcal{W}_+,\ g \in \mathcal{G} \right\}.$$

In the semiparametric terminology W is the parameter of interest and g is the nuisance parameter.

Remark : This model can be extended to noisy models. Although additive noises are generally modelled in an additive way in most applications, it should be noticed that multiplicative and poisson noise models may also be used to take advantage of leading to nonnegative data vectors in the context of NMF. Hence, as highlighted in [6], uniqueness of matrices (W,H) cannot be guaranteed in absence of further assumptions on W and/or H's distribution.

We now set the hypotheses which will be assumed throughout this paper and ensuring the existence and the unicity of the representation in a semiparametric framework.

- **H_1** The matrix W is of full rank K, $1 \leq K \leq F$.
- **H_2** The columns of the matrix W are of unit (euclidian) norm: $\forall k \in \{1, ..., K\}$, $\|W_{.k}\|^2 = 1$.
- **H_3** The columns of the matrix W are sorted by *lexicographic order* of the vectors $(\alpha_{1,k}(W), ..., \alpha_{F,k}(W))$.
- **H_4** The span of the support of the distribution of the v'*is is*denoted by $G(dh)$ is $\mathbb{R}_+^K$.
- **H_5** The distribution $G(dh)$ is such that for any $k \in \{1, \ldots, K\}$, $\text{SUPP}(G) \cap \mathbb{R}_+^* \cdot W_k \neq \emptyset$, where $\mathbb{R}_+^* \cdot w = \{\lambda w = (\lambda w_i) : \lambda > 0\}$ for any $w = (w_i) \in \mathbb{R}^F$ and supp(G) denotes the support of h's distribution.

We will denote by W_+ the set of matrices $W \in M_{FK}(R_+)$ fulfilling assumptions **H_1** – **H_3**.

Theorem (Semi-parametric identifiability in NMF models) Let G be a set of probability distributions on $\mathbb{R}_+^K$. Assume that all the distributions in G fulfill assumptions **H_4** –**H_5**. The family of distributions $\{P_{W,G}^{(i)} : (W,G) \in \mathcal{W}_+ \times \mathcal{G}\}$ is then identifiable.

Now, under assumption **H_1**, the likelihood $p_{W,g}$ is given by [2]

$$p_{W,g}(v) = vol(W^{-1})g(W^{-1}v). \quad (3)$$

In that case the likelihood of this semi-parametric model based on a sample $\mathbf{v}_n = (v_1, ..., v_n)$ of n independent copies of the random variable v, is simply given by

$$L_g(\mathbf{v}_n; W) = \left(\text{vol}(\text{W}^{-1})\right)^n \prod_{i=1}^{n} g(W^{-1}v_i), \quad (4)$$

where $\mathrm{vol}(W^{-1}) = {}^{p}\det(W^{-1}(W^{-1})^{t})$

3. Scores and Tangent spaces of the semiparametric model

The main idea of semiparametric model is to consider square root of density as element of the Hilbert space $L_2(\lambda)$. In the following we will consider densities which are differentiable in quadratic mean (DQM) that is such that for any parametric model p_t $t \in [0,1]$ in $P_{W_+,G}$, there exists a score function s such that

$$\int \left(\frac{p_t^{\frac{1}{2}} - p^{\frac{1}{2}}}{t} - \frac{1}{2} s p^{\frac{1}{2}} \right)^2 d\lambda \underset{t \to 0}{\longrightarrow} 0 .$$

In particular when it is assumed that $g \in G$ and G is regular and differentiable in (that is any density in G admit a score function s_g) then $p_{W,g}(v) = \mathrm{vol}(W^{-1}) g(W^{-1}v)$ is automatically differentiable in quadratic mean. The efficiency bounds and the efficient score may be obtained by computing respectively the scores with respect to the parameter of interest (for us W) and with respect to the nuissance parameter (for us g) and then by projecting the score function with respect to W into the orthogonal space of the tangent space engendered by the scores with respect to nuisance parameters. It follows that

$$s^{eff}_{W,g}(v) = P_{\dot{L}^{\perp}} \sum_{l=1}^{K} \frac{\partial [W^{-1}]_{l\cdot}}{\partial W_{fk}} (v - E\left(v | W^{-1}v\right)) \quad \frac{\frac{\partial g}{\partial h_l}(W^{-1}v)}{g(W^{-1}v)}$$

with

$$\dot{L} = \begin{pmatrix} W^t_{\cdot 1} & 0 & \dots & \dots & 0 \\ 0 & W^t_{\cdot 2} & 0 & \dots & 0 \\ & & & & \\ & & & & 0 \\ 0 & & & & W^t_{\cdot K} \end{pmatrix} \in \mathcal{M}_{K\ F \times K}\ (R_+)$$

and

$$P_{\dot{L}^{\perp}} = I_{F \times K} - \dot{L}^t (\dot{L}\dot{L}^t)^{-1} \dot{L}$$

Estimation of the parameters. Let $v_1, \dots, v_n$ be i.i. d. $P_{W,g}$. In theory an (oracle) estimator would be given by solving the M-equation

$$\sum_{i=1}^{n} s^{eff}_{W,g}(v_i) = 0$$

However since these quantities depends on g as well as some other unknown non-parametric quantities depending on g mainly $E\left(v | W^{-1}v\right)$, $\frac{\partial g}{\partial h_l}(W^{-1}v)/g(W^{-1}v)$, we will replace the efficient score by an estimated score, using the same ideas as in [8]. We will focus here on the simple Nadaraya estimator, with smoothing parameter b_n and kernel density κ given by

$$\widehat{g}_n(x, W) = \frac{1}{n b_n} \sum_{i=1}^{n} \kappa \left(\frac{x - W^{-1}v_i}{b_n} \right)$$

Following [8], it is easy to estimate $E\left(v|W^{-1}v\right)$ by a Nadaraya-Watson estimator, denoted by $\widehat{E}\left(v|W^{-1}v\right)$.

Then we have an estimated efficient score function given by

$$\widehat{s^{eff}_{W,g}}(v) = P_{\dot{L}^{\perp}} \sum_{l=1}^{K} \frac{\partial [W^{-1}]_{l\cdot}}{\partial W_{fk}} (v - \widehat{E}\left(v|W^{-1}v\right)) \ \widehat{l}_{\kappa,n}(v,W)$$

where $\widehat{l}_{\kappa,n}(v,W)$ is the estimated value of $\frac{\frac{\partial g}{\partial h_l}(W^{-1}v)}{g(W^{-1}v)}$.

Let the following conditions:

H$_6$ $a_n \to 0,\ b_n \to 0,\ na_n^3 \to \infty,\ nb_n^3 \to \infty$

H$_7$ κ admits a 3rd order continuous derivative, square integrable and ultimately monotone.

Theorem (Consistency of the estimator) Under the assumptions H_1,...H_7, the estimator which is the solution of the estimating equation say $\widehat{W}_n, \sum_i^n \widehat{s^{eff}_{\widehat{W}_n,g}}(v_i) = 0$, is efficient and asymptotically that is to say :

$$\sqrt{n}(vec(\widehat{W}_n) - vec(W)) \underset{n\to\infty}{\to} N\left(0, V_{P_{W,g}}\left(s^{eff}_{W,g}(v)\right)\right)$$

4. Model selection

Here, we consider the sequence of nested zero-noise NMF models, indexed by $K \in \{1, ..., F\}$, parametrized by the set $\mathcal{W}^{(K)} \subset \mathcal{W}_+^{(K)}$. Let

$$\widehat{K}_n = \underset{1\le K\le F}{\arg\min}\, C_n(K),$$

where $C_n(K) = -2logL_g(v_n; W) = (vol(W^{-1}))^n \Pi_{i=1}^n g(W^{-1}v_i) + a_n(K)$ where $a_n(K)$ is a penalty satisfying the following assumption :

H$_8$ Let the penalization $a_n(K)$, be an increasing sequence (in K) such that, for any $K_1 > K_2 > 0$, $a_n(K_1) - a_n(K_2) \to \infty$ and such that, for any K, $\frac{a_n(K)}{n} \to 0,\ \text{as } n \to \infty$.

This is for instance the case for the BIC penalization criterion obtained with $a_n(K) = \log(n) \cdot d_K$, where the dimension d_K is increasing in K.

We will show the consistency of $\widehat{K}_n$ for general penalization under the following additional assumptions. **H**$_9$ The function $g(W^{-1}v)$ satisfies an uniform Lipschitz condition of the form

$$\left|g(W_1^{-1}v) - g(W_2^{-1}v)\right| \le M_*||W_2 - W_1||_2,$$

for some $M_* > 0$.

This is true in particular, if the derivative of $g(W^{-1}v)$ with respect to W is bounded over the sets W^K, $K = 1,...,F$.

Theorem 1 (Consistent estimation of the cone dimension)

Suppose that the assumptions H_1,...,H_5 are satisfied, as well as the additional assumptions H_8,H_9 and H_{10}. Then, as $n \to \infty$, we almost surely have $K_n \to K^*$.

A toy numerical experiment. We chose $F = K = 2$ and 200 observations have been generated based on model (2), where the h_{kn}'s are i.i.d. r.v.'s drawn

from the Gamma distribution $\Gamma(1,1)$ and the rays $W_{.1}$ and $W_{.2}$ are defined by the angles $\alpha_1 = \pi/12$ and $\alpha_2 = \pi/3$ they respectively form with the v_1-axis. Fig. below shows the related data cloud together with the log-likelihood surface evaluated on a grid of 21x21, between $[0,\pi/2]$, for (α_1,α_2): ML estimates coincide with the generating W.

The model selection is illustrated by Fig. 2, for the same type of model as previously, except that $F = 10$ and $K = 5$. The model selection procedure allows to recover the true dimension.

5. Conclusions

In this paper we have formulated the NMF as a statistical problem. For different generative models, identifiabilty of the related statistical models have been investigated from a semi-parametric angle. A semi-parametric statistical framework have been then proposed, where the proposed Z-estimator is asymptotically consistent. Finally, we have shown how to use information criteria such as BIC or AIC for parametric model selection purposes in the NMF context.

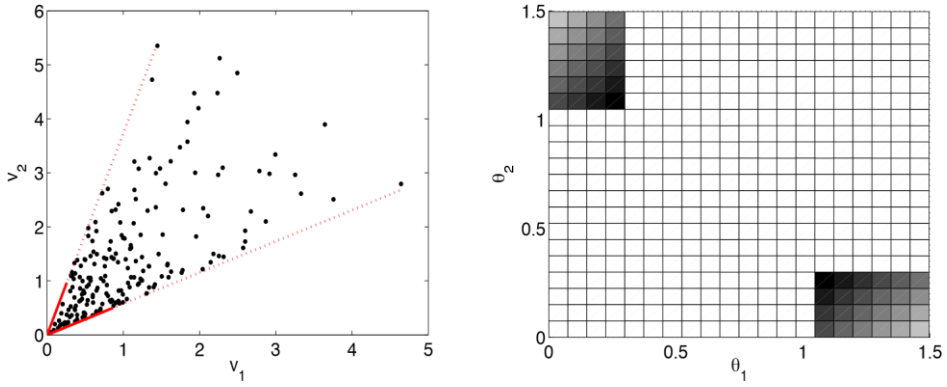


Figure 1: A $2-d$ NMF toy example: cone and likelihood.

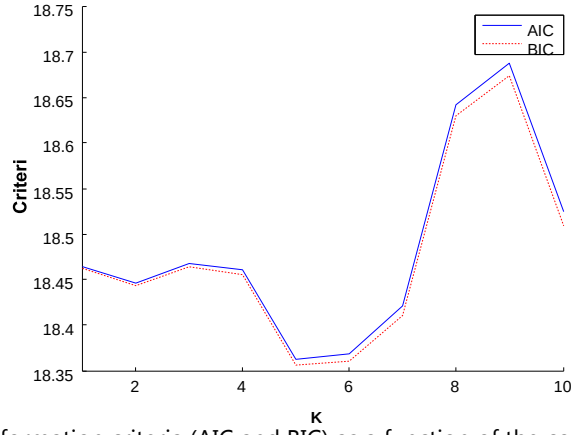


Figure 2: Information criteria (AIC and BIC) as a function of the cone dimension.

References

1. Ben-Israel, A. (1992) A volume associated with $m \times n$ matrices. Lin. Algeb. Appl.
2. Ben-Israel, A. (1999) The change of variables formula using matrix volume. SIAM J. Matrix Analysis.
3. Bickel, P. J. & Klaassen C. AJ & Ritov Y. & Wellner J. A (1998) Efficient and adaptive estimation for semiparametric models. Springer-Verlag.
4. Campbell S.L. & Meyer C.D. (2009). Generalized inverses of linear transformations. Classics in Applied Mathematics. SIAM.
5. Cichocki,A. & Zdunek, R. & Amari,S. (2006) Csiszars divergences for non-negative matrix factorization: Family of new algorithms. In 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA06).
6. Donoho,D. & Stodden,V (2004). When does non-negative matrix factorization give a correct decomposition into parts? In Advances in Neural Information Processing Systems 16, Cambridge, MA.
7. Lee, D. & Seung, H. (1999). Learning the parts of objects by nonnegative matrix factorization. Nature.
8. Ma,Y. & Zhu, L. (2012). A semiparametric approach to dimension reduction. Journal of the American Statistical Association.

A linear mixed model for segmented regression with smooth transition

Julio M. Singer[1], Francisco M.M. Rocha[2], Antonio Carlos Pedroso-de-Lima[1], Giovani, L. Silva[3], Giuliana C. Coatti[4], e Mayana Zatz[4]

[1]Department of Statistics, University of Sao Paulo

[2]Paulista School of Politics, Economics and Business, Federal University of Sao Paulo

[3]Department of Mathematics, University of Lisbon

[4]Institute of Biosciences, University of Sao Paulo

Abstract

We consider random changepoint mixed segmented regression models to analyse data obtained from a study conducted to verify whether treament with stem cells may delay the onset of a symptom of amyotrophic lateral sclerosis in genetically modified mice. The proposed models capture the biological aspects of the data, accommodating a smooth transition between the periods with and without symptoms. An additional changepoint is considered to avoid negative predicted responses. Given the non-linear nature of the model, we adapt an algorithm proposed by Muggeo et al. (2014, Statistical Modelling) to estimate the fixed parameters and to predict the random effects by fitting linear mixed models at each step.

Key words

amyotrophic lateral sclerosis; fitting algorithm; mixed models; random effects.

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is one of the most common adult-onset motor neuron disease causing a progressive, rapid and irreversible degeneration of motor neurons in the cortex, brain stem and spinal cord. In the majority of cases ALS occurs sporadically; in about 10% of the cases it is caused by familial reasons. No effective treatment is available and cell therapy clinical trials are currently being tested in ALS affected patients. The SOD1 gene encodes an important antioxidant human enzyme and mutations in SOD1 represent one of the most frequent causes of ALS.

Among the different animal models for ALS, SOD1 mice are the most used in pre-clinical studies. After the initial tremor in the limbs they develop muscle weakness in early adulthood, become fully paralyzed and die. These mice over-express the human SOD1 gene bearing the G93A mutation, a point mutation found in familial ALS. Interestingly, in this animal model the disease progression is different between the genders. Males have a shorter lifespan and a clinical condition apparently more severe than females and differences

in electrophysiological parameters have also been reported. A comparable effect of gender is also observed in ALS patients.

Treatment of ALS with stem cells is a current research topic. Mesenchymal stromal cells (MSC), specially those derived from adipose tissues, and pericytes have been used in studies that focus on the reduction of the speed of the progression of symptoms of neuro-degenerative diseases. In this context we consider a study conducted in the Human Genome and Stem Cell Research Center, at the Biosciences Institute, University of S˜ao Paulo, Brazil with the objective of comparing MSC cells and pericytes injected in SOD1-G93A mice with respect to their effects on the evolution of some symptoms of ALS. Details may be obtained in Coatti et al. (2017).

Our objective here is to propose models for the statistical analysis of the data.

2. The study

A set of 34 female and 21 male 8 week old SOD1-G93A mice was divided into 3 groups Animals in the first group (12 females and 7 males) were submitted to weekly injections of MSC cells, those in second group (11 females and 8 males), to injection with pericytes while animals in the third group (11 females and 6 males) were submitted to the vehicle (*Hank's balanced salt solution* - HBSS). All animals were followed weekly up to their death for clinical analysis of the progression of the disease by means of four variables, the analysis of one of them, *rotarod* is considered in this study. The *rotarod* test was used to evaluate motor coordination and fatigue resistance. For that purpose, the length of time each animal could remain on the rotating cylinder (3.5 cm) of a *rotarod* apparatus (IITC Life Science model 755) was recorded. The initial speed was 1 rpm and it was increased constantly until a final speed of 30 rpm, after 180 s. Each animal was given three tries and the longest latency to fall was recorded. The specific objectives of the analysis are:

i) Identification of the moment when animals become symptomatic (symptoms onset) forthe six groups defined by the combination of treatment (HBSS, MSCs, pericytes) and sex (male, female).
ii) Estimation of the expected rate of variation in response after symptom onset for eachgroup.
iii) Evaluation of the effects of treatment, sex and their interaction on the expected momentof symptom onset and post onset rate of variation in the expected response.

3. Statistical analysis

Profile plots for the response along with LOESS curves are displayed in Figure 1.

A longitudinal analysis of the behaviour of the response variable corroborates its expected stable level before the onset of the symptom (a decrease in the length of time during which the animal remains in the rotating cylinder). Furthermore, individual differences in the moment where this occurs as well as differences among the accelerations with which the intensity of the symptom progresses are also visible. It also seems reasonable to expect a change in the acceleration with which the intensity of the symptom progresses after the disease onset.

Given that such conclusions are in line with the expected biological behaviour, a random changepoint mixed polynomial segmented regression model may be considered for the analysis.

Such models have an attractive practical appeal in many fields and have been the object of statistical research for a long time as detailed in Muggeo et al. (2014). These authors consider a frequentist approach as opposed to the commonly Bayesian perspective usually employed in the statistical literature.

Keeping in mind the necessarily non-negative nature of the response, we adopt a similar approach and consider an analysis of the ALS data based on the model

$$y_{ijk} = \alpha_{ij} I(t_k < \psi_{2ij}) + \gamma_{ij}[t_k - \psi_{1ij}(\lambda_{ij})]^2 I(\psi_{1ij} \leq t_k < \psi_{2ij}) + e_{ijk} \qquad (1)$$

($i = 1,...,6$, $j = 1,...,n_i$ and $k = 1,...,n_{ij}$) where y_{ijk} denotes the response for the j-th animal observed in the i-th group (defined by the combination of the levels of treatment and sex) at the k-th evaluation instant, α_{ij} is the corresponding stable level of the symptom prior to the first changepoint, γ_{ij} is the coefficient of the quadratic term for the curve that governs the response behaviour post changepoint ψ_{1ij}, with

$$\psi_{1ij}(\lambda_{ij}) = [L_1 + L_2 \exp(\lambda_{ij})]/[1 + \exp(\lambda_{ij})]$$

to restrict the value of ψ_{1ij} to the interval (L_1, L_2) in which the observations are obtained and ψ_{2ij} denotes the instant where the response is null. We assume that $\alpha_{ij} = \alpha_i + a_{ij}$, $\gamma_{ij} = \gamma_i + c_{ij}$, $\lambda_{ij} = \lambda_i + \ell_{ij}$ with $\mathbf{b}_{ij} = (a_{ij}, c_{ij}, \ell_{ij})^\top \sim N(\mathbf{0}, \mathbf{G}_i)$ and $e_{ijk} \sim N(0, \sigma_i^2)$ independent of $\mathbf{b}_{ij}$.

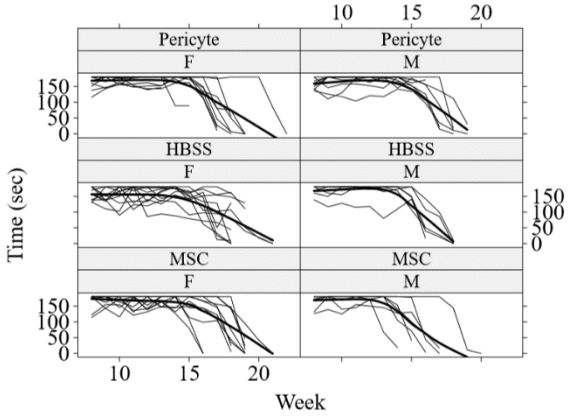


Figure 1: Profile plots for the response along with LOESS curves

This is an extension of the models proposed by Muggeo et al. (2014) where a smooth transition and a second changepoint are incorporated. For the sake of notational simplicity and without loss of generality, we drop the subscript i to specify the the fitting algorithm.

Given that ψ_{2j} corresponds to the instant t_k where $y_{jk} = 0$, we have $I(t_k < \psi_{2j}) = 1$ and $I(\psi_{1j} \leq t_k < \psi_{2j}) = 1$ and consequently, that $\alpha_j + \{\gamma_j[\psi_{2j} - \psi_{1j}(\lambda_j)]^2\} = 0$, implying that

$$\psi_{2j} = \psi_{2j}(\alpha_j, \gamma_j, \psi_{1j}) = \sqrt{-\alpha_j/\gamma_j} + \psi_{1j}(\lambda_j)$$

Following Muggeo et al. (2014) and Fasola et al. (2018), the model, which is non-linear, may be approximated by a first order Taylor expansion of $f[t_k, \gamma_j, \psi_{1j}(\lambda_j)] = \gamma_j[t_k - \psi_{1j}(\lambda_j)]_2 I(\psi_{1j} \leq t_k < \psi_{2j})$.
Explicitly,

$$f[t_k, \gamma_j, \psi_{1j}(\lambda_j)] \approx f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] + (\lambda_j - \widehat{\lambda}_j)\frac{\partial f[t_k, \gamma_j, \psi_{1j}]}{\partial \psi_{1j}}\frac{\partial \psi_{1j}(\lambda_j)}{\lambda_j}\Big|_{\lambda_j = \widehat{\lambda}_j}$$

with

$$\frac{\partial f[t_k, \gamma_j, \psi_{1j}]}{\partial \psi_{1j}} = h_j(\lambda_j) = 2\gamma_j[t_k - \psi_{1j}(\lambda_j)]I[\psi_{1j}(\lambda_j) \leq t_k < \psi_{2j}]$$

and

$$\frac{\partial \psi_{1j}(\lambda_j)}{\partial \lambda_j} = g_j(\lambda_j) = \frac{(L_2 - L_1)\exp(\lambda_j)}{[1 + \exp(\lambda_j)]^2}.$$

Consequently we may approximate model (1) by

$$y_{jk} \approx \alpha_j I(t_k < \psi_{2j}) + f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] - \widehat{\lambda}_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + \lambda_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + e_{jk}. \quad (2)$$

Considering the pseudo observations defined by $y^*_{jk} = y_{jk} + \widehat{\lambda}_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j)$, the model

$$y^*_{jk} = \alpha_j I(t_k < \psi_{2j}) + f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] + \lambda_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + e_{jk}$$

suggests the following algorithm to fit (1)

1) Let $\psi_{1j}^{(0)} = \psi_1^{(0)}$ and $\psi_{2j}^{(0)} = \psi_2^{(0)}$.

2) Fit model $y_{jk} = \alpha_j I(t_k < \psi_{2j}^0) + \gamma_j (t_k - \psi_{2j}^{(0)})^2 I(\psi_{1j}^{(0)} \leq t_k < \psi_{2j}^{(0)}) + e_{jk}$ to obtain $\alpha^{(0)}$, $a_j^{(0)}$, $\gamma^{(0)}$, $c_j^{(0)}$, $\lambda_j^{(0)} = \log[(\psi_{1j}^{(0)} - L_1)/(L_2 - \psi_{1j}^{(0)})]$ and $\psi_{2j}^{(1)} = \sqrt{-\alpha_j^{(0)}/\gamma_j^{(0)}} + \psi_{1j}^{(0)}$.

3) Let $r = 1$.

4) Compute $y_{jk}^{(r)} = y_{jk} + \lambda_j^{(r-1)} h_j(\lambda_j^{(r-1)}) g_j(\lambda_j^{(r-1)})$.

5) Fit model

$$y_{jk}^{(r)} = \alpha_j I(t_k < \psi_{2j}^{(r)}) + \gamma_j [t_k - \psi_{1j}^{(r)}]^2 I(\psi_{1j}^{(r)} \leq t_k < \psi_{2j}^{(r)}) + \lambda_j h_j(\lambda_j^{(r-1)}) g_j(\lambda_j^{(r-1)}) + e_{jk}^{(r-1)}$$

to obtain $\alpha^{(r)}$, $a_j^{(r)}$, $\gamma^{(r)}$, $c_j^{(r)}$, $\lambda^{(r)}$, $\ell_j^{(r)}$, $\psi_{1j}^{(r)} = [L_1 + L_2 \exp(\lambda_j^{(r)})]/[1 + \exp(\lambda_j^{(r)})]$ and $\psi_{2j}^{(r+1)} = \sqrt{-\alpha_j^{(r)}/\gamma_j^{(r)}} + \psi_{1j}^{(r)}$.

6) Stop if some convergence criterion is satisfied, otherwise, let $r = r + 1$ and repeat steps 4-6.

This algorithm, adapted from Muggeo et al. (2014), essentially considers iterative fitting of standard linear mixed models by restricted maximum likelihood. At convergence, we expect a neglible difference between the third and fourth terms in the right hand side of (2) and as a consequence, that the pseudo observations should well approximate the original ones. Given the linear mixed model nature of the proposed fitting algorithm, we may employ the diagnostic procedures outlined in Singer et al. (2017) to check whether the adopted assumptions for the distribution of the random effects or of the random error are reasonable.

Estimates of the parameters of model (1) obtained via fitting the approximation (2) along with the corresponding standard errors are summarized in Table (1).

The results of a Wald test for the homogeneity of the six changepoints ψ_1 ($\chi^2 = 58.30, df = 5, p < 0.001$) suggests further analyses to identify the possible effects of treatment, sex and their interaction. A significant interaction between treatment and sex with respect to the ψ_1 changepoints ($\chi^2 = 13.65, df = 2, p = 0.001$) may be analysed via the multiple comparisons summarized in Table 2 and suggest that the onset of symptoms for the control group (HBSS) males is delayed by 1.7 [CI(95%) = 1.0, 2.4] weeks with respect to the control group females and that treatment with Pericytes (both sexes) or MSC (females) delay the onset of symptoms by 1.3 [CI(95%) = 0.5, 2.2] weeks with respect to HBSS treated males. The changepoint for MSC treated males lies between those for HBSS treated males and females but the small sample size does not lead to a significant difference in either case.

The results for a similar analysis of the acceleration with which the symptom progresses suggest no difference between sexes and an increase in the acceleration of 18.6 [CI(95%) = 17.8, 19.6] sec/week^2 for the experimental

treatments (MSC and Pericytes) relatively to that of the control treatment (HBSS).

An example of predicted subject specific response curves is presented in Figure 2.

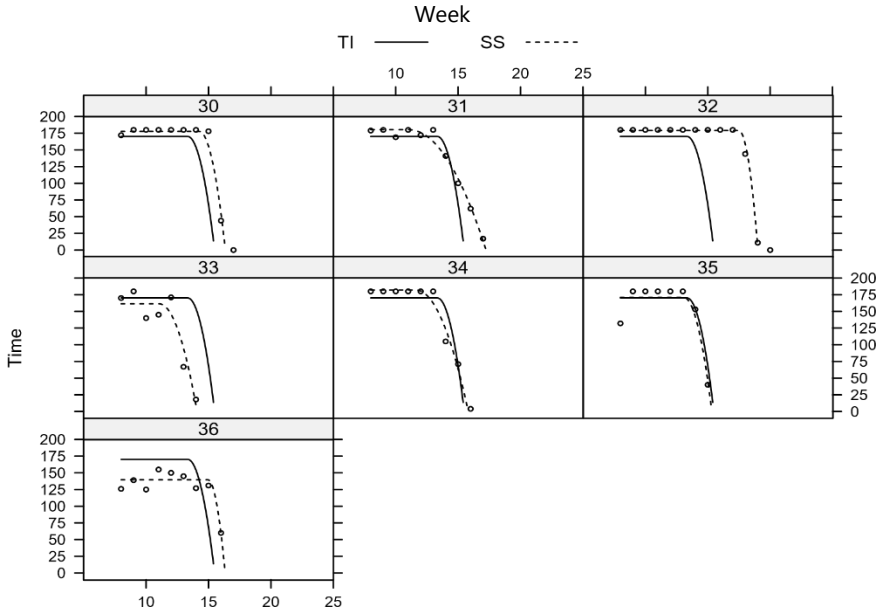
Parameter	Sex	Treatment	Estimate	Std error	
				Model	Robust
Intercept (α)	M	HBSS	166.5	10.0	9.1
	M	MSC	170.2	6.3	5.8
	M	Pericytes	164.5	6.5	6.1
Intercept (α)	F	HBSS	156.6	6.5	6.2
	F	MSC	167.4	3.8	3.6
	F	Pericytes	170.2	3.6	3.4
2nd degree coefficient (γ)	M	HBSS	-18.2	4.7	4.2
	M	MSC	-36.7	11.4	10.4
	M	Pericytes	-23.1	12.5	11.7
2nd degree coefficient (γ)	F	HBSS	-2.8	1.0	0.9
	F	MSC	-26.6	6.3	6.0
	F	Pericytes	-30.0	7.8	7.3
Changepoint 1 (ψ_1)	M	HBSS	13.9	0.2	0.2
	M	MSC	13.3	1.0	0.9
	M	Pericytes	14.9	0.4	0.4
Changepoint 1 (ψ_1)	F	HBSS	12.1	0.3	0.2
	F	MSC	15.5	0.5	0.5
	F	Pericytes	15.2	0.8	0.8
Changepoint 2 (ψ_2)	M	HBSS	16.9	0.3	0.3
	M	MSC	15.5	0.7	0.7
	M	Pericytes	17.6	0.5	0.4
Changepoint 2 (ψ_2)	F	HBSS	19.6	1.1	1.0
	F	MSC	18.0	0.4	0.4
	F	Pericytes	17.6	0.6	0.6

Table 1: Estimates and standard errors for the parameters of model (1) obtained via fitting the approximation (2) along with robust counterpart of the standard errors

Comparison	Changepoint		
	χ^2	df	p-value
Sex within HBSS	23.61	1	< 0.001
Sex within MSC	3.53	1	0.060
Sex within Pericytes	0.14	1	0.713
Pericytes = MSC(F)	0.75	1	0.688
Pericytes + MSC(F) = HBSS(M)	10.16	1	0.001
MSC(M) = HBSS(M)	0.25	1	0.615
MSC(M) = HBSS(F)	1.35	1	0.245

Table 2: Comparisons of changepoints (ψ_1)

Figure 2: Predicted subject specific response curves (MSC males)



Acknowledgements

This research received partial financial support from Conselho Nacional de Desenvolvimento Cient´ıfico e Tecnol´ogico (CNPq, grant 3304126/2015-2) and Fundac¸˜ao de Amparo `a Pesquisa do Estado de S˜ao Paulo (FAPESP, grant 2013/21728-2), Brazil.

References

1. Coatti, G.C. (2015). Avalia¸c˜ao do potencial terapˆeutico de pericitos e de c´elulas mesenquimais no camundongo SOD1, modelo animal para esclerose lateral amiotr´ofica. *Tese de doutorado, Departamento de Bioc^encias, Universidade de S˜ao Paulo*. http://www.teses.usp.br/teses/disponiveis/41/41131/tde-14012016-143346/pt-br.php
2. Fasola, S., Muggeo, V.M.R. and Ku¨chenhoff, H. (2018). A heuristic, iterative algorithm for change-point detection in abrupt change models. *Compuational Statistics* 33, 997-1015.
3. Muggeo, V.M.R., Atkins, D.C., Gallop, R.J. and Dimidjian, S. (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling* 14, 293-313.
4. Singer, J.M., Rocha, F.M.M. and Nobre, J.S. (2017). Graphical tools for detecting departures from linear mixed models assumptions and some remedial measures. *International Statistical Review* 85, 290-324.

Joint asymptotic normality of stopping times and sequential estimators in monitoring autoregressive processes

K. Nagai[1], K. Hitomi[2], Y. Nishiyama[3], J. Tao[1]
[1]Yokohama National University, Yokohama, Japan
[2]Kyoto Institute of Technology, Kyoto, Japan
[3]Kyoto University, Kyoto, Japan

Abstract
We consider the joint asymptotic properties of stopping times and sequential estimators for a stationary rst-order autoregressive process (AR(1)) with independent and identically distributed (i.i.d.) errors with mean 0 and nite variance. Lai and Siegmund (1983) de ned two stopping times based on the observed Fisher information. The rst stopping time is de ned to be the rst time at which the observed Fisher information with known variance of errors exceeds a prescribed level. The second one is de ned by replacing the variance of errors with its estimator. They derived the almost sure convergence of the stopping times to some constant for a stationary AR(1). Using a functional central limit theorem for nonlinear ergodic stationary processes and Skorohod's representation theorem, we show that the stopping times, the sequential least square estimators, and the estimator of the variance of errors have the joint asymptotic normality. We also nd that the asymptotic variance of the rst stopping time is strictly greater than that of the second one.

Keywords
Statistical process monitoring; Observed Fisher information; Fixed accuracy estimation; Functional central limit theorem; Skorohod's representation theorem

1. Introduction

Consider a AR(1) process $\{x_n\}$ on a probability space (Ω,F,P),

$$x_n = \beta x_{n-1} + \epsilon_n, \quad n = 1, 2, \ldots. \tag{1}$$

We assume that $\epsilon_1, \epsilon_2, \cdots$ are independent, identically distributed random variables with $E(\epsilon_1) = 0,\ 0 < E(\epsilon_1^2) = \sigma^2 < \infty$ and that an initial value $x_0 \in L^2$ is independent of $\{\epsilon_n\}$. We consider two cases; the stationary case: $|\beta| < 1$ and the unit root case: $\beta = \pm 1$.

The least square estimate is

$$\hat{\beta}_N = \sum_{n=1}^{N} x_n x_{n-1} \Big/ \sum_{n=1}^{N} x_{n-1}^2 \tag{2}$$

It's well known that when the process is a stationary AR(1), the least square estimate $\hat{\beta}_N$ has asymptotic normality; as $N \to \infty$,

$$\sqrt{N}\left(\hat{\beta}_N - \beta\right) \to N(0, 1-\beta^2). \quad (3)$$

For the case that $\epsilon_1, \epsilon_2, \cdots$ are normally distributed, the observed Fisher information about β is given b

$$I_N = -\frac{\partial^2}{\partial\beta^2}\left(\beta\sum_{n=1}^{N} x_{n-1}x_n - \frac{1}{2}\beta^2\sum_{n=1}^{N} x_{n-1}^2\right)/\sigma^2 = \sum_{n=1}^{N} x_{n-1}^2/\sigma^2. \quad (4)$$

Lai and Siegmund (1983) considered a sequentially observed AR(1) process and proposed to evaluate the least square estimator at the stopping time τ_{1c} de ned by

$$\tau_{1c} = \inf\left\{N > 1 : \sum_{n=1}^{N} x_{n-1}^2/\sigma^2 \geq c\right\}, \quad (5)$$

for some predetermined $c > 0$. Later, we de ne a feasible stopping time τ_{2c} in (10) by replacing σ^2 with its estimator in (5).

For the stopping time de ned in (5), we de ne the sequential least square estimate by setting $N = \tau_{1c}$ in (7). The asymptotic normality of $\hat{\beta}_{\tau_{1c}}$ have been shown by Lai and Siegmund (1983). One of their main results is as $c \to \infty$,

$$\sqrt{c}\left(\hat{\beta}_{\tau_{1c}} - \beta\right) \to N(0,1) \quad (6)$$

uniformly in $\beta \in [-1,1]$, which allows us to obtain the confidence intervals about β with fixed accuracy. [1]

Lai and Siegmund (1983) also showed

$$\tau_{1c}/c \to \sigma^2/\gamma(0) = 1 - \beta^2, \quad (7)$$

where $\gamma(\cdot)$ is the covariance function of $\{x_n\}$.

Suppose we sequentially observe $\{x_n\}$ from the stationary AR(1) model in (1). When the initial value x_0 possesses the stationary distribution, then $\{x_n\}$ has the covariance function

$$\gamma(m) = \beta^{|m|}\sigma^2/\left(1-\beta^2\right). \quad (8)$$

We assume that the initial value x_0 is a L^2 random variable and independent of $\epsilon_1, \epsilon_2, \cdots$. Let s^2_N be the estimator of σ^2;

$$s_N^2 = \sum_{n=1}^{N}\left(x_n - \hat{\beta}_N x_{n-1}\right)^2/N. \quad (9)$$

As well as τ_{1c} in (5), we set a feasible stopping time ;

$$\tau_{2c} = \inf\left\{N > 1 : \sum_{n=1}^{N} x_{n-1}^2/s_N^2 \geq c\right\}. \quad (10)$$

[1] To obtain the uniform asymptotic normality, Lai and Siegmund (1983) assumed that the initial value x_0 is not dependent on β and $\sup_{|\beta|\leq 1} P_\beta\left\{x_n^2 > a\right\} \to 0 \quad as \quad a \to \infty$ for each xed $n \geq 0$. Since these assumptions do not hold when $\{x_n\}$ is a strongly stationary process, we discard them and use the fruitful theory of the ergodic stationary processes. Then we obtain the asymptotic normality of the stopping times instead of the uniformity in the asymptotic normality of $\hat{\beta}_{\tau_{1c}}$.

Our purpose here is to study the asymptotic behavior of $\left(\hat{\beta}_{\tau_{1c}}, \tau_{1c}\right)$ and $\left(\hat{\beta}_{\tau_{2c}}, s^2_{\tau_{2c}}, \tau_{2c}\right)$.

The contributions of the present paper are as follows.

First, for the stationary AR(1), we prove the joint asymptotic normality of the sequential estimators for (β, σ^2) and the stopping times τ_{1c} in (5) and τ_{2c} in (10). Especially, we nd that τ_{1c} has the asymptotic variance strictly greater than τ_{2c}.

Second, we introduce the following new methodology in sequential analysis. We represent random quantities of concern in terms of stochastic processes in $D[0,\infty)$ and apply functional central limit theorems in $D[0,\infty)$ including Theorem 2 which is an extension of Theorem 19.1 in Billingsley(1999, p.197). Together with the limits of the stopping times, we derive the asymptotic properties of the sequential statistics. Skorohod representation theorem (Billingsley (1999, p.70)) makes it possible to evaluate the continuous-time stochastic processes represented by Brownian motions at the limits of the stopping times.

2. Methodology

According to Lai and Siegmund (1983), $\tau_{1c}/c \to 1 - \beta^2$ almost surely. We show the same result holds for τ_{2c}/c. [2]

Theorem 1. Let x_0 be an arbitrary L^2 random variables and independent of $\epsilon_1, \epsilon_2, \ldots$. Then, τ_{1c} defined in (5) and τ_{2c} in (10) satisfy:

$$\lim_{c\to\infty} \tau_{1c} = \lim_{c\to\infty} \tau_{2c} = \infty \ a.s. \tag{11}$$

$$\lim_{c\to\infty} \frac{\tau_{1c}}{c} = \lim_{c\to\infty} \frac{\tau_{2c}}{c} = \frac{\sigma^2}{\gamma(0)} = 1 - \beta^2 \quad a.s.. \tag{12}$$

The sequential estimates of β and σ^2 have strong consistency;

$$\lim_{c\to\infty} \hat{\beta}_{\tau_{1c}} = \lim_{c\to\infty} \hat{\beta}_{\tau_{2c}} = \beta \ a.s \ \text{ and } \ \lim_{c\to\infty} {s^2}_{\tau_{2c}} = \sigma^2 \ a.s \tag{13}$$

Theorem 1 gives the almost sure convergence of the stopping times τ_{1c} and τ_{2c}. Now we provide some asymptotics with respect to the sequential statistics. We consider applying the theory of convergence of random elements in $D[0,\infty)$; $D[0,\infty)$ is the set of the right continuous functions on $[0,\infty)$ with left limits. In a sequential sampling scheme, the space $D[0,\infty)$ is natural to characterize the limiting behavior of sequential statistics, since we consider stopping times with unbounded range of integers.

The following theorem and Skorohod's representation theorem (Billingsley(1999, p.70).) allows us to derive the joint asymptotic normality of

[2] $\lim_{c\to\infty} \tau_{1c}/c = \lim_{c\to\infty} \tau_{2c}/c = \sigma^2/\gamma(0)$ also holds for any stationary p-th order autoregressive process $(\mathrm{AR}(p))\ x_n = \beta_1 x_{n-1} + \cdots + \beta_p x_{n-p} + \epsilon_n$ with covariance function $\gamma(m)$.

the sequential estimates and the stopping times. Here, ||. || denotes the L^2 norm and $\lfloor a \rfloor$ the integer part of a for $a > 0$. Let Z- = {...,−2,−1,0}.

Theorem 2. Let $\Omega = \mathbb{R}^{\mathbb{Z}}$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^{\mathbb{Z}})$, and P be a probability measure on (Ω,F). For $\omega = (\omega_n) \in \Omega$, define the coordinate process $x_n(\omega) = \omega_n$ and assume that x_n is stationary and ergodic. Let $F_n = \sigma[x_k : k \le n]$ and $\xi_n = h(\cdots, x_{n-1}, x_n)$ be L^2-random variables with a common $h : R^{Z-} \to R$. If

$$\sum_{n=1}^{\infty} \| E[|\xi_n| \mathcal{F}_0] \| < \infty \quad , \tag{14}$$

then

$$E[\xi_n] = 0,$$

and the series

$$\nu^2 = E[\xi_0^2] + 2\sum_{n=1}^{\infty} E[\xi_0 \xi_n] \tag{15}$$

converges absolutely. When $\nu > 0$, and $S_n = \sum_{k=1}^{n} \xi_k$, then

$$S_{\lfloor ct \rfloor} / \nu\sqrt{c}c \Rightarrow W(c \to \infty)$$

in the sense of $D[0,\infty)$, where W is a Brownian Motion.

Remark 3. Note that Theorem 2 is a simple extension of Theorem 19.1 in Billingsley(1999, p.197). Unlike Billingsley's theorem, a process to which we apply the functional central limit theorem can be di erent from a process generating ltration in our theorem.

3. Result

The previous section's results lead to the following lemma.

Lemma 4. Suppose x_0 has the stationary distribution in (1). Let

$$\omega^2 = E\left[\left(\epsilon_1^2 - \sigma^2\right)^2\right] < \infty \tag{16}$$

and

$$\nu^2 = E\left[\left(x_0^2 - \gamma(0)\right)^2\right] + 2\sum_{n=1}^{\infty} E\left[\left(x_0^2 - \gamma(0)\right)\left(x_n^2 - \gamma(0)\right)\right] \tag{17}$$

As $c \uparrow \infty$, we have

$$\begin{pmatrix} \sum_{n=1}^{\lfloor ct \rfloor} x_{n-1}\epsilon_n / \sqrt{c} \\ \sum_{n=1}^{\lfloor ct \rfloor} \left(\epsilon_n^2 - \sigma^2\right) / \sqrt{c} \\ \sum_{n=1}^{\lfloor ct \rfloor} \left(x_{n-1}^2 - \gamma(0)\right) / \sqrt{c} \end{pmatrix} \Rightarrow \begin{pmatrix} \sqrt{\gamma(0)}\sigma W^{(1)}(t) \\ \omega W^{(2)}(t) \\ \nu W^{(3)}(t) \end{pmatrix} \tag{18}$$

in the sense of $D[0,\infty)$, where $W^{(1)}, W^{(2)}$ and $W^{(3)}$ are standard Brownian motions with correlation matrix

$$R = \begin{pmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}, \rho_{13} = \frac{2\beta\gamma(0)^{3/2}}{\sigma\nu}, \rho_{23} = \frac{\gamma(0)\omega}{\sigma^2\nu} \quad . \tag{19}$$

Using the above lemma, we obtain

Theorem 5. (Sequential Asymptotic Normality) Let x_0 be an arbitrary L^2 random variable and independent of $\epsilon_1, \epsilon_2, \cdots$ with $\omega^2 = E\left[\left(\epsilon_1^2 - \sigma^2\right)^2\right] < \infty$. As $c \to \infty$, in the sense of $D[0,\infty)$,

$$\begin{pmatrix} \sqrt{c}\left(\hat{\beta}_{\tau_{1c}} - \beta\right) \\ \sqrt{c}\left(\frac{\tau_{1c}}{c} - \frac{\sigma^2}{\gamma(0)}\right) \end{pmatrix} \Rightarrow \begin{pmatrix} W_1^{(1)} \\ -\frac{\sigma\nu}{\gamma(0)^{3/2}} W_1^{(3)} \end{pmatrix}, \tag{20}$$

and

$$\begin{pmatrix} \sqrt{c}\left(\hat{\beta}_{\tau_{2c}} - \beta\right) \\ \sqrt{c}\left(s^2_{\tau_{2c}} - \sigma^2\right) \\ \sqrt{c}\left(\frac{\tau_{2c}}{c} - \frac{\sigma^2}{\gamma(0)}\right) \end{pmatrix} \Rightarrow \begin{pmatrix} W_1^{(1)} \\ \frac{\sqrt{\gamma(0)}\omega}{\sigma} W_1^{(2)} \\ -\frac{\sigma\nu}{\gamma(0)^{3/2}} W_1^{(3)} + \frac{\omega}{\sigma\sqrt{\gamma(0)}} W_1^{(2)} \end{pmatrix}. \tag{21}$$

The following corollary shows the asymptotic variance of τ_{1c} is strictly greater than that of τ_{2c}.

Under the same assumption in Lemma 4. For the stopping time τ_{1c} and τ_{2c} de ned in (5) and (10), as $c \uparrow \infty$

$$\sqrt{c}\left(\frac{\tau_{1c}}{c} - \frac{\sigma^2}{\gamma(0)}\right) \Rightarrow N\left(0, \frac{\nu^2\sigma^2}{\gamma(0)^3}\right), \tag{22}$$

$$\sqrt{c}\left(\frac{\tau_{2c}}{c} - \frac{\sigma^2}{\gamma(0)}\right) \Rightarrow N\left(0, \frac{\nu^2\sigma^2}{\gamma(0)^3} - \frac{\omega^2}{\sigma^2\gamma(0)}\right). \tag{23}$$

Using the following proposition, one can obtain the long-run variance ν^2 in (17).

Proposition 6. For a strongly stationary AR(1) process $\{x_n\}$ de ned in (1) with the covariance function $\gamma(m)$ and $\mu_4 = E(\epsilon_n^4)$ is nite. Let $\gamma_{x^2}(m) = cov\left(x_0^2, x_m^2\right)$,

$$\gamma_{x^2}(m) = 2\gamma(m)^2 + \left(\mu_4 - 3\sigma^4\right)\beta^{2m}/\left(1 - \beta^4\right).$$

When $\mu_4 = 3\sigma^4$, then $\gamma_{x^2}(m) = 2\gamma(m)^2$, $\gamma_{x^2}(m) = \beta^{2m}\gamma_{x^2}(0)$ and the long-run variance in (17)

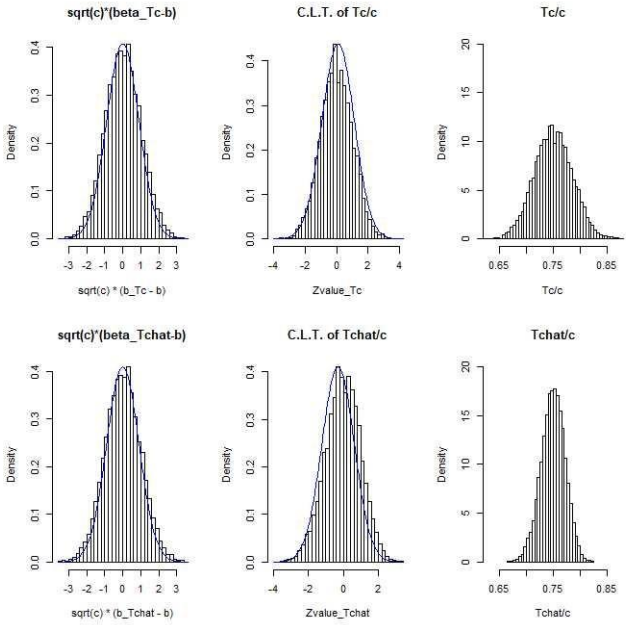
$$\nu^2 = \frac{2\left(1 + \beta^2\right)\sigma^4}{\left(1 - \beta^2\right)^3}. \tag{24}$$

4. Discussion and Conclusion

Now, we provide a simulation study to examine our main results in Lemma 4 and Theorem 5. The simulation setting is $\beta = 0.5$, $\epsilon_n \sim_{i.i.d.} N(0,1)$, $c = 2000$ and the number of replication is 10,000.

Figure 1 presents the simulated results: Tc stands for τ_{1c} and Tchat stands for τ_{2c}. The rst and second columns show the simulated histograms of $\hat{\beta}_{\tau_{1c}}$ $\hat{\beta}_{\tau_{2c}}$ and stopping times τ_{1c} τ_{2c} after normalization, and the blue curves are the density of standard normal distribution. We could see that both sequential estimators $\hat{\beta}_{\tau_{1c}}$, $\hat{\beta}_{\tau_{2c}}$ and stopping times τ_{1c}, τ_{2c} are well approximated by standard normal distribution. From the histograms in the third column, it's obvious that the variance of τ_{1c} is larger than τ_{2c}.

Figure 1: Histograms of sequential estimators and stopping times



We also examine the correlation $\rho_{12},\rho_{23},\rho_{13}$ of $W^{(1)},W^{(2)}$ and $W^{(3)}$ in (18). The simulated results are well approximated to the theoretical values.

Table 1: The simulation results of correlation $\rho_{12},\rho_{13},\rho_{23}$ in Lemma 4

	theoretical values	simulation results
ρ12	0	-0.0111
ρ13	0.6325	0.6328
ρ23	0.7746	0.7736

This paper investigate the joint asymptotic normality of stopping times and sequential estimators possessing fixed accuracy. The functional central limit theorem (Theorem 2) and Skorohod's representation theorem give the methodology to analyze the asymptotic properties of linear or nonlinear time series. Using that we nd the asymptotic distributions of the two stopping times τ_{1c} and τ_{2c} are normal, while the stopping time τ_{1c} with the true σ^2 has a larger variance than the stopping time τ_{2c} with the estimator of σ^2. A similar investigation can be extended to p-th order autoregressive process (AR(p)).

We present the ideas and the theoretical results here. The application should be also considered. For example, combing the results of companion paper (K.Nagai, Y. Nishiyama, and K. Hitomi (2018)), the sequential detection for the order d of Integrated AR(p) process is developed and examined by simulation. Based on our methodology, sequential analysis and statistical process monitoring should be developed for linear and nonlinear time series, such as autoregressive moving average model (ARMA), autoregressive conditional heteroscedasticity model (ARCH), generalized ARCH model (GARCH).

References

1. Billingsley, P. (1999) Convergence of Probability Measures, 2nd Ed., John-Wiley and Sons.
2. Galtchouk, L. and Konev, V. (2004). On uniform asymptotic normality of sequential least squares estimators for the parameters in a stable AR(p), J. Multivariate Anal. 91, p. 119-142.
3. K.Nagai, Y. Nishiyama, and K. Hitomi (2018). Sequential test for unit root in AR(1) model, Kyoto Institute of Economic Research Discussion Paper, No. 1003, Kyoto University
4. Lai, T.L. and D. Siegmund (1983). Fixed accuracy estimation of an autoregressive parameter, Annals of Statistics 11, 478-485.
5. Sriram, T.N. (1987). Sequential estimation of the mean of a rst order stationary autoregressive process. Ann.Statist.15, p. 1079-1090.
6. Sriram, T.N., I. V. Basawa, R. M. Huggins (1991). Sequential Estimation for Branching Processes with Immigration, Ann. Stat., Vol. 19, No. 4, pp. 2232-2243.

Measurement of multidimensional child poverty in Morocco 2001-2014: Methodology and results

Abdeljaouad EZZRARI, Khalid SOUDI
High Commission of Planning, Rabat, Morocco

Abstract

The objectives of this study are: (i) proposing a multidimensional measure of poverty approach to quantify the extent of this phenomenon; (ii) determining the dimensions and factors that contribute to its social reproduction; and (iii) determining the Child Poverty profile. Among other things, this mainly aims to respond to the following questions: Who are these poor children? Why are they poor? What are the correlates and reproductive factors in child poverty? is there a generational transmission of poverty? What are the individual, family and community determinants of child poverty? How has this poverty evolved? The methodological approach developed as part of this investigation was based on the theory of fuzzy sets and on MPI Alkire and Foster approach. By combining these two approaches, the final approach is: 1) determining the weights of the dimensions defining the space of the well-being of children; 2) standardizing dimensional indices defining the well-being of children; 3) calculating the composite index of deprivation according to the approach of fuzzy sets; 4) and calculating of the indices of multidimensional poverty according Alkire and Fooster approach.

The results of this study show a general improvement of social children welfare. The evolution of the composite index of deprivation highlights the continuing decline of the situation of children deprivation, of all ages: it dropped by nearly half, from 0.295 in 2001 to 0.128 in 2014. Along with this trend, multidimensional child poverty knew a strong downward trend. The prevalence of poor children evolved from 43.6% in 2001 to 24.1% in 2007 and 11.0% in 2014. From the outset, the share of severely poor children moved from 24.5% in 2001 to 9.7% in 2007 and 2.6% in 2014. It is in rural areas where this form of poverty is most striking: in 2014, it was 5.4%, while 0.3% in urban areas. These indices were respectively 45.1% and 3.3% in 2001.

Poverty experienced in childhood is a social reproduction of adult poverty and a consequence of poor living conditions. The risk of multidimensional child poverty is strongly differentiated by socio-professional category of the head of household. Similarly, education and knowledge are also proving essential determinants in improving children's standard of living.

1. Introduction

Improving the understanding of the Moroccan children's situation of the well-being, and the inherent issues and challenges, is a statistical framework necessary to any actions aimed at breaking the cycle of intergenerational transmission of vulnerability and poverty, strengthening the pro-poor quality of public policies, reducing the inequality of opportunities faced by children, and supporting poor households to raise up their children.

To do this, this study is framed around the following: (i) methodological and analytical framework of multidimensional child poverty: this axis presents the approach developed in this study to measure the multidimensional child poverty; (ii) Comparative Profile welfare and multidimensional child poverty.

2. Methodological Framework: proposing of a measurement approach to multidimensional child poverty

Attempts to conceptualize child[1] poverty are many and lead to a series of deprivation that prevent this segment of the population to enjoy their rights. Basically, different definitions opt for three well-being dimensions to define child poverty: (i) lack of survival means: in other terms, growing up without access to financial and nutritional resources necessary for survival and development; (ii) family and community structures' failure to protect children (social resources); and (iii) lack of opportunities to participate in political life (political resources).

3. Measurement methodology of multidimensional child poverty: combined approach of fuzzy sets and Alkire-Fooster

The measurement of multidimensional poverty of children is based on the combination of fuzzy set approach and Alkire and Foster approach. It consists of distributing individuals along a continuum of well-being (between 0 -a maximum well-being, and no deprivation- and 1 -a minimum welfare marked by absolute deprivation). The counting method of this measure goes through four stages, namely: 1) determining the weight of defining the dimensions of well-being of children; 2) standardization of variables defining the well-being of children; 3) calculation of the composite of deprivation index 4) calculation of the indices of multidimensional poverty according Alkire and Fooster approach.

[1] *The notion of the Child adopted by the present work is consistent with that of United Nations agencies and especially UNICEF. Thus we consider child any person whose age is strictly less than 18 years.*

Step 1:Weighting Scheme

Determining the weight of dimension is the main concern of the multidimensional poverty measurement. The choice of an appropriate weight is one of the fundamental steps in the calculation of poverty composite indices of. As part of this work, we use the method proposed by Cerioli and Zani, which evolves around the following relationship:

$$w_j = \ln\left[\frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} x_{ij} n_i}\right] \qquad (1) \qquad \text{avec} \qquad \sum_{i=1}^{n} x_{ij}\, n_i \; > 0$$

x_ij represents the score of the i-th individual in relation to the j-th dimension, and n_i is the weight of an individual or group of individuals. This means that the weighting of each dimension is weighted by the logarithm of the inverse of the frequency of non-full or partial fulfillment of this dimension (the dimension of deprivation score).

Step 2: Standardization of measurement variables: Determination of the the score function.

This function is defined as follows:

$$\varphi_{ij} = \begin{cases} 1 \text{ si } \varphi_{ij} = \varphi_j^{min} \\ \dfrac{\varphi_j^{max} - \varphi_{ij}}{\varphi_j^{max} - \varphi_j^{min}} \text{ si } \varphi_j^{min} \le \varphi_{ij} \le \varphi_j^{max} \\ 0 \text{ si } \varphi_{ij} = \varphi_j^{max} \end{cases} \qquad (2)$$

with Φ_ij the score of the i-th individual in relation to the variable j-th; ϕ_j ^ min ^ ϕ_j and max are the minimum and maximum values. Each score is associated with a value between 0 and 1, representing this variable in a given individual or household, the degree of deprivation.

For each dimension with more than one variable, a weighted score is calculated as follows:

$$\boldsymbol{S_{ij}} = \sum_{\boldsymbol{p=1}}^{\boldsymbol{n_k}} \boldsymbol{r_p \varphi_{ip}} \qquad (3)$$

With n_k number of variable dimension j, r_p is the relative weight assigned to the variable p with $r_p \ge 0$ and $\sum_{p=1}^{n_k} r_p = 1$, and φ_{ip} is the membership function of household i for the variable p.

The r_p weight is obtained from equation (1) by replacing j with p.

Step 3: Calculate the composite deprivation index (CDI)

After calculating the weight assigned to each attribute (variable) or dimension, the last step is the determination of composite indices (fuzzy) deprivation. To do this, we must first calculate the deprivation composite index of each individual or household a_i through the following relationship:

$$\mu_B(a_i) = \frac{\sum_{j=1}^{m} x_{ij} w_j}{\sum_{j=1}^{m} w_j}, \qquad 0 \le \mu_B(a_i) \le 1 \qquad (5)$$

Blur poverty index of the subset P is determined from the relationship:

$$\mu_B = \frac{\sum_{i=1}^{n} \mu_B(a_i) n_i}{\sum_{i=1}^{n} n_i} \qquad (6)$$

The multidimensional index for each dimension or variable is determined as follows:

$$\mu_B(X_j) = \frac{\sum_{i=1}^{n} x_{ij} n_i}{\sum_{i=1}^{n} n_i} \qquad (7)$$

Step 4: Indices of multidimensional poverty according to the Alkire and Fooster method

By construction, ICP summarizes all the hardships experienced by children. By classifying them according to the degree of deprivation, this index allows to assess all the indices of multidimensional poverty, including the indices which are the most recognized by Alkire-Fooster, namely:

- Headcount ratio of Multidimensional poverty (H): it gives the proportion of poor children, that is to say, children who accumulate a number of deprivations greater to the poverty line -at least 30% of dimensions deprivation of well-being;
- The deprivation intensity (A): This index provides information on the gaps faced by poor children in a simultaneous manner. It has the merit to account for the deprivation acuity in children in multidimensional poverty situation;
- The Multidimensional Poverty Index (MPI): It's a generalization of the intensity of deprivation among all children, whether in poverty or not;
- The index of severity: it gives the proportion of children in situations of deprivation which at least 50% of the spatial dimensions of well-being of the child. It provides information on the share of the poorest children;
- The index of vulnerability to poverty: it gives the share of children whose level of deprivation oscillates in a range between 20% and 30% of well-being dimensions. It provides information on the non-poor children's risk of falling into poverty.

Throughout this work, we will calculate, analyze and compare over time all of these indices (CDI, H, A and MPI).

4. The key results from this study are available in:

1- Composite index of children's social well-being

Between 2001 and 2014, the situation of children in Morocco has experienced a significant improvement in all areas. It is characterized by a tendency to generalize the education of young children aged 6-11 years in primary school, by improving net enrollment rates in other levels: preschool, secondary college and qualifying secondary by a narrowing of the urban / rural and boy / girl disparities in children's access to education, by improving their

health and nutrition, and housing conditions which are more decent offering more comfort and security, etc.

This development has resulted in an overall improvement of the children's socioeconomic welfare. The analysis of the deprivation composite index highlights the continuing decline in the situation of children deprivation in all ages, over time. Thus, this index decreased by almost half, from 0.295 in 2001 to 0.128 in 2014, an average annual decrease of about 7.0% during this period. This improvement has concerned the two residence areas: between those dates, the average level of deprivation decreased from 0.115 to 0.052 in urban areas and 0.47 to 0.221 in rural areas.

The spatial distance (urban / rural) of the average level of deprivation shows that deprivation remains a rural phenomenon.

2- Multidimensional Indices of Poverty

With emphasis on the fringe of children subject to deprivation in at least 30% of the size of the welfare area, it appears that the multidimensional child poverty recorded strong downward trend. The prevalence of poor children increased from 43.6% in 2001 to 24.1% in 2007 and 11.0% in 2014. With these rates, the number of children in poverty decreased from 4.9 million children in 2001 to 1.24 million children in 2014, an average annual reduction of 10.0% of the total number of poor children.

By controlling the area of residence, the prevalence of multidimensional poverty decreased from 11.8% in 2001 to 6.1% in 2007 and 2.4% in 2014 in urban areas. These indices are respectively 74.6%, 46.9% and 22.0% in rural areas. It shows that child poverty remains predominantly a rural phenomenon. The difference between these two indices reflects the high concentration of child poverty in rural areas.

With regard to children vulnerability to impoverishment, the risk of falling into poverty reached 23.4% in the countryside, against 5.6% in the urban areas. However, it is clear that if the risk of being vulnerable to multidimensional poverty declined in urban areas, 7.2% in 2001 versus 5.6% in 2014, it has, however, increased in rural areas from 12.5% to 23.4% between 2001 and 2014.

The rate of children escape from poverty is at different speeds across regions of the country. Between 2001 and 2014, the largest decline was recorded in the regions where the phenomenon is more widespread: it fell from 59.8% to 16.5% in "Marrakech-Safi", 46.4% to 10.0% in "Tanger-Tetouan-Al Hoceima", from 44.6% to 14.6% in "Béni Mellal-Khénifra".

The examination of the correlates between the conditions of children and the conditions of their family homes shows that multidimensional poverty affects more 5 to 6 years old children, a poverty rate of 21.0%, while the 7-14 years old children are the least affected by this form of poverty (7.3%). Despite

the narrowness of the poverty rate of 7 to 14 years old, the poor children within this age group contribute about 28.3% of the total child poverty.

Another approach to the issue of children in poverty is to focus on the demographic and economic conditions of their households. Poverty experienced in childhood is a social reproduction of adult poverty and a consequence of poor living conditions.

Thus the number of children in the household notoriously makes a difference on the children well-being. In 2014, the poverty rate has more than quadrupling depending on whether the household has one child (6.5%) or 6 children and more (28.0%).

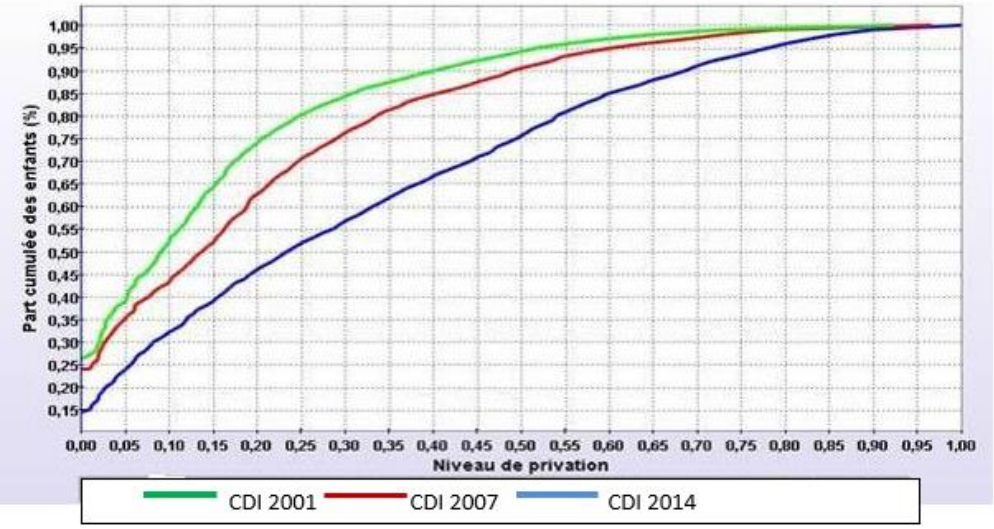
5. Intergenerational transmission of poverty: the socio-economic situation of the household and especially parents have a strong impact the child's destiny.

The head of household sex differently impacts the situation of children with regard to poverty. In 2014, the poverty rate is 11.2% among children whose head of household is a man, against 8.6% in Guiders' children. Children headed by women have a 3.1% chance to be among children little or no at all deprived than their counterparts living in households headed by a man.

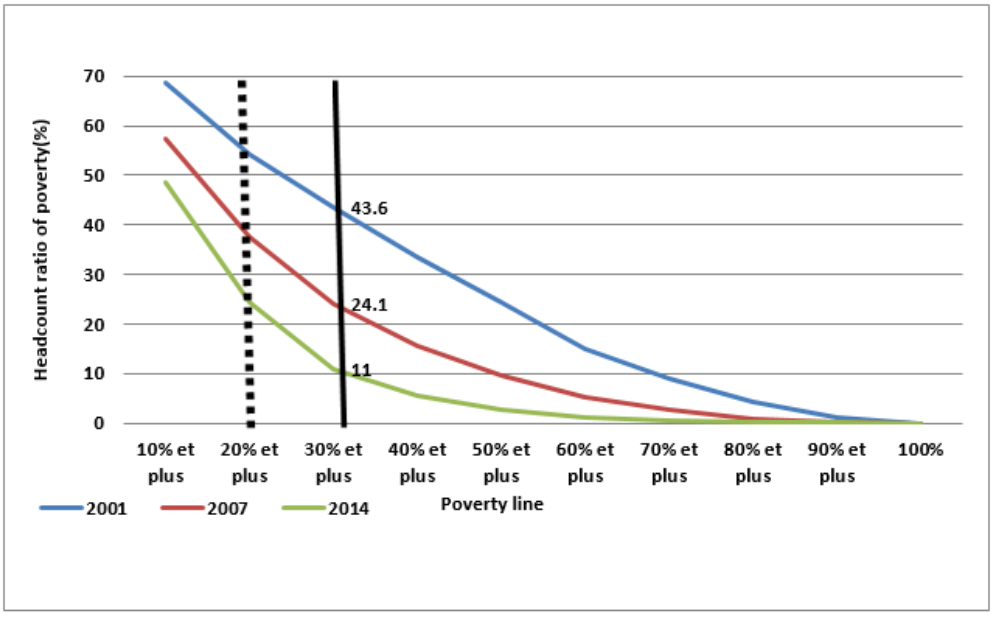
Education and knowledge are also proving essential determinants in improving children's living standard. The level of household head education has a central role in determining the level of child poverty: the incidence of child poverty is 0.5% for children of households headed by a leader with a higher education, against 16.4% of households headed by children without school level.

The risk of multidimensional child poverty is strongly differentiated by the head of household's socio-professional category of. In 2014, the incidence of child poverty is more marked among children in households headed by "farmers", a 25.4%poverty rate, "farm workers and laborers and fishing" (24.3%), "small trades" (11.3%) and "craftsmen and skilled workers" (8.9%). The risk for a child to fall in poverty or vulnerable to poverty is 19.1% higher in children "farm workers and laborers" than in children "managers and senior managers".

Graph 1: Distribution curve of CDI



Graph 2 : Multidimensional poverty incidence curves



Measuring the statistical capacity of nations[1]

Grant J. Cameron, Hai-Anh H. Dang, Mustafa Dinc, James Foster, Michael M. Lokshin
World Bank

Abstract
The international development community has used the World Bank's Statistical Capacity Index since its inception in 2004. The Sustainable Development Goals create new challenges for national statistical systems to produce high-quality and internationally comparable data. We review measurement methodologies, posits desired attributes, and presents theoretical and empirical frameworks for the new, improved index to monitor progress in the statistical capacity of nations. We illustrate the properties of the updated index with global data from 2016.

Keywords
statistical capacity; statistical indicators; statistical index; national statistical system; data

1. Introduction

The national statistical system, or NSS, plays a crucial role in modern economies. It provides stakeholders, ranging from policy makers to stock market analysts and the general public, with the latest data on the country's socio-economic developments. At the international level, monitoring progress on global undertakings such as the recently established Sustainable Development Goals (SDGs) requires high-quality data that must be produced consistently across different national statistical systems. Assessing and improving the capacity of a country's NSS has long been a part of the global agenda. The international Partnership in Statistics for Development in the 21st Century (PARIS21) Task Team was established in 2002 to help measure country statistical capacity. Over the subsequent years, a few capacity assessment tools have been developed to identify the weaknesses and strengths of national statistical systems[2].

[1] This is a shortened version of a paper of the same title, Cameron et al. (2019) which is World Bank Policy Research Working Paper no. 8693.

[2] Statistical capacity is usually interpreted as the ability of an NSS to meet user needs for relevant and good quality statistics in a timely manner. An NSS often consists of a number of different data-producing agencies and departments (such as the national statistical office, the central bank, and statistical departments within other line ministries), which renders the task of directly measuring statistical capacity a difficult one.

The World Bank's Statistical Capacity Index (SCI) is one such tool that has been widely employed.[3]

Several international and national agencies have adopted the SCI for measuring progress in statistical capacity building and related investments. The United Nations, for example, uses the SCI to measure trends in the development of national statistical capacity (United Nations, 2016). The SCI is used to evaluate the efficiency of statistical support provided to a country as well as the need to further develop its statistical capacity (PARIS21, 2002). Some regional organizations use the SCI to identify areas of improvement in their member countries (OIC, 2012), while researchers use the SCI as a benchmark to validate their new statistical indexes (Sanga et al., 2011). The World Bank mainstreamed the SCI in its monitoring and assessment framework and has adopted it as a baseline indicator in various projects at the country level[4]. The SCI is based on publicly available data, and this has various advantages over other indexes of statistical capacity. A key advantage of the SCI is that it can provide assessment of a country's statistical capacity in an internationally comparable and cost-effective manner.

Existing efforts in building indexes to assess statistical capacity have focused on the practical aspects such as data collection, organization, and legal issues, paying little attention to the underlying theoretical principles that are indispensable for the construction of a reliable, transparent, and consistent statistical capacity index. For example, the UNECE, in a recent Global Assessment report, discusses only the legal basis, description of the statistical system, data source, and processing of the target country (UNECE, 2014). The FAO, in its guidelines for assessing country capacity in producing agricultural statistics, provides instructions on completing the questionnaires and on compiling the assessment indicator (FAO, 2014), but pays no attention to the axiomatic principles of these indicators. The U.S. Census Bureau developed and recently updated (2017) the Tool for Assessing Statistical Capacity (TASC) with a primary objective of measuring the overall capacity of an NSS by providing a breakdown of the areas of strength and weakness. However, the focus of this instrument is on measuring the capacity of an NSS to conduct household-based surveys and censuses[5]. To our knowledge, only Sanga, Dosso, and Gui-Diby (2011) discuss the technical framework behind the African Statistical Development Index (ASDI).

[3] For brevity, we refer to both the Statistical Capacity Indicators and the Statistical Capacity Index as the SCI in the rest of the paper. We will make it clear where we refer to either the indicators or the index. We similarly refer

[4] For other recent examples that use the SCI, see: Beegle et al. (2016) for an analysis of the relationship between good governance and statistical capacity in African countries; Tapsoba, Noumon, and York (2017) for the impacts of statistical capacity on reducing procyclical fiscal policy; and UNICEF (2018) for the role of statistical capacity in tracking the SDG for child development.

[5] We return to provide more discussion on the SCI and these other methods in Section 2 below.

In this paper, we aim at laying out the conceptual foundation behind statistical capacity indexes, and construct a new index based on practical and theoretical considerations. We review existing measurement methodologies, posit desired attributes, and propose updated indicators, and an updated Statistical Capacity Index (hereafter referred to as the Statistical Performance Index, or SPI). On the empirical front, we expand the number of indicators in the old SCI by almost twice, and we extend the sample of covered countries by one-half to all countries in the world.

2. Methodology

In order to construct a measure that is policy relevant it is helpful to follow a series of basic steps.[6]

The first step asks the question: what phenomenon is being measured? A clear conception helps orient the process by which the measure is assembled and will prove valuable in communicating its underlying meaning.

The second step asks: for what purpose or purposes is the index being sought? Knowing how the index will be used can greatly affect subsequent choices in its construction, and its eventual suitability. In particular, it will help define the unit of analysis both for data gathering and reporting purposes.

The third step identifies a list of essential characteristics, or desiderata, that the methodology should exhibit. This list of "pre-axioms" helps orient the construction process and define what success means.

A fourth step identifies the conceptual space in which measurement is to take place. If there are multiple conceptual dimensions, consideration must also be given to the relative importance of each.

The fifth step selects the form of the variables to be used and the aggregation method to be employed – how the variables are to be combined into an overall measure.

The sixth step identifies a set of axioms that the resulting index should satisfy to have the greatest practical utility. Axioms are not sterile mathematical requirements, but rather contain the salient nuggets of policy required of the index: which aspects of the data should be ignored, which should be reflected, and helpful consistency requirements over subsets of data. Together, these six steps comprise the core theoretical elements of our proposed measurement technology.

We briefly summarize the main ideas of the sixth step in our proposed methodology below. Interested readers are referred to the full version of the paper for more discussion and further technical details on the other steps; the equation and section numbers below refer to those in the full paper version.

[6] This process is similar for many types of measurement exercises. See for example Alkire et al. (2015) in the context of multidimensional poverty measurement.

Axioms are rigorous properties for an index to satisfy, and they are more formalized and generalized than the properties discussed earlier in Subsection 3.2.5. Axioms help in understanding what an index is actually measuring and in deciding which index to use. Knowing which properties an index satisfies can help in interpreting the empirical results obtained using that index; certain forms of policy analysis become possible only when the index satisfies a given property. Some axioms can be interpreted as "nuggets of policy" that specify the kinds of changes that should leave the index value unchanged and those that should alter it. Others break down the index value to help understand how dimensions contribute to that value. Our proposed index satisfies three axioms, which include symmetry, monotonicity, and subgroup decomposability.

As noted in Foster et al (2013), axioms can be usefully grouped into three categories: invariance axioms, which indicate what not to measure; dominance axioms, which indicate what the index should measure; and subgroup axioms, which break down or build up indices by variables or units of analysis. The three axioms our proposed index of statistical capacity satisfies are closely related with these three groups of axioms. In what follows, a generic index of statistical capacity over profiles $a = (a_1, \ldots, a_V)$ will be denoted by F.

Symmetry: in other measurement environments where the number of people, dimensions, or other factors may differ across comparisons, invariance axioms are often used to ensure consistency. In the present context, the index F is being applied to one country's data with a fixed number of dimensions and dichotomous variables, so properties of this sort are not needed. A second common form of invariance axiom is anonymity or symmetry whereby the index value is unaffected when variable levels are switched. In the present context, where the variables have a structure as represented by hierarchical tree T and partition P, universal symmetry is not appropriate. Motivated by Basu and Foster (1998), one might consider a weaker form of symmetry that is contingent on variables being "similarly placed" in the variable structure. We say that profile b is obtained from profile a by a *basic switch* if $b_v = a_{v'}$ and $b_{v'} = a_v$ for some $v \neq v'$ in the same basic group, while $b_{v''} = a_{v''}$ for all other v''. In other words, the only difference between b and a is that two variable values in the same basic group have been switched. A statistical capacity measure F satisfies *basic symmetry* if $F(a) = F(b)$ whenever b is obtained from a by a basic switch. Notice that any nested counting index C satisfies basic symmetry because it has the same weight on every variable in the same basic group.

Monotonicity: the main axiom for F is an intuitive dominance axiom requiring the index value to reflect improvements in variables. We say that profile b is obtained from profile a by an *improvement* if $a_v \geq b_v$ for all v, and $a \neq b$ or, in other words, if profile a vector dominates profile b. A statistical capacity index F satisfies *monotonicity* if $F(a) > F(b)$ whenever a is obtained

from b by an improvement. This simple but significant requirement ensures that the index value rises whenever one variable rises from 0 to 1 and the rest of the variables do not fall in value. The index $C(a; w)$ satisfies this property since each w_v is strictly positive. Notice that monotonicity supports the incentive compatibility criterion, since it ensures that a country is not penalized when it successfully raises its profile.

*Subgroup decomposability: s*ubgroup axioms allow the index to be divided into salient sub-indices and linked back to the original index for policy analysis. In the present case, the main decomposition is over the basic groups given in partition $P = (p_1, \ldots, p_K)$. A statistical capacity index *F* satisfies *basic decomposability* if there exist weights $\rho_k \geq 0$ summing to 1 and sub-indices $c_k(p_k)$ such that

$$C(a) = \sum_{k=1}^{K} \rho_k \, C_k(p_k) \qquad (10)$$

In other words, there is a collection of indices, one for each basic group of variables, such that *C* can be expressed as a weighted average of these basic indices. This is clearly the case for the nested counting index $C(a; w)$, as it is based on a weighted mean. Likewise, Equation (5) (after Proposition 1) follows from Equation (10) by aggregating across basic groups within each dimension, so that the overall index value is just the average of the dimensional index values. These decompositions can help inform why one country is doing better than another or help describe how a single country is progressing over time.

As noted above, the single country index $C(a; w)$ can be expanded into an index $C(A; w)$ that covers all countries in a region or even the universe of covered countries. The formula used to do this – Equation (7) – doubles as another form of decomposition that expresses the aggregate index and an average of the country indices. Since $C(a; w)$ is the index of primary interest here, the equation will not be expressed as a formal property here. However, the fact that Equation (7) and $C(A; w)$ are available allows users to have a better understanding of regional levels and trends in statistical capacity.

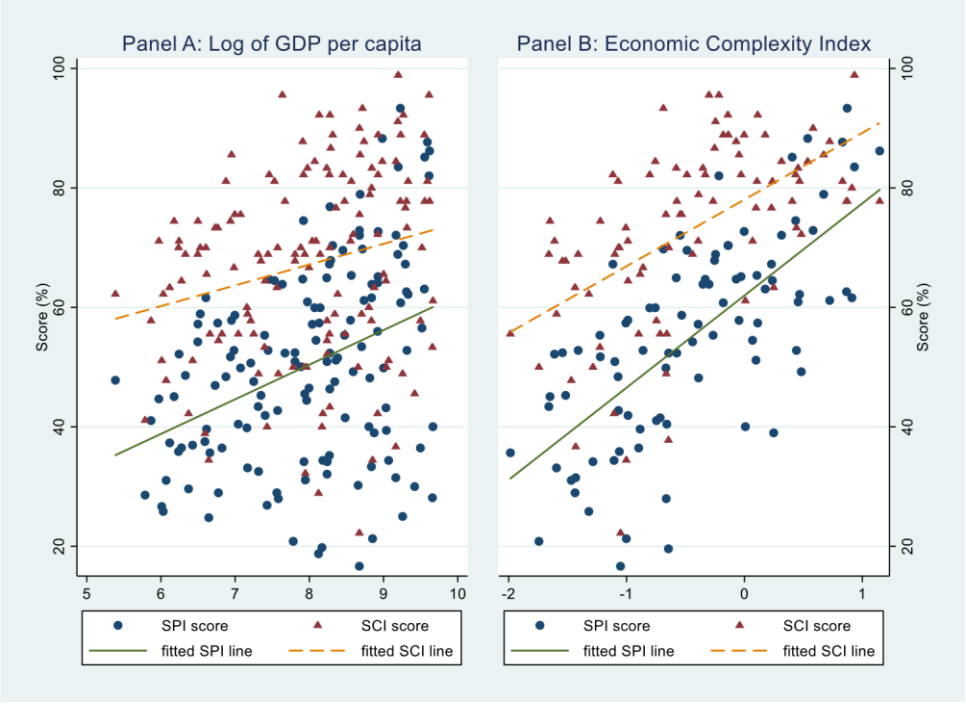
3. Result

The SPI has several advantages over the SCI, particularly in terms of data coverage. In particular, it has

i. Richer and more comprehensive dimensions covering different data aspects ranging from data generation, curation, and dissemination to data analysis.
ii. More indicators: the SPI has 42 indicators (of which 39 are used for scoring), versus 25 indicators in the SCIs.
iii. More countries: the SPI covers more than 200 countries, especially including high-income countries, while the SCI covers fewer than 150 countries and includes no high-income countries.

More importantly, the SPI is built on the conceptual and theoretical framework laid out above, while the theoretical principles of the SCI are not clearly formulated. For comparison, we plot in Figure 1 the SPI and SCI scores against log of GDP per capita (Panel A) and economic complexity index (Panel B) for the 146 countries covered by both indexes. The slopes of the fitted lines for the SPI scores are clearly steeper than those for the SCI scores in both these graphs.

Figure 1: SPI and SCI Scores versus Country Income and Country Economic Complexity, 2016



We also run regressions of the SPI and SCI scores on these country characteristics. Estimation results suggest that richer countries, or countries with a more complex economy tend to have higher SPI scores. Indeed, for this sample of countries, a 10 percent increase in a country's GDP per capita is associated with approximately a 0.6 percentage point increase in its SPI score, twice the corresponding increase for the SCI. Similarly, a 0.1 increase in a country's ECI being associated with a 1.5 percentage point increase in its SPI score but only a 1.1 percentage point increase in its SCI score. (Notably, these correlations are stronger for the larger sample of more than 200 countries that the SPI covers).

Another strength of the SPI is that it is explicitly built around four main dimensions: i) Methodology, Standards and Classifications (MSC), ii) Censuses and Surveys (CS), iii) Availability of Key Indicators (AKI), and iv) Dissemination Practices and Openness (DPO). (We further discuss each dimension, its indicators, and other practical implementation details in the paper). But suffice

it now to examine a disaggregation of the scores by regions in all four dimensions, which can help us disentangle which dimensions drive these differences and can be improved.

Figure 2: Dimension SPI Scores by Region, 2016

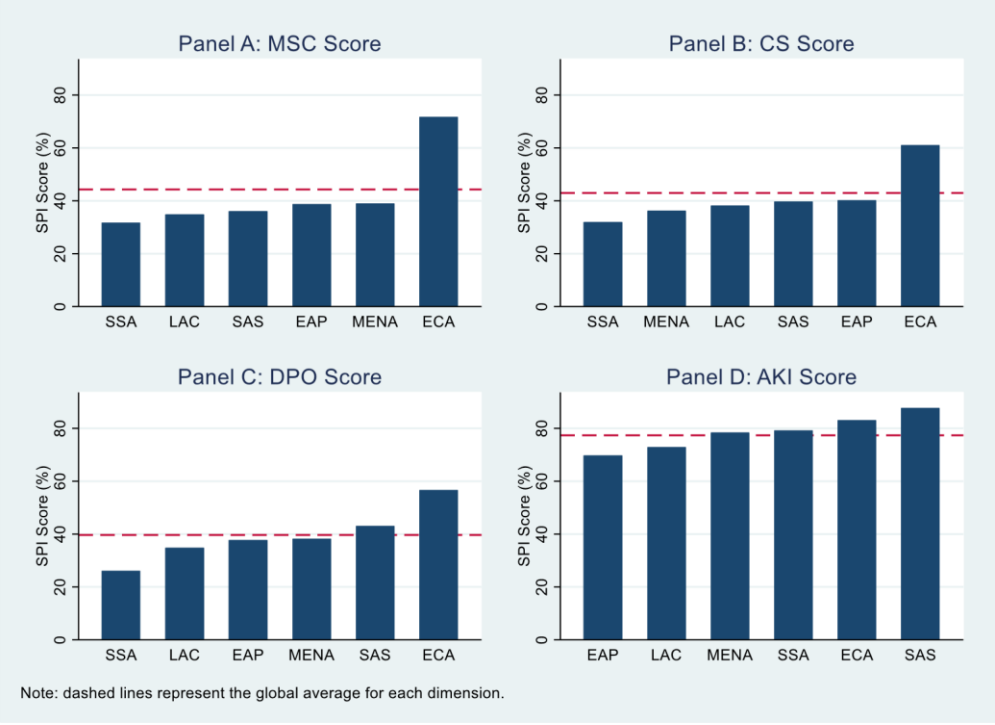


Figure 2 offers several interesting observations. First, a country's score on a certain dimension can be quite different from that of its overall SPI score. In particular, all dimensions—except for the DPO dimension—have a ranking order for country performance that is different from that with the overall score. An interesting example is the ECA region, which is consistently the best performer on all dimensions but AKI, where the SAS region is now the best performer. On the other hand, SSA performs slightly better than the global average on the AKI dimension, but is the weakest performer on the rest. Second, countries generally perform best in the AKI dimension, where their average dimension score is 77 percent, which is almost the corresponding figure for the other dimensions.

4. Discussion and Conclusion

This new index could be seen as the first step before more resource-intensive country-specific assessments to inform multi-year improvement plans. Our proposed framework is also flexible enough to allow for future revisions as the global data landscape evolves. For example, we can incorporate new indicators such as whether an NSO uses cloud computing to store their data or implements household panel surveys in the relevant dimensions without creating major changes to the total scores. Our

framework may also be relevant to the construction of other indexes in related areas, such as tracking the global SDGs or child development. Since the SPI is currently available only for 2016, another promising direction is to collect time series data for the new index, both going forward and for several years past, and expand the current framework to allow for dynamic changes over time.

References

1. Alkire, Sabina and James Foster. (2011). "Counting and multidimensional poverty measurement." *Journal of Public Economics,* 95(7): 476-487.
2. Alkire, Sabina, José Manuel Roche, Paola Ballon, James Foster, Maria Emma Santos, and Suman Seth. (2015). *Multidimensional Poverty Measurement and Analysis*. United Kingdom: Oxford University Press.
3. Basu, K., & Foster, J. E. (1998). On Measuring Literacy. *Economic Journal*, 1733-1749.
4. Beegle, K., Christiaensen, L., Dabalen, A., & Gaddis, I. (2016). *Poverty in a Rising Africa*. Washington DC: The World Bank.
5. Grant Cameron, Hai-Anh Dang, Mustafa Dinc, James Foster, and Michael Lokshin. "Measuring the Statistical Capacity of Nations". *World Bank Policy Research Paper # 8693*.
6. Food and Agriculture Organization of the United Nations (2014). *Guidelines for Assessing Country Capacity to Produce Agricultural and Rural Statistics*. June. http://gsars.org/wpcontent/uploads/2014/09/Guidelines_Country-Assessment_FINAL.pdf .
7. Foster, James, Suman Seth, Michael Lokshin, and Zurab Sajaia. (2013). *A Unified Approach to Measuring Poverty and Inequality: Theory and Practice*. Washington D.C.: World Bank.
8. Hidalgo, César A. and Ricardo Hausmann. (2009). "The Building Blocks of Economic Complexity." *Proceedings of the National Academy of Sciences,* 106(26): 10570-10575.
9. Organisation of Islamic Cooperation (2012). Current State of Statistical Capacity in the OIC Member Countries. February.
10. Partnership in Statistics for Development in the 21st Century (PARIS21) (2002). *Statistical Capacity Building Indicators Final Report*. https://www.paris21.org/sites/default/files/scbi-final-en.pdf.
11. Sanga, D., Dosso, B., and Gui-Diby, S. (2011). "Tracking Progress towards Statistical Capacity Building Efforts: The African Statistical Development Index". *International Statistical Review*, 79(3): 303-329.
12. Tapsoba, Sampawende J-A., Codjo Neree Noumon, and Robert C. York. (2017). "Can Statistical Capacity Building Help Reduce Procyclical Fiscal Policy?" *Journal of International Development,* 29(4): 407-430.

13. UNICEF. (2018). *Progress for Every Child in the SDG Era*. New York: UNICEF.
14. United Nations (2016). *Evaluation of the Contribution of the United Nations Development System to Strengthening National Capacities for Statistical Analysis and Data Collection to Support the Achievement of the Millennium Development Goals (MDGs) and other Internationally-agreed Development Goals.* Technical report.
15. United Nations Economic Commission for Europe. (UNECE). (2014). Global Assessment Report: National Statistical System of Mongolia. https://www.unece.org/fileadmin/DAM/stats/documents/technical_coop/GA_Mongolia_EN.pdf.
16. United States Census Bureau. Tool for Assessing Statistical Capacity (TASC)—Version 2.0. https://www2.census.gov/software/tasc/tasc-booklet.pdf
17. World Bank. (2018a). *Country Income Classification*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-andlending-groups
(2018b). *World Development Indicators Online*.

A comparison of ordinal regression in an analysis of factors associated with family well-being

Noor Azlin Muhammad Sapri[1], Kamarulzaman Ibrahim[2]
[1]National Population and Family Planning Development Board of Malaysia (NPFPDBM), Kuala Lumpur, Malaysia
[1,2]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia

Abstract
The aim of this study is to determine the factors which are significantly related to the level of family well-being by using ordinal regression model. The data for this study is obtained from the Family Well-Being Index Survey for the year 2011, conducted by the National Population and Family Development Board of Malaysia. It involved a stratified random sampling of 2,808 households and contains information on level of family well-being, demographic, socioeconomic and social characteristics of each household. The ordinal regression model involving four different link functions which include logit, probit, clog-log and nlog-log are fitted to the data for determining factors which have a significant relationship with the level of family well-being. Based on Akaike Information Criteria (AIC), it is found that ordinal regression model with logit link function provides the best fit for the data. The factors such as ethnic, family relationship, economy situation, health status, safety, religion practice, community relationship and housing and environment are significantly related to family well-being.

Keywords
Ordinal Regression Model, Ordinal Data, Family Well-Being, Malaysia.

1. Introduction

Ordinal regression models are playing a central role in studies where the researcher needs to quantify subjective variables. In the case of social sciences research, ordinal regression models are now becoming a very powerful tool in analysing data from questionnaires (Javali & Pandit 2010; Mouriño 2013). The main idea is to model an ordinal response variable as a function of some explanatory variables. In its basic formulation, the ordinal regression model can be viewed as an extension of the logistic regression model for binary response variables.

In Malaysia, little is known about the well-being of families since there is not much research done in this topic especially at the national level. Malaysian Quality of Life Index and Malaysian Youth Index are the two examples of local study. But both of it did not stress on the family well-being. However, in 2011,

there was a Family Well-Being Index Survey that was carried out by National Population and Family Development Board of Malaysia (Noor et al. 2014). The purpose of the study is to develop a set of indicators for measuring the well-being of families among Malaysian and to produce a composite Index of Family Well-Being. Confirmatory Factor Analysis (CFA) is used to identify the significant domains of family well-being which are family relationships, economic situation, health status and safety, community relationship and religion/spirituality.

There are several different tools for measuring level of family well-being depending on the different types of data collected and the scale of measurement adopted for capturing perception of respondent. Regression methods such as linear, logistic and ordinal regression are example of useful tools to analyze the relationship between multiple explanatory variables and level of family well-being (Fagerland & Hosmer 2016). In this study, the ordinal regression method is used to model the relationship between the ordinal outcome variable which is overall level of family well-being with several demographic and social characteristics variables.

This article is organized as follows. Section 2 explains the ordinal regression model which has been further discussed by Agresti (2007, 2011) and Liu & Agresti (2005). Section 3 describes the application of ordinal regression model on the family well-being data, while the results found are reported in Section 4. Finally, the conclusion is discussed in Section 5.

2. Ordinal Regression Model

The ordinal regression model is a generalisation of the logistic regression model, where the dependent variable is ordinal. There are several types of ordinal regression models which can be described based on the specific scenario of the data (Mccullagh 1980; Mccullagh et al. 2014; Mckelvey & Zavoina 1975). However, the aim of this study is to model the dependence of an ordinal response on discrete or continuous explanatory variables. The proportional odds model which is considered to summarize the relationship between the ordinal response and the explanatory variables, is as given in Equation (1).

$$g_j(x) = log\left[\frac{\Pr(Y \leq j|\mathrm{X})}{\Pr(Y > j|\mathrm{X})}\right] = \alpha_j - \beta'\mathrm{X}, \quad j = 1,2,\dots,c-1 \qquad (1)$$

Following the ordinal model above, let Y denotes an ordinal response variable with c levels (1,..,c) which in this case is the level of family well-being and x=(x_1,x_2,...,x_p)' be the vector of p explanatory variables. The higher value of ordered response category for family well-being indicates the high level of satisfaction of family well-being. For this study, x_i consists of demographic, socioeconomic and social characteristics of the selected households. The relationship between the response variable and the

explanatory variables is described through c-1 logits denoted as g_1 (X),g_2 (X),…,g_(c-1) (X), which relate a set of intercepts (α_s) and regression coefficients (β_s) to the probability of the response levels. The [exp (β_s)] is interpreted as the odds ratio for a one-unit increase in the given explanatory variable, which is a measure used for comparing two response levels or two sets of response levels.

In particular, there is no clear-cut procedure on how to choose the appropriate link-function to analyse the ordinal regression model (Christensen 2015). Accordingly, the main objective of this study is to employ the ordinal regression model with different link functions for modelling the relationship between level of family well-being and demographic, socioeconomic and social factors. The four link functions that are considered for this analysis, are as given in Table 1.

TABLE 1. Summary of various link functions

Function	Distribution	Link Function
Logit	Logistic	Log $(\frac{x}{1-x})$
Probit	Normal	$G^{-1}(x)$
Complementary log-log	Gumble (min)	Log$[-\text{Log}\,(1-x)]$
Negative log-log	Gumbel (max)	-Log$[-\text{Log}\,(x)]$

3. Application

The data for this study is obtained from the Family Well-Being Index Survey, which had been carried out by the National Population and Family Development Board of Malaysia in 2011. It consists of information on 2,808 households which contains some details on family well-being, demographic, socioeconomic, and social characteristics. The selection criteria of households were parents with child age at least 13 years old. The study involved a stratified random sampling design where the samples were selected throughout Malaysia according to strata (urban and rural) and three ethnic groups in Malaysia (Bumiputra, Chinese and Indian). For this study, the sample of either father or mother interviewed in each household involves the fraction of 1,484 fathers (52.8%) and 1,324 mothers (47.2%).

The response variable (Y) for the ordinal regression model under study is expressed as the Level of Satisfaction of Family Well-Being. Respondents were asked on overall satisfaction level of their family well-being. The question was measured on a 3-point scale (low, moderate, and high). There are two domains of explanatory variables for the study that is socio demographic and social characteristics. The socio demographic variables consist of Strata (urban/rural), Ethnicity (Bumiputra, Chinese and Indian), Type of Family (Nuclear, Blended, Single and Extended), Level of Education (Primary, Secondary, Tertiary and No

Formal Education) and Household Income. While, social characteristics contain overall satisfaction level on Family Relationship, Economic Situation, Health Status, Safety, Community Relationship, Religion Practice, and Housing and Environment. Each explanatory variable for social characteristic was measured using a 3-point scale (low, moderate, and high). The list of variables and associated classes are shown in Table 2.

TABLE 2. List of dependent and independent variables and the associated classes for the study

Variable	Class	Variable	Class
Family Well-Being Satisfaction (Y)	Low Moderate High	Economic situation (x_7)	Low Moderate High
Strata (x_1)	Urban Rural	Health (x_8)	Low Moderate High
Ethnic (x_2)	Bumiputra Chinese Indian	Safety (x_9)	Low Moderate High
Household Type (x_3)	Nuclear Extended Single Blended	Community Relationship (x_{10})	Low Moderate High
Educational Level (x_4)	Tertiary Secondary Primary No Formal Education	Religion Practice (x_{11})	Low Moderate High
Household Income (x_5)	<RM2000 RM2001-RM4000 RM4001-RM7000 > RM7000	Housing and Environment (x_{12})	Low Moderate High
Family Relationship (x_6)	Low Moderate High		

4. Results and Discussion

This section provides some comparison in terms of goodness of fit test (Table 3), pseudo R-square statistics (Table 4), parameter estimates, log likelihood and AIC values for the fitted models with four link functions. Table 5 shows the results of parameter estimates for ordinal regression models with four link functions, while Table 6 shows the comparison of log likelihood and AIC values.

TABLE 3. Goodness of fit of the ordinal regression model with different link function using Pearson chi-square

Ordinal Model with Selected Link Functions	Pearson chi-square	df	p-value
Logit	4357.87	2693	0.000
Probit	7810.00	2693	0.000
Clog-log	4083.16	2693	0.000
Nlog-log	269025.68	2693	0.000

TABLE 4. Pseudo R-square Statistics with different link function

Ordinal Model with Selected Link Functions	Pseudo R-square	
	Cox and Snell	Nagelkerke
Logit	0.444	0.573
Probit	0.433	0.559
Clog-log	0.431	0.556
Nlog-log	0.400	0.517

The result of goodness-of-fit demonstrated that all models with four different link functions are not well fitted ($p < 0.000$). Literature has mentioned that chi-square is likely to be highly significant with a large sample size, which in this study should be achievable as 2,808 samples has been used (Agresti, 2007; Lall, 2004). Moreover, chi-square is sensitive to empty cells especially when dealing with large number of categorical predictors. Therefore, other methods of assessing the goodness of fit, such as measures of association, like the pseudo R^2, are instead used. The results of R^2 value as shown in Table 4 indicated that the ordinal model with logit link function has the highest value of R^2 which is 0.573.

TABLE 5. Parameter estimates of determinants of level of family well-being using ordinal regression model with different link functions

Variables	Ordinal Regression Model with Selected Link Functions							
	Logit		Probit		Clog-log		Nlog-log	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Strata								
Rural (Ref)	-	-	-	-	-	-	-	-
Urban	0.032	0.155	0.009	0.062	0.092	0.085	-0.038	0.054
Ethnic								
Indian (Ref)	-	-	-	-	-	-	-	-
Chinese	0.296	0.189	0.162	0.95	0.198	0.136	0.039	0.091
Bumiputra	0.364*	0175	0.199	0.103	0.283*	0.127	0.050	0.084
Education Level								
Primary (Ref)	-	-	-	-	-	-	-	-
Tertiary	0.256	0.186	0.118	0.083	0.227	0.135	0.0130	0.087
Secondary	0.213	0.354	0.142	1.00	0.204	0.115	0.116	0.071
No Formal Education	0.275	0.248	0.172	0.136	0.185	0.188	0.339	0.117
Type of Family								
Nuclear (Ref)	-	-	-	-	-	-	-	-
Extended	0.216	0.254	0.096	0.083	0.074	0114	0.064	0.070
Blended	0.074	0.274	0.006	0.150	0.022	0.190	-0.105	0.137
Single-Parent	0.061	0.558	0.067	0.305	-0.058	0.354	0.111	0.293

Variables	Ordinal Regression Model with Selected Link Functions							
	Logit		Probit		Clog-log		Nlog-log	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Household Income								
≤RM2000 (Ref)	-	-	-	-	-	-	-	-
RM2001-RM4000	-0.040	0.136	0.027	0.074	0.173	0.193	0.039	0.116
RM4001-RM7000	-0.037	0.191	-0.017	0.103	0.107	0.136	-0.008	0.088
>RM7000	0.252	0.261	0.114	0.139	0.083	0.099	-0.028	0.064
Family Relationship								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	1.085***	0.228	0.556***	0.130	0.411***	0.153	0.650***	0.147
High	2.462***	0.233	1.294***	0.132	1.459***	0.156	1.189***	0.145
Economic Situation								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	0.481**	0.158	0.253**	0.089	0.269	0.114	0.245*	0.085
High	1.152**	0.172	0.635**	0.095	0.807*	0.135	0.538*	0.087
Health								
Low (Ref)	-	-						
Moderate	0.441*	0.200	0.218*	0.133	0.414*	0.153	0.123	0.113
High	0.940*	0.508*	0.276*	0.120	0.718*	0.157	0.401*	0.115
Safety								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	0.040	0.231	0.049	0.131	-0.108	0.159	0.199	0.132
High	0.529*	0.240	0.276*	0.130	0.299	0.173	0.322*	0.134
Community Relationship								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	0.381**	0.208	0.183	0117	0.164	0.142	0.174	0.116
High	0.620**	0.225	0.323*	0.124	0.359*	0.164	0.282*	0.120
Religion Practice								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	0.492*	0.199	0.261*	0112	0.181	0.135	0.304	0.114
High	1.139*	0.214	0.667*	0.121	0.709*	0.155	0.563*	0.118
Housing and Environment								
Low (Ref)	-	-	-	-	-	-	-	-
Moderate	0.070	0.167	0.008	0.093	-0.020	0.118	0.069	0.089
High	0.630*	0.183	0.391*	1.00	0.585*	0.140	0.376*	0.089
Intercept 1 (Low|Moderate)	1.451***	0.284	0.757***	0.159	0.048***	0.197	1.249***	0.174
Intercept 2 (Moderate|High)	4.914***	0.319	2.565***	0.172	2.445***	0.212	2.732***	0.187

*Significant at 5% level of significance ($p<0.05$); S.E = Standard Error

As can be seen from the Table 5, six out of seven social determinants which are family relationship, economic situation, health status, community relationship, religion practice and housing and environment are significant in determining the level of family well-being in all four link functions of the model. However, as shown in Table 6, the ordinal regression model with logit link function is found to provide the best fit, with the highest value of log-

likelihood (-1268.5) as compared to probit link-function (-1294.5), clog-log link-function (-1373.6) and nlog-log link function (-1301.5). This is also supported by the smallest AIC value for ordinal regression model with logit link function (2593.0) as compared to ordinal regression model with other link functions; probit link function (2645.0), clog-log link function (2658.9) and nlog-log link function (2803.2). Therefore, this finding suggested that the ordinal regression model with logit link function is the best fit model as compared to model with probit, clog-log and nlog-log to represent the level of family well-being data.

These results show that all social factors are identified as significant factors to determine the level of family well-being except safety. In contrast, comparing other models with different link function, the demographic factors such as ethnic is found to have a significant relationship with the level of family well-being. Interestingly, the significant parameter estimates are all positive, which means, each of the category of the contributed factors are more likely to have a higher-level satisfaction of family well-being.

TABLE 6. Performances ordinal regression model with different built-in link functions

Ordinal Model with Selected Link Functions	Log-Likelihood	Akaike Information Criteria
Logit	-1268.5	2593.0
Probit	-1294.5	2645.0
Clog-log	-1373.6	2658.9
Nlog-log	-1301.5	2803.2

5. Conclusion

The ordinal regression model with logit link function is found to be the best model for describing the relationship between the level of family well-being and the covariates which include ethnic, family relationship, economic situation, health status, community relationship, religion practice and housing and environment. It is clear that social determinants are the significant factors for explaining the level of family well-being. However, it is a bit surprising to see that demographic factors except for ethnic and socioeconomic factors are not significant. This could possibly be due to the Malays which represent the major race in the country who are usually easily get contented with the little things that they have.

Acknowledgement

This article uses sample data from the Malaysian Family Well-Being Index Survey (MFWIS) 2011 conducted by the National Population and Family Development Board Malaysia (NPFDBM).

References

1. Agresti, A. 2007. *An Introducing To Categorical Data Analysis*. John Wiley and Sons Ltd: t.pt.
2. Agresti, A. 2011. Examples of Using R for Modeling Ordinal Data. 2010.
3. Christensen, R.H.B. 2015. Regression Models for Ordinal Data.
4. Fagerland, M.W. & Hosmer, D.W. 2016. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation* 86(17): 3398–3418.
5. Javali, S.B. & Pandit, P. V. 2010. A comparison of ordinal regression models in an analysis of factors associated with periodontal disease. *J Indian Soc Periodontol* 14(3): 155–159.
6. Liu, I. & Agresti, A. 2005. The Analysis of ordered categorical data: an overview and a survey of recent developments. *Sociedad de Estadıstica E Investigacion Operativa Test* 14(1): 1–73.
7. Mccullagh, P. 1980. Regression models for ordinal data. Jil. 42.
8. Mccullagh, P., Journal, S., Statistical, R. & Series, S. 2014. Regression Models for Ordinal Data Regression Models for Ordinal Data. 42(2): 109–142.
9. Mckelvey, R.D. & Zavoina, W. 1975. A statistical model for the analysis of ordinal level, dependent variables. *Journal of Mathematical Sociology* 4(1): 103–120.
10. Mouriño, H. 2013. Ordinal regression models to describe tourist satisfaction with Sintra's world heritage. *AIP Conference Proceedings* 1558 1885–1888. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84887541632&partnerID=40&md5=6195187ed03422779e99f2331b29a13a.
11. Noor, N.M., Gandhi, A.D., Ishak, I. & Wok, S. 2014. *Development of Indicators for Family Well-Being in Malaysia*. *Social Indicators Research*, Jil. 115. t.tp.: t.pt. http://link.springer.com/10.1007/s11205-012-0219-1.

Item analysis in the assessment of knowledge in biostatistics among the postgraduate students of a medical college in northeast India

Dr. Rajkumari Sanatombi Devi, Dr. V.K Mehta
Sikkim Manipal Institute of Medical Sciences, Sikkim Manipal University, 5th Mile, Tadong, Gangtok, East Sikkim, Sikkim, Pin: 737102, India,

Abstract

The objectives of the study was to assess the quality of a multiple choice test items based on knowledge on Biostatistics using difficulty and discriminating indices and to find the correlation between the two indices. The test consisted of 38 items having one correct response and three wrong answers. The test was administered among the postgraduate medical students attending the research methodology classes conducted by the Department of Community medicine, SMIMS. Thirty-one items (81%) were within the acceptable range of difficulty index (0.20 to 0.80) and 7 (18%) items were discarded due to easy (> 0.80) and very difficult items (> 0.20). Twenty-seven items (71%) were falls within the acceptable range (0.20 to above 0.40) of the discriminating power of the test and 7 items had poor discriminating power (> 0.20). Two items each had 0 and negatives discriminating power. The mean (SD) score of the difficulty index was 47.53% ± 20.96%) while for discriminating index, it was 0.28 ± 0.20. Using Pearson correlation formula, it was observed that the two indices was strongly positively correlated which was significant at the 0.01 level of significant (r = 0.52, P=0.001). Hence, a significant positive correlation was observed between these two indices and the strength of correlation was strong in the study.

Keywords

Difficulty index; Discriminating index; Correlation; Multiple choice questions

1. Introduction

The adequacy of a test –whatever its purpose depends upon the care of with which the items of the test have been chosen (Garrett, 1966). Multiple choice question is an efficient tool in identifying the strengths and weaknesses in students, as well as providing guidelines to teachers on their educational protocols (Tan & McAleer, 2008). Ebel (1972) "Item analysis indicates the difficulty level of each item and discriminates between the better and poorer examinees. Thus, it helps in selecting and retaining the best test items in the final draft of the test rejecting poor items and also shows the need to review and modify the items". The difficulty of an item (problems or question) may be determined in several ways but the number right, or the proportion of the

group which can solve an item correctly, is the "standard" method for determining difficulty in objective examination. The proportion of passing an item is an index of item difficulty. If 90% of a standard group pass an item, it is easy; if only 10% pass, the item is hard. Discrimination index also known as validity of the test items or point biserial correlation is the other parameter used in item analysis. Item discrimination determines whether those who did well on the entire test did well on a particular test's item. The size of an acceptable validity index will depend upon the length of the test, the range of the difficulty indices, and the purposes for which the test is designed (Garrette, 1966). Fowell et al., (1999) Discrimination index (DI) describes the ability of an item to distinguish between high and low scorers. Discrimination index (DI), ranges between -1.00 and +1.00. It is expected that the high-performing students select the correct answer for each item more often than the low-performing students. If this is true, the assessment is said to have a positive DI (between 0.00 and +1.00), indicating that students who received a high total score, chose the correct answer for a specific item more often than the students who had a low overall score. If, however, the low performing students got a specific item correct more often than the high scorers, then that item has a negative DI (between -1.00 and 0.00). Garrette (1966) as a general rule, items with validity indices of 0.20 or more are regarded as satisfactory, but items with lower indices will often serve if the test is long. Item having zero validity are, of course, useless. These items and items having negative validity (a larger percent right in the bottom group than the top) must be discarded; or they must be carefully examined for ambiguities, inaccuracies and other errors. Carroll (1993) the difficulty and discrimination indices are often reciprocally related. However, this may not always be true. Questions having high p-value (easier questions), discriminate poorly; conversely, questions with a low p-value (harder questions) are considered to be good discriminators. In the present study, the definition of Difficulty index (DIF I) given by Frank S. Freeman defined as the proportion of certain sample of subjects who actually know the answer of an item was used. Index of difficulty for each test item can be calculated as

DIF I = (Ru + Rl) / (Nu + Nl)
DIF I = item difficulty
Ru = the number of students in the upper 27% who responded correctly
Rl = the number of students in the lower 27% who responded correctly
Nu = the number of students in the upper group
Nl = the number of students in the lower group
The difficulty index (DI) for an item was categorizes as followed

Cut off point of Difficulty index (Dif I)	Quality of item
Below 0.20	Very difficult
0.20 0.50	Good
0.50 – 0.80	Best
Above 0.80	Very easy

Discrimination power (DI) defined by Blood and Budd (1972) as the ability of an item on the basis of which the discrimination is made between superiors and inferiors was used in assessing the discriminated power of the test items in the present study. The formula for calculating DI was given as

DI = (Ru - Rl)/ 0.5 N, where
DI = discrimination index
N = total no. of correct responses
Ru = the number of students in the upper 27% who responded correctly
Rl = the number of students in the lower 27% who responded correctly
Ebel and Frisbie, (1986) rule of thumb for determining the quality of items with respect to their discrimination index was used in the present study. It was given as

Cut off point of Discrimination index (DI)	**Quality of items**
Above 0.40	Excellent
0.30 to 0.39	Good
0.20 to 0.29	Mediocre
Below 0.20	Poor
0.0	No discrimination
Negative value	Worst

2. Justification of the study

An achievement test is one of the important aspects of teaching and learning process which involves both the teacher and the learner. The basic purpose of administering the multiple choice questions/items is to measure whether the students are able to understand the content of the subject matter of a particular subject. One of the important characteristics ofa good tests are the reliability and validity of the test items. It is very difficult to understand whether the test items are reliable and valid by simply looking at the options given in a test. A reliable test items is free from bias that there is no chance of influencing unsystematic errors in measurement even when the test are administered to others groups in different condition by different

administrators. On the other hand, validity of a test is influences both the unsystematic and systematic error of measurement. Hence, a test may be reliable without being valid, but a test can't valid without being reliable. Therefore, reliability is a necessary condition but not a sufficient condition of validity of a test. Item analysis is a procedure that provides information regarding the reliability and validity of a test. Hence, analysing the questions/items related to questions to knowledge on Biostatistics will help the researcher to improve her skill in constructing a good quality multiple choice questions, to identify the weak and strong areas of the course content of the subject Biostatistics while teaching the PG students in future.

Objectives

1. To assess the quality of a multiple-choice test items related to the questions on knowledge on Biostatistics using difficulty and discriminating indices of item analysis
2. To find out the correlation between difficulty and discriminating indices of the test items

3. Methodology

The study was a descriptive cross-sectional survey study. All the PG (MD/MS/DNB) students enrolled during 2013 to 2017 under Sikkim Manipal Institute of Medical Sciences, Sikkim Manipal University students were the target population. There were 38 multiple choice questions (MCQs) having one correct answer and three false choices measuring knowledge on Biostatistics. It was developed by the researcher herself with the help of subject expert and as well as from literature searched from internet similar to the objective of the present study. After obtaining the requisite permission from the institution Ethics committee, the researcher distributed the questionnaire to the PG students while they were attending the research methodology classes conducted by the department of Community Medicine, SMIMS. The fill up form was collected directly from them by the researcher. Thirty minutes was given to fill up the form. Verbal consent was taken in collecting the data students after explaining the purpose of the study. Data was entered into Microsoft Excel sheet and analysed using SPSS Version 16.0

4. Result

A total of 38 MCQs were constructed and assessed the correct response among the 64 PG students. There was no negative mark for wrong answers. For classifying the students into two different groups, scores on knowledge on Biostatistics were ranked in descending order from highest score of 55 to lowest score 8. The first 27% students from 64 students (27/100 X 64 = 17) were included in high achiever group (HAG) and the last 27% (27/100*64=17)

were included in low achiever group (LAG). Students in between these two groups that are the middle 30 students were excluded in the evaluation as their pattern of mark scores remain nearly the same (range 16 to 20 scores). Fig.1 showed that of total 38 items, 4 items were of very difficult (Dif I < 0.20) and 3 items were very easy (Dif I above 0. 80) items, 14 items were good (Dif I 0.20 to 0.50), while 17 items were best (Dif I 0.50 to 0.80). Hence, 7 items (4 very difficult and 3 very easy items) were found to be poor in the study. In the present study, the overall mean (SD) of correct responses was 18.05 ± 4.44 (maximum 38 marks). Correct mean scored of HAG and middle and LAG were 23.53 ± 1.77, 18.07 ± 1.48 and 12.59 ± 2.60 respectively. Mean Difficulty index was 47.53 ± 20.96%. Mean Difficulty index of 31 items within the acceptable range of 0.20 to 0.80 (good and best items) was 48.67 ± 15.47%. Mean of 7 poor items based on 3 very easy and 4 difficult items was 42.44 ± 38.53%

Fig.1 Bar diagram showing the selected items based on item difficulty

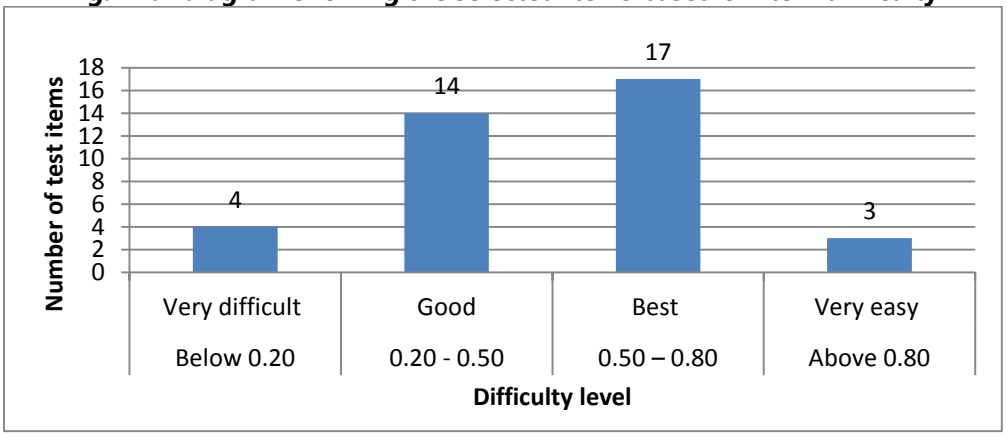


Fig.2 showed that discriminating power of 10 (26%) items was excellent, of 6 (16%) items was good, 11 (30%) items had average discriminating power, and of 7 (18%) items was poor. Two items each were having 0 and negative discriminating power. Overall mean DI was 0.29 ± 0.20. Mean DI for 16 ideal items based on 10 excellent and 6 good items was 0.46 ± 0.12. Mean DI of 11 mediocre and 7 poor items were 0.28 ± 0.02 and 0.10 ± 0.05 respectively.

Fig.2 Bar diagram showing the selected items based on power of discrimination

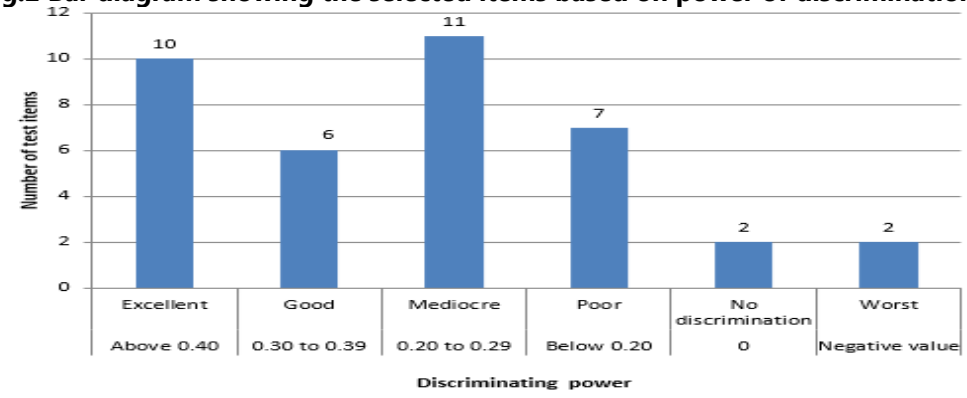
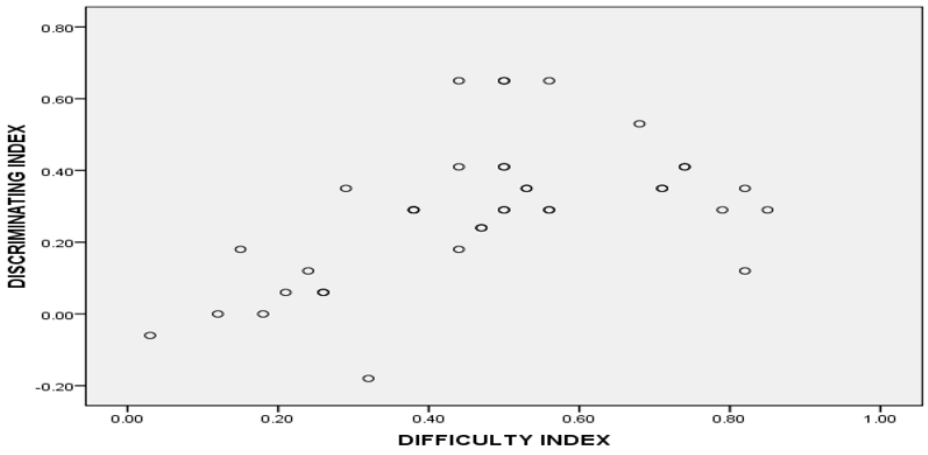


Fig.3 showed the scatter diagram of correlation between the DIF I and DI for each test items. The figure showed that the relationship between the DIF I and DI are not linear, data points for scores showed a curvilinear relationship between the DIF I and DI. The maximum discriminating power of item fall within the DIF I of 0.40 to 0.60. The discriminating power decreases in the difficulty range as compared with the range of easy items of DIF I. By applying the Pearson correlation coefficient "r", a strong positive correlation was observed between the two indices (r = 0.521, P = 0.001) which was statistically significant at 0.01 level of significant.

Fig.3: Scatter diagram showing the correlation between the difficulty index and discriminating for each test item



5. Discussion

According to Brown and Frederick (1971), item analysis has two purposes: First, to identify defective test items and secondly, to indicate the content the learners have or have not mastered. In the present study, the overall mean value of correct responses was 18.05 ± 4.44% (maximum 38 marks). Correct mean scored of high and middle and low groups were 23.53 ± 1.77%, 18.07 ± 1.48% and 12.59 ± 2.60% respectively. The values of standard deviations showed that the variability was very high in low group as compared with high and middle groups of the students. Overall mean Dif I was 47.53 ± 20.96%. Mean Dif I of 31 items within the acceptable range of 0.20 to 0.80 (good and best items) was 48.67 ± 15.47%. Mean Dif I of 7 poor items (very difficult and very easy) was 42.44 ± 38.53%. Hence, 31items were retained and 7 items which were very easy (3 very easy and 4 very difficult items) were rejected as these items contribute little to the discriminating power of an item. Overall mean DI was 0.29 ± 0.20. Out of 38 items, 10 (26%) items had excellent DI with mean value of 0.52 ± 0.12 and 6 (16%) items had good DI with mean 0.35 + 0.00. Combining the two indices, 16 (42%) items could be called as ideal items of the test. Mean DI of 16 ideal items was 0.46 ± 0.12. Hence, based on the DI of the present study, 27 (71%) items (10 excellent plus 6 good plus 17

mediocre items) were retained and 7 (18%) poor items need to be revised due to poor discriminating power between the HAG and LAG. These items need improvement in choosing the correct key and options given in the test items. Two items each were having 0 and negative discriminating power in the study. These 4 items need to be completely removed in the construction of test items. In a study of MCQs having 5 options, the mean score was 27.31 ± 5.75 (maximum 50 marks). Mean Dif I-value and DI were 54.14 ± 17.48 and 0.356 ± 0.17, respectively. Seventy eight per cent items were of recommended difficulty with mean 51.44 ± 11.11. Sixty two per cent items had excellent DI (0.465 ± 0.083) (Hingorjo & Jaleel, 2012). The mean Dif I and DI were higher in their study as compared with the present study but the variability in both Dif I and DI were higher in the present study. In the present study, the relationship between Dif I and DI was not in linear. It was in curvilinear form. Pearson correlation coefficient "r" showed a significant positive correlation between these two indices and the strength of correlation was strong (r = 0.51, P = 0.001). However, a wide scattering of dots were observed between the test items which indicate the guessing practice done by the students. Suruchi and Rana (2014) found that out of 120, one very difficult and seven very easy test items were rejected for the final draft of achievement test based on the difficulty index. Fourteen items were found to be good, 2 items needed improvement and 6 items which fell below 0.20 of discriminating index were rejected. One item was found to have zero discriminating power and none of the test item showed negative discrimination. However, in their study no wide scattering data points was observed which was in contrast of the present study. Thus, 9 items were recommended to reject for the final draft of achievement test. In their study, the relationship between Dif I and DI was a moderate negatively correlated (r = - 0.3711). Mitra et a.l, (2009) reported that difficulty and discrimination indices are reciprocally related and poor correlation was observed in their study. Kheyami et al., (2018) found a significant dome-shaped correlation between DIF I and DI (r = 0.162; P = 0.010), with the highest DIs occurring in the acceptable DIF I range and decreasing for DIF Is in the difficult range which is similar with the finding of the present study.

6. Conclusion

The finding of the study signifies about the importance of item analysis in the selection of a reliable and validity test items in a multiple-choice types questions. Based on the difficulty index of the test items, 31 items that comprised of 14 good and 17 best items were retained, and 4 difficulties and 3 very easy items were rejected. Discriminating power of the test items indicates that 10 excellent, 6 good and 11 mediocre items were retained, and 7 poor items need to be revised due to the poor discriminating power of the

test items. The reasons of the poor discriminating power may be due to the confusing meaning of the options used test items. Sixteen items can be considered as ideal test items based on 10 excellent and 6 good items. Two items were having zero discriminating power between the high achieving group and low achiever group. It was also observed that 2 items were having negative values which indicate the low performing students got a specific item correct more often than the high scorers. The reasons of having the negative values may be due to the mis-keyed in test items. Hence, these 4 items need to be discarded in the construction test items. It was also clear that there was a strong and significant positive correlation between the difficulty and discriminating indices of the test items. The difficulty and discrimination indices are often reciprocally related is not coming true in this study.

References

1. Blood, D.F., & Budd, W.C. (1972). *Educational measurement and evaluation*. New York: Harper and Row.
2. Brown, Frederick G. (1981). "*Measuring Classroom Achievement",* Halt Richard and Winston, U.S.A.", pp. 101-110, 224p.
3. Carroll, R.G. (1993). Evaluation of vignette-type examination items for testing medical physiology. *Am J Physiol, 264*, S11- 5.
4. Kheyami, D, Jaradat A, Al-Shibani T, Ali F A.(2018).Item analysis of multiple choice questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Med J*, *18* (1), pp. e68–74.
5. Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of Educational Measurement* (5th Ed.).New Delhi: Prentice Hall of India Pvt. Ltd.
6. Ebel, R.L. (1972). *Essentials of Educational Measurement* (1st Edition). New Jersey: Prentice Hall.
7. Fowell, S.L., Southgate, & L.J., Bligh, J.G. (199). Evaluating assessment: The missing link? *Med Educ*, *33*, 276-81.
8. Freeman, F.S. (1962). *Theory and Practice of Psychological Testing*. New Delhi: Oxford &Ibh publishing.
9. Garrett, H.E. (1966*). Statistics in Psychology and Education*. Paragon International Publishers; pp.362.
10. Hingorjo, M.R., & Jaleel, F. (2012). Analysis of one-best MCQs: The Difficulty Index, Discrimination Index and Distracter Efficiency. *J Pak Med Assoc*, *62*,142-7.
11. Mitra,N.K., Nagaraja, H.S., Ponnudurai, G. & Judson, J. P. (2009). The levels of Difficulty and Discrimination Indices in type a multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *Int e-J Sci Med Educ, 3* (1), 2-7.
12. Suruchi, & Rana, S.S. (2014).Test Item Analysis and relationship between Difficulty Level and Discrimination Index of test items in an achievement test in Biology. *Paripex - Indian journal of research*, *3*(6), 56-58.
13. Tan, L.T., McAleer, J.J; & Final FRCR Examination Board. (2008).The introduction of single best answer questions as a test of knowledge in the final examination for fellowship of the Royal College of Radiologists in Clinical Oncology. *Clin Oncol (R Coll Radiol), 20*, 571-6.

Subjective and community wellbeing interaction in multilevel spatial modelling framework

Wlodzimierz Okrasa, Dominik Rozkrut
Statistic Poland

Abstract

Analyzing the cross-level interaction between individual and community well-being requires joint involvement of both 'vertical' and 'horizontal' perspectives. While multilevel modelling separates the effects resulting from personal characteristics from those resulting from community features, accounting for spatial variation and geographic membership prove that space and place matter too. To this aim, the explicitly spatial multilevel model is developed that allows to identify both types of effects (space and place-related) using the hierarchical (nested) data structure, with the lowest level administrative areas (NUTS5 units/ Nomenclature of Units for Statistical Purposes), communes/gminas) as the aggregate-level context for its members/residents. There are two kinds of well-being measures used in the ensuing analysis: individual (subjective) well-being measure derived from the nation-wide Time Use Survey data, replaced occasionally by 'life satisfaction' type of selfreported measures, and multidimensional index of local deprivation composed of eleven domain-scales. An empirical application of the multilevel spatial modelling (which constitutes the major portion of the remaining part of the paper) is preceded by searching for main factors and auxiliary covariates affecting individual (subjective) well-being, while looking after the issue of endogeneity. When expressed in a way analogous to so-called basic 'life-satisfaction equation', subjective well-being might be treated as a function of residents' income and hours of work vis-a-vis the impact of community well-being (or deprivation) through employing causal type of reasoning using path analytic version of structural model. Another important factor at the community level (referred often to social cohesion) is social capital the relative impact of which (weighted against individual income) is checked using the 'compensating variation' approach. The spatial multilevel modelling is finally extended by an attempt to assess the spatial interaction effect on the cross-level relationships. Its inclusion is recommended in the concluding this paper discussion suggesting a more systematic efforts toward a spatially integrated approach to such a type of modelling problems.

Keywords

spatial analysis; measuring subjective well-being; community deprivation; social capital

1. Introduction

There are several reasons for focusing on community and individual well-being relationships, especially in the local development context. Many of them have been recognized and discussed thoroughly in the literature, challenging the tradition of using GDP and other economic indicators as measures of social progress, (Stiglitz et al., 2009), while including subjective values based self-reported feeling about selected aspects of wellbeing in connection with community, eg. Phillip and Wong (2017) . In the presented modelling approach, an empirical application is preceded by discussion of the measurement and data issues, including problem of creation an analytical multi-source database (through 'bottom up' integration of units from different surveys) and construction of the major wellbeing measures: (i) multidimensional index of local deprivation encompassing eleven components , each of them being constructed from public-use data file (Local Data Bank, Statistics Poland), using 'confirmatory' version of factor analysis (for all 2478 communes (gminas)), and (ii) individual (subjective) well-being measure derived from the nation-wide Time Use Survey which is substituted in some contexts by self-reported measures from national surveys on Social Cohesion or Social Diagnosis). An important methodological question that arises in modeling the processes underlying cross-level relationships in the spatial perspective concerns the omitting variable. Since it can be associated with both micro- (or 'response' variable) and macro- (or predictors) level, the issue of endogeneity becomes troublesome and demands the evaluation of possible effect of the crucial omitted variable. Two aspects of this issue is considered here. On the one hand, it is hypothesized that the level of community deprivation (or well-being) is a mediating factor that modify the way in which the residents' material status (income) affects the individual (subjective) well-being. A path analytic version of structural model is employed to decompose total effect of the independent variable into the natural direct and indirect effects (Hong, 2015; Okrasa and Rozkrut, 2018). On the other hand, social capital - indicated by the intensity of the third sector organizations' presence in a community - can be interpreted as the amount of money required to compensate a person for a possible loss in utility (for instance, like when price is rising). The 'compensating variation' approach to social capital allows to identify the utility gain derived from a unit increase in social capital (Anand and Montovani, 2018, Okrasa, 2018). Following exploration of spatial patterning, clustering and spatial dependence (with GeoDa procedures - Fischer and Getis, 2010) a direct assessment of the spatial interaction effect on the cross-level relationships is also attempted (Patuelli and Arbia, 2016) using flow-type data from between-community migration public statistics. In conclusions, a spatially integrated approach to vertical (multilevel) and horizontal (across areal units) relationships between individual

(subjective) and group (community) measures of well-being is discussed toward elaborating a comprehensive methodological framework (encompassing the relevant issues involved in such a type of joint modelling approach).

2. Methodology : Conceptualization and Operationalzation . Data and Models

Increasing focus on well-being (along the beyond-GDP paradigm) results also in several guidelines and recommendations on the measurement of subjective well-being in public statistics - eg. OECD (2013, 2015), Stone and Mackie (2014); Kalton et al (2015). While there is a consensus in the literature regarding individual (subjective) well-being measures that they are supposed to cover all or some aspects of its conceptually triadic structure - evaluation (eg. Satisfaction from Life) experience (eg. How did you feel yesterday) and eudaimonic (eg. Sense of Life) - the community well-being measurement approaches still awaits similar elucidation (eg. Kim and Ludvigs (2017)), although several country-specific approaches are already well developed within public statistical systems (to mention Australia, Canada, USA, UK, and others).

a. Individual (Subjective) well-being: Time Use Survey/TUS data-based measures

Since psychometric, self-reported data-based measures of well-being are often criticized for their arbitrariness and low reliability, data from time use surveys (collected with day reconstruction metod/DRM) are recommended instead - see Kahneman and Krueger (2006). Amount of time spent by respondent on performing an activity with information on emotion (negative-neutral-positive) s/he associates with this activity ('time of unpleasant state') is reflected by the value of U-index :

$\boldsymbol{U}i = (\boldsymbol{\Sigma}j\ \boldsymbol{I}\_ij\ \boldsymbol{h}\_ij)/\boldsymbol{\Sigma}_j\boldsymbol{h}_{ij}$ (in TUS conducted in 2013: I = -1. 0. +1) (1)

And $U = \Sigma\_i\ (\Sigma\_j\ I\_ij\ h\_ij)/\ \Sigma_j h_{ij})$/ N for N-persons / group in population

b. ***Community Well-Being (CWB)*** is a multifaceted and multilevel concept, hardly covered by standardized procedures of operationalization and measurement. It is a "concepts developed by synthesizing research constructs related to resident's perceptions of the community, … needs fulfillmemnts, observable community conditions, and the social and cultural context…" (Sung and Phillips 2016:2 [in Phillpis and Wong2017: xxix]). Among the important features of CWB is often included community cohesion (or local , spatial cohesion), which is here interpreted as any of the possible configuration of the economic cohesion and/or social cohesion and/or territorial cohesion (following Kearns and Forest (2001), Both types of measures - individual and community well-being - constitute the main input of the Analytical Multi-source Database (AMDb), embracing Multidimensional Index of Local

Deprivation (MILD) for 2478 communes /gminas (NUTS5), composed of eleven (pre-selected) domains of deprivation - each characterized by a number of original items: ecology – finance – economy – infrastructure – municipal utilities – culture – housing – social assistance – labour market – education – health [65 items]

c. Individual and community level factors of subjective well-being

Basic Wellbeing Equation – important types of tradeoff/ balance.

- Approximation of wellbeing equation (originally 'life satisfaction equation', eg. Clark (2018)) allows to consider first the classic hypothesis of work (time, h- hours) vs. earning (Y) tradeoff, including also auxiliary covariates (X):

  Well-Being = $\beta_1 Y + \beta_2 h + \theta' X + \varepsilon$ (2)

- Complementary to the above considerations lead to checking the role of community's social capital, the role of which can (hipothetically) be interpreted in terms of 'compensating variation' (CV) as discussed by Anand and Montovani (2018).

Formally, a life satisfaction equation can be re-written as:

$U^0(y^0, SC^0) = U^1(y^0, CV, SC^1)$ (3)

where y is household income, SC stands for SC, and CV for compensating variation (of CV for y), which can be can be obtained by identifying the utility gain derived from a unit increase in social capital. Accordingly, the expected utility given any particular value of social capital can be written as:

$E(Ui|SCi,yi,Xi) = \beta 0 + \beta yyi + \beta scSCi + \gamma' Xi + \varepsilon i$ (4)

where X represents all additional covariates. Following Anand and Montovani (2018), CV can be be defined as

$CV = \beta SC / \beta y.$ (5)

d. *Individual well-being and community well-being relationship* - a multilevel model (Subramanian (2010), Lloyd (2011)). Using notations:

- $\boldsymbol{y_{ij}}$; well-being of i individual in j commune/gmin ;
- $\boldsymbol{x_{1ij}}$ predictor of indywidual (level-1) – such as: age. education. or satisfaction (e.g. from life in a community. family life . etc.)
- predictor of level-2 / (macro-level): Multideminsonal Index of Local Deprivation for jcommune/gmina /MILDj

Model for level-1: $y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{0ij}$ (6)

where: β_{0j} – refers to x_{0ij} average score on a well-being scale in j-th commune/gmina (eg.. . 'less affluent' or 'low-income'. etc.. for cases < Me. x_{0ij} =1);

β_1 – average differentiation of individual well-being associated with individual material status . (x_{1ij}). across all territorial units (communes/gminas;); e_{0ij} – residual term for the level-1. Two-level model can be specified as below:

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{1ij} + a_1 w_{1j} + a_2 w_1 x_{1ij}} + (u_{0j} + u_{1j} x_{1ij} + e_{1ij} x_{1ij} + u_{2ij} x_{2ij}) \quad (7)$$

- where w_{1j} is a 2-level predictor. i.e. the index of local deprivation. $MILD_{1j}$. The following model was calculated using data from Time Use Survey 2013 (22 695 and 24 065 persons surveyed for weekdays and for weekends/holidays. respectively):

$$IBW(U-index)_{ij} = \beta_{00} + \beta_{10} education_{ij} + \beta_{20} age_{ij} + a_1 MILD_j + a_{11} education_{ij} * MILD_j + a_{21} age_{ij} * MILD_j + u_{1j} education_{ij} + u_{2j} age + u_{0j} + e_{ij} \quad (8)$$

[It is assumed that] Such a specification of cross-level (between individual and community/gmina measures of well-being) with cross-level interaction effect should ensure robust estimation (e.g.. Subramanian. op. cit.. p. 521; Hox et al.. 2018

e. *Spatial aspects -* *checking for spatial dependence Estimation of the spatial regression model parameters*: (notation for individual observation i):

$$y_i = \rho \sum^n j = 1 W_{ij} y_j + \sum^k r = 1 X_{ir} \beta_r + \varepsilon_i \quad (9)$$

$where: y_i -$ the dependent variable for observation $i; X_{ir}\, k -$ explanatory variables. r = 1. …. k with associated coefficient $\beta_r; \varepsilon_i$ is the disturbance term; ρ is parameter of the strength of the average association between the dependent variable values for region/observations and the average of them for their neighbours (eg.. LeSage and Pace. 2010. p. 357). The above specification of the spatial regression model assumes that ε_i is meant as the *spatially lagged* term – versus *spatial error* formulation - for the dependent variable (which is correlated with the dependent variable). that is: $\varepsilon_i = \rho W_i . y_i + X_i \beta + \epsilon_i$. Both types of models are used below to check *how* and *why* 'place' and '*space*' matter.

3. Result

Some at a glance results and observations [comments to be added]

3.1 *Impact of income vs. work time* (acc. to 'well-being equation'): Opposite directions of influence of income and time-in-work on wellbeing , according to U-index: while greater income is positive for individual wellbeing, the increased amount of time spent on work is negative (U- index increases) - question arises about the point of balance (trade-off between the two factors of wellbeing). (See Kahneman and Deaton 2010 for comparison of income effect)

Table 1. Approximation of *wellbeing equation*(or 'life satisfaction equation' work *vs.* earning *tradeoff* hypothesis

Model /predictors (of Uindex)	Unstand. Coefficients		Stand. Coeff.	t	Sig.
	B	Std. Error	Beta		
(Constant)	0.014	0.025		0.553	0.580
• Job-time (main and additional)	0.005	0.000	0.297	26.766	0.000
• Income of H'holdpc - monthly (average in year)	-1.836E-05	0.000	-0.087	-7.353	0.000
• MILD_2014 /Multidimensional Index of Local Deprivation	0.000	0.000	0.117	6.865	0.000
• Subsidies Real < Simulated as 'fair'	-0.011	0.002	-0.068	-7.043	0.000
• Risk associated w/deprivation in Local Social Welfare	-0.036	0.002	-0.637	-15.921	0.000
• Risk associated w/deprivation in local labour market		0.003	0.799	18.934	0.000
• Ratio of 'in-work' to 'not-in-work'	-0.010	0.001	-0.078	-6.995	0.000
• Rural	-0.005	0.003	-0.023	-1.599	0.110
• U-R mixed	-0.013	0.002	-0.067	-5.262	0.000
Adjusted R Square = 0.178;		F(9,11101) 268.594; p< .000			

3.2. Community vs. individual 'assets' - relative impact of social capital vs. income. Substantial potentially 'compensating' effect of the community's social capital on individual well-being (acc. to Uindex)

Table 2. The *Wellbeing Equation* expanded by community cohesions-relevant variables

Model /predictors (of U-index)	Unst. Coefficients		St. Coef	t	Sig.
	B.	St. Error	Beta		
(Constant)	0.029	0.027		1.068	0.285
Job-time (main and additional)	0.004	0.000	0.285	24.630	0.000
Income of H'hold pc - monthly	1.841E-05	0.000	**-0.087**	-6.987	0.000
MILD_2014 /Multidimensional Index of Local Deprivation	0.000	0.000	0.118	6.630	0.000
Subsidies Real < Simulated as 'fair'	-0.011	0.002	-0.070	-6.887	0.000
Risk associated w/deprivation in Local Social Welfare	-0.036	0.002	-0.649	-15.626	0.000
Risk associated w/deprivation in local labour market	0.050	0.003	0.809	18.454	0.000
Ratio of 'in-work' to 'not-in-work'	-0.010	0.001	-0.080	-6.900	0.000
Rural	-0.007	0.003	-0.030	-1.978	0.048
U-R mixed	-0.014	0.002	-0.074	-5.547	0.000
Trust in local authority	-0.002	0.001	**-0.032**	-3.468	0.001
Satisfaction from the place of residence	-0.002	0.001	-0.017	-1.898	0.058

R Sq $_{Adj}$ = 0.18 ; F (11,10 095) = 198.387; p< .000 CV = -0.032/ -0.087

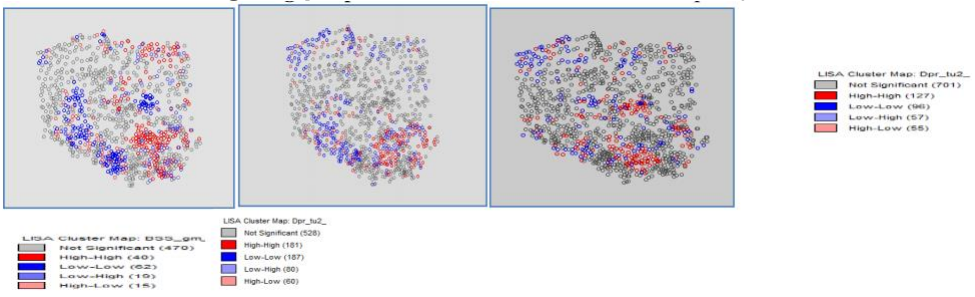
3.3. Cross-level relationships - individual and community level factors.

Table 3. Multilevel regression of individual well-being - U-index – on individual and commune/*gmina* level variables with cross-level interaction terms.

Model /predictors	Weekdays		Weekend /holiday	
	Beta	t	Beta	t
Constant	(.726)**	(6.316)	(.333)**	(3.515)
Education	-.085	-1.136	-.089	-1.209
Age	-.299**	-4.015	-.008	-.105
Multidimensional Index of Local Deprivation /MILD	-.098	-2.556	-.046	-1.209
Education * MILD	.142*	1.900	.145*	1.97
Age * MILD	.115	1.497	-.029	-.383
Urban (rural omitted)	.011	1.280	.016*	1.966
	F (6. 22698) = 174.860**		F (6. 24 068) = 23.515**	

3.4. Spatial autocorrelation and spatial clustering. Moran's I for presented below maps (from the left): (a) I=0.20 for local deprivation (MILD); (b) I=0,09 for U-ndex; (c) I= 0,10 for 3rd sector units.

Fig. 1. Spatial autocorrelation - Moran's maps



The spatial patterns of local deprivation and subjective well-being (both interpreted in 'negative' terms show one important feature in common - they both tend to cluster around high or low values of each of these measures in similar part of the country. In south-east dominate cluster of high deprived communes and also of communes with residents high on the U-scale (unpleasant state). Therefore the joint spatial distribution of communes (gminas) according to both measures is presented at the panel (c).

Table 4. Spatial dependence / spatial regression of SWB on commmune's attributes and compositional characteristics

SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Dependent Variable : *U –index* Number of Observations: 937; Number of Variables : 8; Degrees of Freedom : 929 **Lag coeff. (Lambda) : 0.43**; R-squared : 0.12

Variable	Coefficient	Std.Error	z-value	Probability
Constant	0.523731	0.042847	12.2233	0.00000
Monthly income	-0.002730	0.001960	-1.40359	0.16044
Age_avg (%)	-0.014313	0.005653	-2.53177	0.01135 *
Education_hs+ (%)	0.000381	0.000222	1.71849	0.08571 *
Not working pop. (%)	-0.001304	0.000273	-4.77623	0.00000 *
Index of loc.depr.-ecology	0.000560	0.000462	1.21309	0.22510
Index of loc. depr._Soc. Welfare	-0.000415	0.000312	-1.32693	0.18453
Subsidies_pc	1.2323e-005	1.1588e-05	1.06344	0.28758
Lambda	0.431769	0.0677941	6.36883	0.00000

4. Discussion and Conclusion

Research on individual and community well-being requires data from both individual and community level and both objective and subjective measures in order to explore effectively relationship in which they remains, and are influenced by such crucial factors as community cohesion, including social capital. Bringing space into analysis gives insight into processes which actually take place on a larger scale than own community –spatial dependency confirms this, suggesting spatio-temporal analytical framework. In particular, for the purpose of rational policy design and evaluation. Individual wellbeing increases along with greater household income. However, community deprivation reinforces significantly the subjective well-being effect of individual income. Also, deprivation in several domains shows negative association with U-index (such as risk associated with deprivation in local social welfare). Working with existing databases. eg. public files of official statistics has its advantages and disadvantages, which needs to be recognized to enhance integration procedures in constructing multisource analytical database.

References

1. Anand P., Montovani I., (2018) The Value of Individual and Community Social Resources. In New Frontiers of the Capability Approach(eds) F, Fennell S, and Anand, PB Cambridge U.Press
2. Clark A. E., (2018) Four Decades of the Economics of Happiness: Where Next? Review of Income and Wealth. Volume 64, Issue 2. https://doi.org/10.1111/roiw.12369
3. Fischer M.M., Getis A.,(eds). (2010) Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer.
4. Hong G., (2015) Causality in a Social World: Moderation, Mediation and Spill-over. Wiley.
5. Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective wellbeing. Journal of Economic Perspectives, 20, 3-24.
6. Kalton G., Mackie Ch., Okrasa W., eds. (2015) The Measurement of Subjective Well-Being in Survey Research. Statistics in Transition new series. Vol. 16, No. 3.
7. Kim Y., Ludwigs K., (2017) Measuring Community Well-Being and Individual Well-Being for Public Policy. In: R. Phillips & C. Wang (eds.), Handbook Of Community Well-Being Research. Springer.
8. Lloyd C.D., (2011) Local Models For Spatial Analysis. CRC Press Taylor & Francis Group.
9. Okrasa W., Rozkrut D., (2018). The Time Use Data-based Measures of the Wellbeing Effect of Community Development. Proceedings of the FCSM2018/Federal Committee on Statistical Methodology Research Conference. [in press].
10. Patuelli R., Arbia G., (Eds), (2016)Spatial Econometric Interaction Modelling. Springer
11. Subramanian S.V., (2010) Multilevel Modeling [in] Fischer

Longitudinal analysis of financial ratios and economic indicators of Italian firms in the period 2008-2017

Matilde Bini[1], Lucio Masserini [2], Alessandro Zeli[3]
[1] Department of Human Sciences, European University of Rome
[2] Department of Economics and Management, University of Pisa
[3] Division for data analysis and economic, social and environmental research, ISTAT

Abstract
The Great Recession derived from USA subprime crisis involved the European countries in two different steps: in the first phase the finance contraction and the bank failures spread out across the whole Atlantic area involving, above all, the financial market but also influencing the real economy. The second phase involves in depth the Euro zone and sovereign debts of the Countries until 2015. Also for Italy, the crisis period was from 2008 to 2011 characterized by a deep negative conjuncture until 2009 and by a slight recovery until the first half of 2011, and from 2011 to 2015, characterized by an intense recession. The aim of this work is to collect evidence on the Italian manufacturing system with the following goals: to calculate the main financial ratios related to firms' riskiness and distress risk trend by means of the book-value data; to detect the guide-variables outlining the firms' riskiness and distress risk trend in the period 2008-2017 to investigate the riskiness-distress risk trend for manufacturing industries and try to understand the effects of the Great Recession on this important business indicator trend. To perform this analysis a Latent Growth Curve Model is proposed, using an important Italian private database containing the book-value data of the joint-stock company Italian firms.

Keywords
Firms' riskiness; Latent Growth Curve Model; Longitudinal model; Panel data

1. Introduction

The Great Recession derived from USA subprime crisis involved the European countries in two different steps: in the first phase the finance contraction and the bank failures spread out across the whole Atlantic area involving, above all, the financial market but also influencing the real economy. The second phase, after the temporary recovery in 2010, involves more and more in depth the Euro zone and sovereign debts of the Countries in that area and it lasted at least until 2015. Also for Italy, the crisis period from 2007 to 2014 can be divided in two sub-periods: the first, from 2008 to 2011, is characterized by a deep negative conjuncture until 2009 and by a slight recovery until the first half of 2011, the second, characterized by an intense

recession, begun in 2011 and it endured until 2014. The aim of this work is to collect evidence on the Italian manufacturing system with the following goals: to calculate the main financial ratios related to firms' riskiness and distress risk trend by means of the book-value data (Zeli & Mariani, 2009; Zeli, 2014); to detect the guide-variables outlining the firms' riskiness and distress risk trend in the period 2008-2017 (Di Clemente, 2008; Kudlyak & Sánchez, 2017); and to investigate the riskiness-distress risk trend for manufacturing industries and try to understand the effects of the Great Recession on this important business indicator trend (Graham et al. 2011; Giroud & Mueller, 2015). To perform this analysis a Latent Growth Curve Model is proposed, using an important Italian private database containing the book-value data of the joint-stock company Italian firms.

2. Measures of firms' riskiness-distress

A large part of literature is aimed to properly classify the "signal" of bankruptcy coming principally from book values and standard financial statements (Altman et al., 1994; Bottazzi et. Al, 2011). A lot of indicators are considered in literature, among these, the most largely used ones cover area of economic enterprises' accounting related with financial distress such as: liquidity, leverage, profitability.

The interest coverage compares net profits to interest on loans and thereby expresses the firm's vulnerability linked to liquidity. It can be seen as a short-term risk indicator (the lower interest cover index, the higher the probability of financial distress). The leverage indicates the best indicators of the financial distress, because the possibility to pay back the debts decreases when the leverage increases (the lower leverage, the lower the probability of financial distress). Profitability, measured by means of ROE, assesses the ability to achieve a minimum profit share level, after covering costs.

In literature there are three approaches to bankruptcy prediction: accounting approach, analytical approach and statistical one. This last approach offers many statistical models that use balance sheet data. Statistical procedures (multiple discriminant analysis, logit or probit) were the most used methods in this kind of problem. Among them the factorial and latent analysis can be applied. A latent growth curve model is, innovatively, proposed to analyse riskiness-distress trend for manufacturing firms in 2008-2017 period. Unlike traditional longitudinal data analysis techniques, LGM allows researchers to make inferences about individual level effects as well as group effects.

3. The Data

We utilize an important private database AIDA (Analisi Informatizzata delle Aziende Italiane) containing the book-value data of the joint-stock company Italian firms over the period 2008-2017. Data were preventively checked and controlled according to the following steps: longitudinality check: only firms with at least 5 presences in the period were considered; coherency check: all firms with fake values (i.e. negative sales, and so on) were eliminated. Data of around 9,300 firms were used in the statistical analysis. They generated about 88,000 observations over 10 years. The following graph shows the trend in Interest coverage, Leverage and ROE (Leverage = right axis) in the analysed sample of firms in the period 2008-2017. Leverage and interest coverage have, mostly, opposite trends, starting 2011 these ratios ameliorated. From 2014 onward the improvement trend stands out in concomitance with a change of sign of the ROE (from negative to positive).

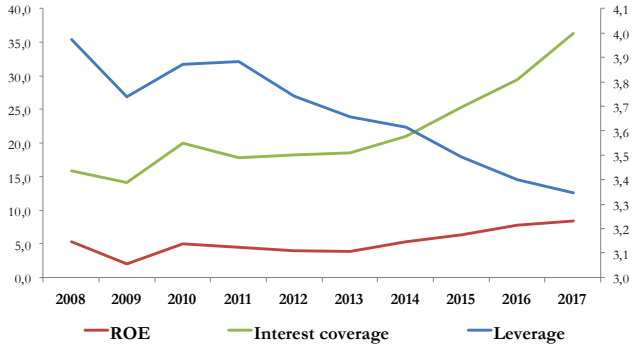


Figure 1: trend in Interest coverage, Leverage and ROE in the period 2008–2017.

4. Latent Growth Curve Models

Longitudinal studies are common in social sciences research. In these studies, individuals are observed at more than one point in time and interest is focused on the analysis of change or growth. Several statistical approaches are available for the analysis of change in longitudinal research: Autoregressive Models, Multilevel Models, Generalized Estimating Equation and Latent Growth Curve Models, among others

During the last thirty years, Latent Growth Curve Models (LGCMs) has become popular in the analysis of longitudinal and panel data (Meredith and Tisak, 1990; Singer and Willet, 2003; Bollen and Curran, 2006) for the study of individual change, and represent an effective method for examining interindividual differences in intra-individual change or growth. Latent Growth Curve Models under the Structural Equation Modeling framework adopt a latent variable approach and assume the existence of latent trajectories (i.e., underlying factors) for each individual, which are observed indirectly with the repeated measures (Bollen, 2002). All individuals are assumed to have

developmental curves of the same functional form and individual differences both in the initial status and in the growth rates are included into the model as latent variables: latent variable means for the intercept and slope factors describe the averages of initial status and growth rates, respectively; inter-individual differences in the growth curve parameters are modeled as the (co)variances of the intercept and slope factors. Given y_i a T × 1 vector of repeated observed measures for individual i at time points t = 1,2,...,T, the model can be expressed in matrix notation by (Bollen and Curran, 2006): a trajectory equation, expressed in terms of a confirmatory factor model, conditional to a vector of time-varying covariates, $\boldsymbol{w_i}$, in which the latent factors (η) represent the growth curve components (intercept and slopes)

$$\boldsymbol{y_i} = \boldsymbol{\Lambda_{\eta i}} + \boldsymbol{K_{wi}} + \boldsymbol{\varepsilon_i}$$

a structural model, to define the underlying latent growth factors in terms of means and individual deviations from the means, conditional to a vector of observed time-invariant predictors, $\boldsymbol{x_i}$

$$\boldsymbol{\eta_i} = \boldsymbol{\mu_\eta} + \boldsymbol{\Gamma x_i} + \boldsymbol{\zeta_i}$$

Here below the detailed contents of the matrices:

$$\underset{(T\times 1)}{\mathbf{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \quad \underset{(T\times m)}{\mathbf{\Lambda}} = \begin{bmatrix} 1 & 0 & 0 & \vdots \\ 1 & 1 & 1^2 & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T-1 & (T-1)^2 & \vdots \end{bmatrix} \quad \underset{(m\times 1)}{\boldsymbol{\eta}_i} = \begin{bmatrix} \eta_\alpha \\ \eta_\beta \\ \cdots \\ \eta_m \end{bmatrix} \quad \underset{(T\times 1)}{\mathbf{w}_i} = \begin{bmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{iT} \end{bmatrix} \quad \underset{(T\times 1)}{\boldsymbol{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}$$

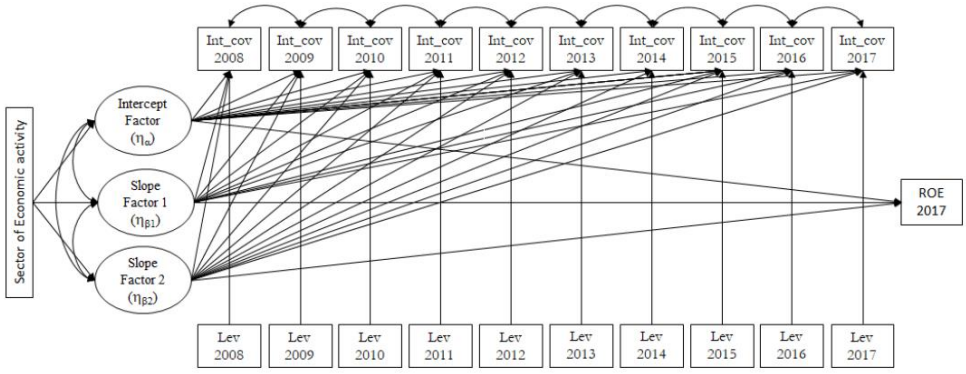
$$\underset{(m\times 1)}{\boldsymbol{\mu}_\eta} = \begin{bmatrix} \mu_\alpha \\ \mu_\beta \\ \cdots \\ \mu_m \end{bmatrix} \quad \underset{(m\times k)}{\boldsymbol{\Gamma}_i} = \begin{bmatrix} \gamma_{\mu_{\alpha 1}} & \gamma_{\mu_{\alpha 2}} & \cdots & \gamma_{\mu_{\alpha k}} \\ \gamma_{\mu_{\beta 1}} & \gamma_{\mu_{\beta 2}} & \cdots & \gamma_{\mu_{\beta k}} \end{bmatrix} \quad \underset{(k\times 1)}{\mathbf{x}_i} = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{bmatrix} \quad \underset{(m\times 1)}{\boldsymbol{\zeta}_i} = \begin{bmatrix} \zeta_{\mu_{\alpha i}} \\ \zeta_{\mu_{\beta i}} \\ \cdots \\ \zeta_{\mu_{mi}} \end{bmatrix}$$

$$\boldsymbol{\zeta}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$$

Where η_i is $m \times 1$ vector of growth factors, delta is $T \times m$ matrix of factor loadings for T time points, $\boldsymbol{w_i}$ is $T \times 1$ vector of time-varying covariates, **K** is *T×T* matrix of regression coefficients of the repeated measures of the time-varying covariates, ε_i is the *T×1* random vector of time-specific residuals, μη is m×1 vector of growth factor means, $\boldsymbol{x_i}$ is *K×1* vector of time-invariant covariates for the latent variables, gamma is $m \times k$ matrix of regression coefficients between the latent factors and the observed covariates; and ζ_i is $m \times 1$ vector of residuals, capturing individuals variation in growth factor means, a single distal outcome, indicated with u_i, the models can be extended as follows:

$$u_i = \tau_u + \beta_\alpha \eta_{i\alpha} + \beta_1 \eta_{i\beta 1} + \beta_2 \eta_{i\beta 2} + \upsilon_i$$

Here, the effects of the growth factors on the distal outcome $uuii$ are summarized by the corresponding regression coefficients $\beta\alpha$, $\beta 1$, $\beta 2$. The analysis includes various types of variables, each with a different role. Response variable is the Interest Coverage from 2008 to 2017, measured on a continuous scale; it represents a short-term risk indicator. Time-varying covariate is the Leverage which is the indicator of financial distress and it varies across firms and time. Time-invariant covariates are Sector of Economic Activity (Food and Tobacco, Textile and Leather, Wood, Publishing and Paper, Refining, Chemistry and Rubber, Metallurgy and Steel industry, Electric machines and Mechanical, Means of transport and Other industries and maintenance as reference category); they vary across firms but not in time. Distal outcome is the ROE 2017, which is an outcome of firms' performance that is predicted from the growth of Interest Coverage. Here following the path diagram of the LGCM with ten repeated measures of the outcome variable (Interest coverage: Int_cov), a time-varying covariate (Leverage: Lev), the time invariant covariates (Sector of economic activity) and a distal outcome (ROE 2017):



5. Result

Here below we show the results from the analysis performed by fitting a LGCM. A set of alternative unconditional LGCMs with correlated measurement errors between adjacent time point were first estimated (linear, quadratic, Piecewise linear with 2 knots and latent basis), in order to identify the more suitable functional form for the individual latent trajectories. Based on the model goodness of fit the quadratic form LGCM was preferred (RMSEA=0.028, CFI=0.980, TLI=0.976). The corresponding model parameters were estimated by using the Full Information Maximum Likelihood (FIML; Arbuckle, 1996) method with robust standard errors:

Parameters	Estimate	P-value
Intercept Mean (μ_α)	16.136	0.000
Intercept Variance ($\psi_{\alpha\alpha}$)	2906.039	0.000
Slope1 Mean ($\mu_{\beta 1}$)	-0.853	0.040
Slope1 Variance ($\psi_{\beta 1\,\beta 1}$)	401.514	0.000
Slope2 Mean ($\mu_{\beta 2}$)	0.300	0.000
Slope2 Variance ($\psi_{\beta 2\,\beta 2}$)	4.839	0.000
Correlation intercept vs slope1 ($\psi_{\alpha\,\beta 1}$)	-0.358	0.000
Correlation intercept vs slope2 ($\psi_{\alpha\,\beta 2}$)	0.197	0.000
Correlation slope1 vs slope2 ($\psi_{\beta 1\,\beta 2}$)	-0.906	0.000

Results show that the starting point of interest coverage is 16.136. The initial decrease of this indicator (-0.853) is followed by an increase (0.300). These growth parameters show a significant variability. Differences from baseline level in growth curve parameters by Sector of Economic Activity can be analyzed by sector economic activity:

EconomicActivity	Intercept(η_α)	Slope1 ($\eta_{\beta 1}$)	Slope2 ($\eta_{\beta 2}$)
Food and Tobacco	-1.035	1.048**	0.040
Textile and Leather	-5.295	0.662	0.015
Wood, Publishing and Paper	2.950	-0.258	-0.030
Refinig	1.054	1.870	0.116
Chemistry and Rubber	-2.256	0.782	0.127***
Metallurgy and Steel industry	-1.371	0.876**	0.037
Electric machines and Mechanical	7.655**	0.944*	0.137***
Means of transport	2.741	-0.131	0.062
Other industries and maintenance (reference)	27.387	-1.149	0.319

p < 0.10 (*); p < 0.05 (**); p < 0.01 (***)

The effect of Leverage on Interest coverage during the period:

Year	Estimate	P-value
LEV_08	-2.508	0.000
LEV_09	-2.843	0.000
LEV_10	-2.160	0.000
LEV_11	-2.427	0.000
LEV_12	-3.077	0.000
LEV_13	-3.722	0.000
LEV_14	-4.170	0.000
LEV_15	-4.509	0.000
LEV_16	-5.125	0.000
LEV_17	-5.323	0.000

Growth parameters of Interest Coverage helps to predict the firms' performance at the end of the observed period, measured by ROE 2017:

Coefficients	**Estimate**	**P-value**
Intercept (τ_u)	-3.520	0.381
Intercept factor ($\beta\beta_{\alpha\alpha}$)	0.064	0.000
Slope factor 1 ($\beta\beta_1$)	-1.269	0.022
Slope factor 2 ($\beta\beta_2$)	24.822	0.001

The model's goodness-of-fit was evaluated based on the most commonly used criteria (Bagozzi and Yi 1988):

	Value
Chi-square	1368.3
p-value	0.000
Degree of freedom	211
RMSEA (Root Mean Square Error of Approximation)	0.028
CFI (Comparative Fit Index)	0.922
TLI (TuckerLewis Index)	0.907
SRMR (Standardized Root Mean Square Residual)	0.039

These results show an adequate fit of the estimated model

6. Conclusions

The major results achieved by the analysis can be synthetized as follows: LGCM approach detects successfully a time trend in riskiness-distress risk. The relationship between interest coverage and leverage can represent the riskiness over the period well and it grows stronger over the period. The relationship between interest coverage and leverage is well fitted by a quadratic curve; it means that the estimated riskiness-distress risk time trend

is consistent with the general economic cycle. The level of profitability (ROE) is well predictable by the same quadratic function. The industries in which the Italian manufacturing is stronger (i.e. mechanical and chemical productions) are the ones for which the recovery in terms of riskiness indicators after the crisis is faster and stronger. From an economic point of view: it is relevant in the period the impact of double crisis (subprime and sovereign debt). This implies a W-shaped trend for quite all economic indicators. This is particularly true for firms' riskiness and the indicators representing it. The estimated curve parameters imply a first period of increasing riskiness and a second period characterized by a more and more decreasing riskiness. The second period can be partitioned, in turn, into two sub periods in which two different factors impact on riskiness. Right after the sovereign debt crisis (2011-2014) the growth of riskiness was stopped by the effects of ACE (*Aiuto alla Crescita*) provision, that strongly boosted the deleveraging process in the larger firms (Zeli, 2018). The coming of the global economic recovery in 2014-2015 further improved the riskiness indicators.

References

1. Altman, E., Marco, G., & Varetto, G. (1994). Corporate distress diagnosis: comparison using linear discriminant analysis and neural networks (the Italian experience). Journal of Banking & Finance., 18, 505–529.
2. Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G.A. Marcoulides & R.E. Schumacker [Eds.] Advanced structural equation modeling. Mahwah, New Jersey: Lawrence Erlbaum Associates.
3. Bagozzi, R.P., Yi, and Y. (1988). On the evaluation of structural equation models. Academy of Marketing Science, 16(1): 76–94.
4. Bellovary, J.L., Giacomino, D.E. & Akers, M.D. (2007). A review of Bankruptcy Prediction Studies: 1930 to Present. Journal of Financial Education, 33, 1–42.
5. Bollen, K. A. (2002). Latent variables in psychology and the social sciences. Annual Review of Psychology, 53, 605–634.
6. Bollen, K. A., and Curran, P. J. (2006). Latent Curve Models. New York: Wiley.
7. Bottazzi, G., Grazzi, M., Secchi, A., & Tamagni, F. (2011). Financial and economic determinants of firm default. Journal of Evolutionary Economics., 21, 373–406.
8. Di Clemente, A. (2008). Rischio d'insolvenza e ciclo economico: un'analisi di macro stress-testing per le imprese non finanziarie italiane In: Proceedings of the 49th Annual scientific conference of Società Italiana degli Economisti – Perugia University - 24th October.

9. Meredith, W. M. and Tisak, J. (1990). Latent curve analysis. Psychometrika, 55(1): 107–122.
10. Singer, J. D. and Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press.
11. Giroud, X. & Mueller, H. M. (2015). Firm Leverage and Unemployment during the Great Recession. NBER W.P. n. 21076.
12. Graham, J.R., Hazarika, S. & Narasimha, K. (2011). Financial Distress in the Great Depression. Financial management, 40(4), 821-844.
13. Kudlyak, M. & Sánchez, J. M. (2017). Revisiting the behavior of small and large firms during the 2008 financial crisis. Journal of Economic Dynamics and Control, 77, 48-69.
14. Zeli, A., & Mariani, P. (2009). Productivity and profitability analysis of large Italian companies: 1998–2002. International Review of Economics, 56, 175–188.
15. Zeli, A. (2014). The Financial Distress Indicators Trend in Italy. An Analysis of Medium–size Enterprises. Eurasian Economic Review, 4(2), 199–221.
16. Zeli, A. (2018). The impact of ACE on investment: the Italian case. Economia Politica, 35(3), 741–762.

Deep learning the mca dot sign of acute ischemic stroke on non-contrast ct images

Jia You, Philip L.H. Yu
Department of Statistics and Actuarial Science, The University of Hong Kong

Abstract

The hyperdense middle cerebral artery (MCA) dot sign has been reported as an important factor in the diagnosis of acute ischemic stroke due to large vessel occlusion. Interpreting the initial CT brain scan in these patients requires high level of expertise and has high inter-observer variability. An automated computerized interpretation of the urgent CT brain image, with an emphasis to pick up early signs of ischemic stroke will facilitate early patient diagnosis, triage, and shorten the door-to-revascularization time for these group of patients. In this paper, we present an automated detection method of segmenting the MCA dot sign on non-contrast CT brain image scans based on powerful deep learning technique.

Keywords

Deep learning, Segmentation; Medical imaging; Hyperdense middle cerebral artery dot sign; Acute stroke.

1. Introduction

Acute ischemic stroke (AIS) has becoming a leading cause of morbidity and mortality worldwide and recent advances in endovascular thrombectomy (EVT) for treatment of AIS caused by large vessel occlusion (LVO) have been widely accepted around the world (Powers et al., 2018; Malhotra & Liebeskind, 2015). The hyperdense middle cerebral artery (MCA) dot sign has been reported as an important factor in the diagnosis of acute ischemia, especially in LVO cases (Lim et al., 2018). Fast diagnosis and localization of MCA sign can largely save patients' rescue time, thus lower the probability of severe effect. However, it is fairly challenge to detect MCA sign due to the subtlety of the pathological intensity changes and low signal to noise ratio (Fig. 1). Available data on large vessel occlusion stroke is based on western populations and the respective incidence in Asian countries is largely unknown. So far, this study is the first application of deep learning with specified interest to the hyperdense MCA sign.

We adopted deep learning model as a feature extractor in this study, as well. Recent years saw the availability of large amounts of annotated training sets and the accessibility of affordable parallel computing resources via Graphics Processing Units (or GPUs) have made it feasible to train deep neural

networks with huge amount of data and parameters, which have revolutionized the artificial intelligence in achieving the outstanding results on many challenging tasks. The medical imaging community has taken notice of these pivotal developments. The deep learning has been widely applied to medical image analysis, demonstrating the state-of-the-art performances on many medical image analysis tasks, including classification, detection and segmentation.

2. Methodology

The Hong Kong Hospital Authority's Clinical Management System (CMS) has well-established records of all patients admitted to the public hospitals for all types of acute ischemic stroke in 2016. The study population was stratified using disproportionate random sampling methods, and the patients' selection criterion is announced in another published paper (Tsang et al., 2019). Total 150 patients were sampled from the ischemic stroke database in CMS. The data includes both CT images and some on-set clinical information, e.g.: side of weakness. The samples were then randomly split into 120 for model training and 30 for validation.

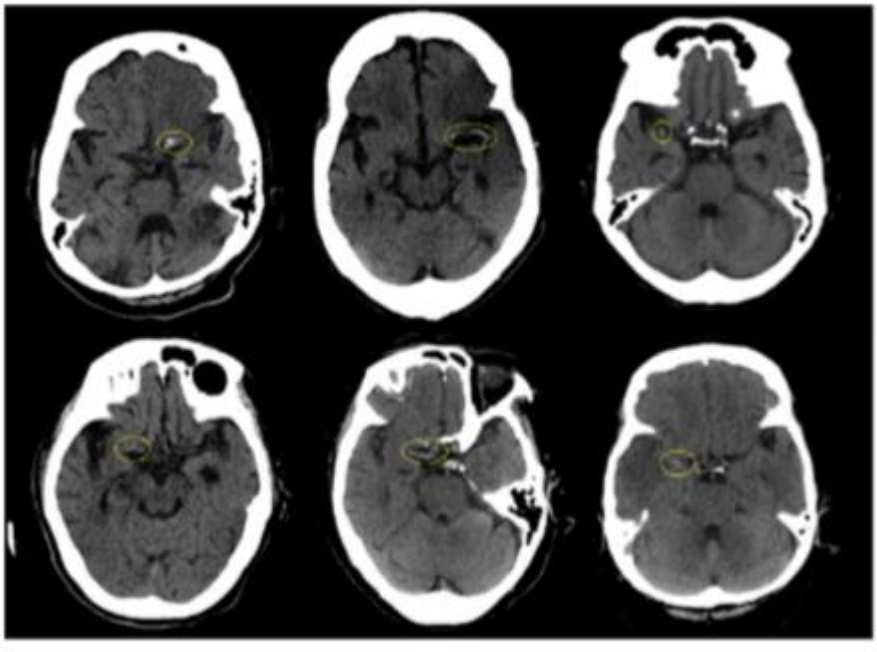


Fig. 1. Sample MCA Dot signs

The MCA ground truth was independently evaluated by two cerebrovascular disease specialists. Any discrepancies were resolved by consensus. The segment labels were manually drawn through software FSL

(Smith et al., 2004; Woolrich et al., 2009). Besides imaging data, it also involves the structural data such as patients' side of limb weakness at A & E admission.

The CT images have similar quality, spatial resolution and field- of-view. The in-plane resolution is 0.426*0.426 mm. The slice thickness is 5.0 mm for all cases, and the number of slices is around 26 to 32. Each axial slice has identical resolutions of 512*512.

The existence of hyperdense MCA dot signs can be directly visualized as thromboembolic material within the lumen, which is largely course in a plane perpendicular to the transverse plane of imaging (Fig. 2). Thus, the recognition of the MCA dot signs can be localized within a specified area of the scans, and extraction the specified regions of interest will largely help eliminate useless information. We found all MCA dot signs are localized between the 4th and 10th slices after registration to a template. For both training and testing phase, CT scans were pre-processed using the fully automatic pre-processing pipeline through FSL and Nibabel library under python 3.5.

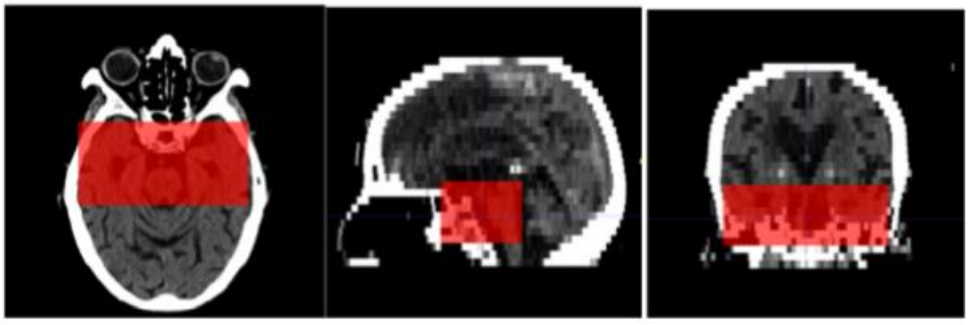


Fig. 2. Regions of Interest for MCA Signs

As shown in pre-processing flow chart (Fig. 3), the first step is brain extraction to strip the skulls. In the second step, all CT scans are rotated and translated through a rigid-body 2D registration procedure in order to make sure all brains within images are horizontally symmetric. All the MCA dot signs have H.U. index between 35 and 60; thus, a threshold of 20 to 80 is utilized in order to eliminate the irrelevant image information and histogram equalization is applied to increase the contrast. To better specify the region where MCA dot sign, we localize a bounding box to subtract the region of interest as Fig. 2. The coloured bounding box has size of 128*128; while two colours indicating left and right hemispheres. The location of MCA within different hemispheres would cause corresponding side of weakness for patients. Given clinical information for different side of limb weakness, we can better localize the infracted hemisphere, coloured in blue and yellow. After extraction of potential infarcted hemisphere, histogram equalization was applied to ROI images to enhance the contrast of MCA dot signs.

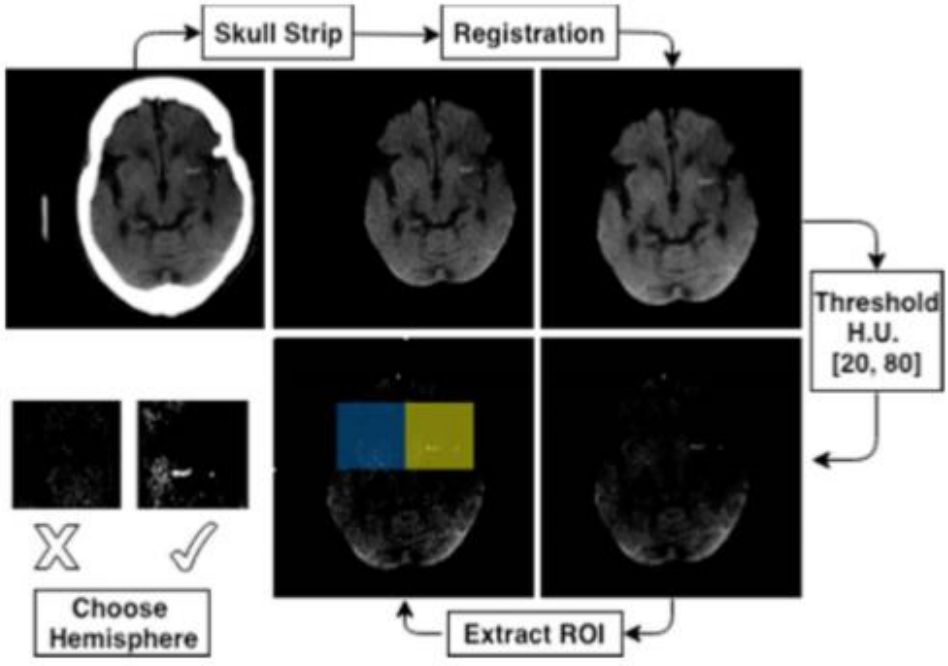


Fig. 3. Pre-processing Flowchart

3. Experiments

The proposed architecture belongs to the category of fully convolutional networks (FCN) (Long & Darrell, 2015) that extends the convolution process across the entire image and predicts the segmentation mask as a whole. This architecture consists of an encoding part and a decoding part, shown as Fig. 4. The encoding part resembles a traditional convolutional neural networks (CNN) (Krizhevsky et al., 2012) that extract a hierarchy of image features from low to high complexity. The decoding part then transforms the features and reconstructs the segmentation label map from coarse to fine resolution. The model contains skip connections, which is pretty similar to the U-net (Ronneberger, et al., 2015), which is one of the most popular architecture for biomedical imaging segmentation tasks. The long-range connections across the encoding part and the decoding part enable high resolution features from the encoding part can be used as extra inputs for the convolutional layers in the decoding part.

Less than half patients, 74 of out 150, in our database has MCA dot sign, and the slice containing ground truth is quite imbalance to empty slices. Due to the limited sample size, we applied data augmentation with randomly zoom in, shift, rotation and horizontal flip of the input images as the final pre-processing step. Moreover, the transfer learning with pre-trained weights could also help during training. Therefore, our encoding structures are exact the same as VGG16 and initial weights are pre-trained on ImageNet dataset.

Our deep learning model used Adam optimizer with 1e-5 initial learning rate and trained on 200 epochs with Tesla K80 GPU.

To evaluate the performance of the deep learning architecture in segmenting the hyperdense MCA dot signs, the dice similarity coefficients (DSC) is utilized as the evaluation metric for goodness of fit. The DSC is defined as

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B represent the regions of all voxels of ground truth and segmentation respectively.

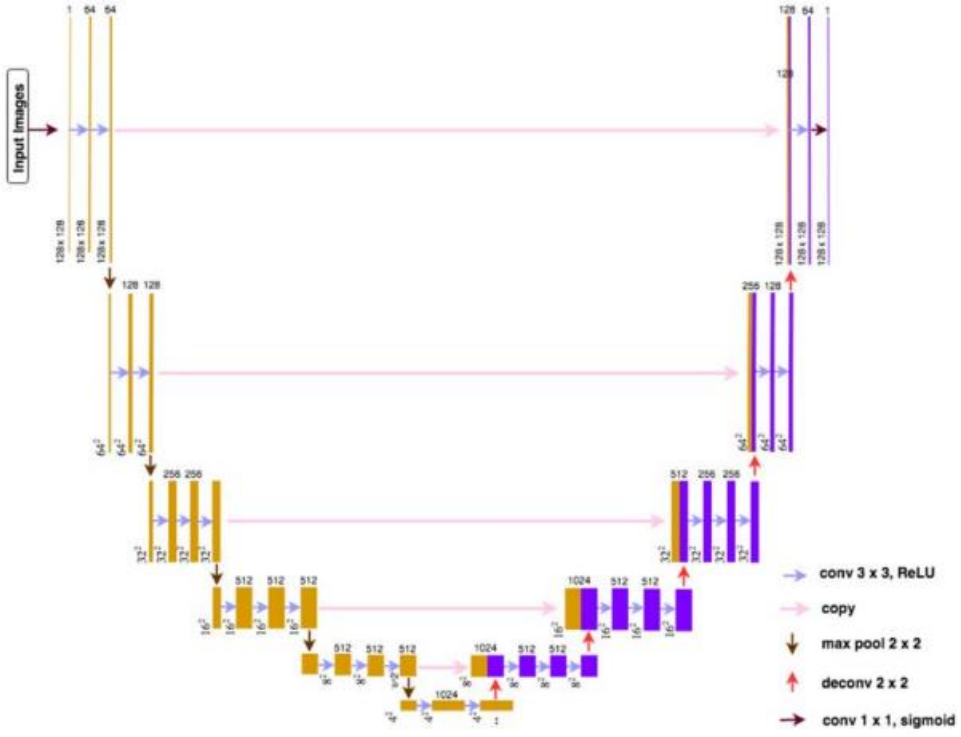


Fig. 4. Deep Learning Architecture

4. Result

Within 150 CT scans, 74 patients were diagnosed to have hyperdense MCA dot sign and the rest were empty. Among the 74 positive MCA sign subjects, 63 were within training and the rest 11 were in the testing. As shown in Table 1, patients without side of weakness do not have MCA dot sign; thus the model involved side of weakness and a filter to select the potential subjects might have MCA. We subtract cropped bounding boxes within specified regions of interest for each set of CT scans. If patient suffer side of weakness, we extract his or her ROI within corresponding left or right hemisphere; if patient suffer both side of weakness, we use both ROIs in two hemispheres; otherwise,

patient without any side of weakness was not considered in model construction. Total Among 120 training subjects, 95 patients (79.16%) were recorded as side of weakness, and 3 patients suffer both sides of weakness. Thus, total 588 slices were fed into model training, and only 84 slices containing MCA sign ground truth.

	Side of Weakness	w/o Side of Weakness	
MCA	63	0	63
w/o MCA	32	25	57
	95	25	110

Table 1. MCA vs Side of Weakness in Training Data

Our model achieves dice similarity coefficient 0.686 on the testing data. The result is satisfied since the MCA sign is extremely small and quite hard to gain relative high DSC.

5. Discussion & Conclusion

MCA dot signs are extreme small in size and quite low signal to noise ratio, the essential step in MCA segmentation task is the localization of specified regions of interest, which would largely ignore that irrelevant information.

The DSC is not high enough mainly from two aspects. The first is due to its size is pretty small that even a subtle miss would cause large effect on the prediction. Total positive ground truth label is less than 0.1% of negative label. Another is the false positive predictions accounts for large inaccuracy that largely due to proximity of bone and the similarity to normal age-related vascular calcification. However, our prediction has high sensitivity that able to right predict all MCA dot signs in our testing case.

Further post-processing step to distinguish MCA dot signs and false positive predictions will largely help enhance the model's performance. Adequate data with more positive labelled ground truth will enable model to learn more and become more robust, as well.

Overall, we present an automated method for identifying the hyperdense MCA dot sign on Noncontrast CT scans. The study can be further reinforced with additional data input.

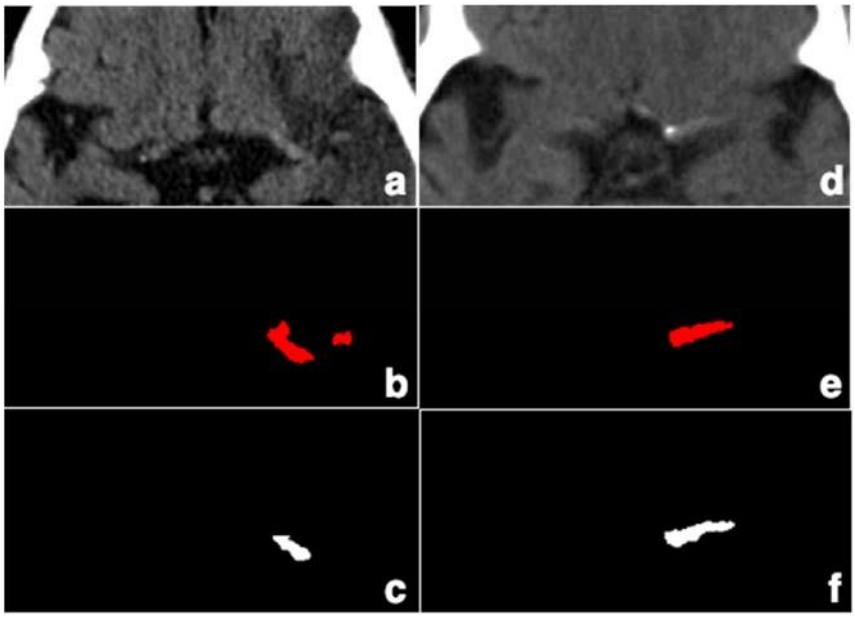


Fig. 4. **a** & **d**: Raw ROIs; **b** & **c**: Ground Truth; **c** & **f**: Predictions

References

1. Lim, J., Magarik, J. A., & Froehler, M. T. (2018). The CT Defined Hyperdense Arterial Sign as a Marker for Acute Intracerebral Large Vessel Occlusion. Journal of Neuroimaging, 28(2), 212-216.
2. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241). Springer, Cham.
3. Powers, W. J., Derdeyn, C. P., Biller, J., Coffey, C. S., Hoh, B. L., Jauch, E. C., ... & Meschia, J. F. (2015). 2015 American Heart Association/American Stroke Association focused update of the 2013 guidelines for the early management of patients with acute ischemic stroke regarding endovascular treatment: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 46(10), 3020-3035.
4. Malhotra, K., & Liebeskind, D. S. (2015). Imaging in endovascular stroke trials. Journal of Neuroimaging, 25(4), 517-527.
5. Tsang, A. C. O., You, J., Li, L. F., Tsang, F. C. P., Woo, P. P. S., Tsui, E. L. H., ... K., G. K. (2019). Burden of large vessel occlusion stroke and the service gap of thrombectomy: A population-based study using a territory-wide public hospital system registry. To appear in International Journal of Stroke.

6. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 10971105).
7. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).
8. Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... & Niazy, R. K. (2004). Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage, 23, S208-S219.
9. Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... & Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. Neuroimage, 45(1), S173-S186.

Computing employment multipliers in the context of Malaysian economy

Siti Nurliza Samsudin, Akmalia Hanifah
Core Team Malaysian Bureau of Labour Statistics, Department of Statistics Malaysia

Abstract

The aim of this article is to explain the computing of employment multipliers in the context of Malaysian Economy by industry. The data used is sourced from the Input-Output (I-O) Tables, Economic Census Reports, Labour Force Survey Report and the Labour Productivity Report released by Department of Statistics Malaysia in 2015. Computing the employer multipliers uses the Leontief Inverse Model from the I-O Tables, as well as w, physical labour coefficient vector which is computed using the ratio of employment to output. Furthermore, the article will study the figures obtained to interpret the number of new jobs created to meet increased final demand for new output. Lastly, the article seeks to compare the behaviours of industries based on the multipliers obtained using K-means clustering.

Keywords

Employment Multiplier; Input-Output Table; Labour Productivity; Labour Demand; Job Creation

1. Introduction

To compute employment multiplier in the context of the Malaysian Economy, the following data were used:

(a) Employment
(b) Value-added
(c) Inverse Matrix of Domestic Production

While (b) and (c) can be easily obtained from the Input-Output Tables published by DOSM, employment data is difficult to compute.

According to the Labour Force Survey (LFS), DOSM, an employed person is defined as a person, who at any time during the reference week worked at least one hour for pay, profit or family gain either as an employer, employee, own-account worker or unpaid family worker. An employed person can either be in the formal sector or the informal sector, where the formal sector is defined as activities carried out by registered organizations in Malaysia and the informal sector is defined as activities carried out by nonregistered individuals and organizations. These definitions can also be referred from the

Employment and Salaries & Wages Statistics Report (2016), which is based on the Economic Census (EC) 2015.

The challenge in computing employment data is due to the multiple sources available. Particularly at DOSM, two approaches are used, which are the household approach (LFS) and the establishment approach (Economic Census, Monthly Manufacturing Statistics (MMS) and Quarterly Services Statistics (QSS)) -- all of which employment data is collected but is referred to different coverage of the labour force. The household approach covers the entire labour force, including those employed in the informal sector whereas the establishment approach covers only registered organizations and hence by definition, those employed only in the formal sector. While the LFS covers more, its data is based on a sample of which the sample stratification does not include industry. Hence there is high possibility of error when using data at a very granular level. For industries of which there is more formal employment such as Finance and Insurance, data from the Census can be more reliable.

In order to compute employment multiplier, which reflects creation of jobs based on the industrial nature, employment data needs to be adjusted and estimated. This study will compute employment multiplier based on employment data estimated using combination of LFS 2015 and EC 2015, by considering the structure of formal-informal employment in the sector.

The employment multiplier is important as it is a way to measure number of jobs created in economy, resulted from an increase of the output. In addition, employment multiplier provides opportunity to the country to reduce widespread unemployment and to improve people's wellbeing (Ntibanyurwa, 2008). There are two types of employment multiplier, Type 1 and Type 2 multiplier. Type 1 multiplier captures the indirect effects, meanwhile Type 2 multiplier captures both indirect and induced effects. According to Cetnarski (2011), direct jobs are related to the specific industry, indirect jobs are those that support the industry while induced jobs are those that are a result of direct or indirect employee's spending money in the community. Generally, industries with a higher multiplier are more desirable (Cetnarski, 2011).

2. Methodology

There are a few of economic models to estimate employment multiplier, one of them is using Leontief inverse model. According to Siti Rahmah and Nurul Naqiah (2015), the basic equation of Leontief inverse which also known as the basic multiplier model in input-output is written as follows:

$Q_t = (I-A)^{-1} * f_t$

Q = Vector of domestic product

I = Identity matrix

A = Domestic input coefficient matrix
f = Vector of final demand
t = 2015

In this study, we want to examine the employment multiplier. Thus, the physical labour input coefficient, L_t is used instead of monetary labour input coefficient. Mathematically, L_t is $e^{*}x^{-1}$ which is computed from a vector of employment divided by value added. Hence, the final equation to compute employment multiplier written as follows:

$Qt = Lt^{*}(I-A)\text{-}1^{*}\ ft$

L_t = Vector of employment divided by value-added

t = 2015

As mentioned, employment data here is obtained and estimated from the EC and LFS, reference year 2015. To make the estimates, the industries are first grouped according to formality of employment; if mainly formal, we use the EC and if mainly informal, we use the LFS. The estimates are also compared to employment data in the Labour Productivity Report, however not directly due to the difference in industrial grouping (2-digit level).

The inverse matrix and value-added data are obtained by subtracting Total Input from Total Output from the Domestic Use Table at Basic Prices -- both obtained from the Input-Output Tables, 2015. When all three variables had been obtained, the industrial code used for the employment data (5-digit) needs to be mapped to the I-O Tables (aggregated 5-digit). There are 124 aggregations in total. Then finally, the employment multiplier is computed using the formula.

3. Result

Table 1: Employment Multipliers Calculated Based on Employment Data Estimated from Labour Force Survey 2015 and Economic Census 2015, by industry

Rank	Industrial Code	Industry	Employment Multiplier
1	122	Non-Profit Institutions Serving Households	0.081
2	5	Rubber	0.081
3	124	Other Private Services	0.079
4	116	Business Services	0.072
5	1	Paddy	0.067
6	23	Bakery Products	0.049
7	33	Wearing Apparel	0.048

8	95	Food and Beverage	0.040
9	88	Sewerage, Waste Management and Remediation Activities	0.037
10	96	Land Transport	0.033
115	30	Preparation, Spinning and Weaving of Textiles	0.006
116	121	Other Public Administration	0.006
117	114	Scientific Research and Development	0.006
118	7	Flower Plants	0.005
119	45	Basic Chemicals	0.005
120	44	Coke and Refined Petroleum Products	0.004
121	29	Tobacco Products	0.004
122	112	Ownership of Dwellings	0.003
123	101	Highway Operation Services, Bridge and Tunnel	0.002
124	13	Crude Oil and Natural Gas	0.002

Table 2: A Snapshot of Groups of Employment Multiplier based on K-Means Clustering, k=4

Group 1	Group 2	Group 3	Group 4
Non-Profit Institutions Serving Households	Finishing of Textiles	Bakery Products	Residential Buildings
Rubber	Rental and Leasing	Wearing Apparel	Specialised Construction Activities
Other Private Services	Steam Generators	Food and Beverage	Wooden Containers and Other Wood Products
Business Services	Footwear	Land Transport	Furniture
Paddy	Arts, Entertainment and Recreation	Sewerage, Waste Management and Remediation Activities	Processing and Preserving of Seafood
	Wiring Devices, Electric Lighting Equipment and Other Electrical		Postal and Courier Activities
	Publishing Activities		Repair & Installation of Machinery and Equipment
	Domestic Appliances		Sawmilling and Planning of Wood
	Soft Drinks, Mineral Waters and Other Bottled Waters		Other Livestock

	Pharmaceuticals, Medicinal Chemical and Botanical Products		Health
	Other Manufacturing		Processing and Preserving of Fruits & Vegetables
	Basic Precious and Other NonFerrous Metals		Other Mining and Quarrying
	Services Incidental to Water and Air Transportation		Accommodation
	Computer and Information Services		Oil Palm
	Optical Instruments, Photographic Equipment, Magnetic and Optical Media		Reproduction of Recorded Media
	Fertilizers and Nitrogen Compounds		Veneer Sheets and Wood-based Panels
	Motion Picture, Programming and Broadcasting Activities		Vegetable & Animal Oils and Fats
	Fruits		Prepared Animal Feeds
	Dairy Products		Processing and Preserving of Meat
	Vegetables		Activities Auxiliary to Financial Service and Insurance/ Takaful
	Other Textiles		Professional
	Air Transport		Structural Metal Products, Tanks, Reservoirs and
	Fishing and Aquaculture		Non-Residential Buildings

Table 3: Number of Industries and Estimated Centre in Each K-Means Cluster, k=4

	Number of Cases in each Cluster	Centre
Cluster 1	5.000	.075878
2	73.000	.009657
3	5.000	.041341
4	41.000	.020329
Valid	124.000	
Missing	.000	

4. Discussion and Conclusion

Table 1 shows the industries with the ten highest and ten lowest employment multipliers.

The largest job creation can be seen in Non-Profit Institutions Serving Households, of which data is sourced from the LFS, whereas the lowest is in Crude Oil and Natural Gas sourced from the Economic Census. According to Anushree, Avantika and Rajesh (2015), larger employment multipliers refer to those labour-intensive, and lower ones refer to capital-intensive industries.

In Malaysia, Agriculture, Manufacturing and Services sectors are generally labour-intensive, whereas Mining & Quarrying and Construction are mostly capital-intensive.

Table 2 shows a snapshot of results of K-means clustering of the employment multipliers. K-means clustering here uses k=4, which is chosen based on a dendogram obtained from hierarchical clustering. This clustering is an attempt to group the industries based on the employment multipliers. Based on Table 3, it can be observed that Group 1 is made up of industries with the highest figures and Group 2 is with those with the lowest figures. As Group 2 is the largest group, it may be concluded that the majority of the industries have low employment multiplier; thereby suggesting that jobs created are mostly in certain industries in Malaysia.

One of the limitations of this study is that the grouping of industries based on the structure of formalinformal employment is largely based on the typical perception of the industry in Malaysia. Informal employment, by its nature, is difficult to measure; and even more so at a granular level of industry. However, in understanding job creation in for formulation of economic policies, informal employment cannot be ignored so as to not leave anyone behind.

To enhance this study, employment multipliers can be conducted by estimating further informal and formal employment at each industry.

Furthermore, direct and indirect impacts may be calculated if one were to analyse the I-O tables at a more segregated level.

References

1. Anushree, S., Avantika, P. and Rajesh, J. (2015). Employment Dimension of Infrastructure Investment - State Level Input-Output Analysis. Working Paper No. 168. Employment Policy Department. International Labour Office. Geneva.
2. Bivens, J. Updated Employment Multipliers for the U.S. Economy. (2003). Working Paper No. 268. Economic Policy Institute. Washington, DC.
3. Cetnarski, E. (2011). The Employment Multiplier: An Important Tool for Promoting the Burgeoning
4. Green Economy. Presidio Graduate School's MBA program. https://www.triplepundit.com/2011/05/employment-multiplier-green-economy/
5. Department of Statistics Malaysia. (2016). Employment and Salaries & Wages Statistics Report 2015. Putrajaya, Malaysia.
6. Department of Statistics Malaysia. (2016). Input-Output Tables Malaysia 2015. Putrajaya, Malaysia.
7. Department of Statistics Malaysia. (2018). Labour Productivity Third Quarter 2018. Putrajaya, Malaysia.
8. Department of Statistics Malaysia. (2016). Labour Force Survey Report 2015. Putrajaya, Malaysia.
9. Dyrstad, E. H. (2014). Local Employment Multiplier in Norway: A Comparative Study of Norway, Sweden and the United States. Department of Economics. University of Olso.
10. Hussain, A. B. (2011). Output, Income and Employment Multipliers in Malaysian Economy: InputOutput Approach. International Business Research. Vol. 4. No. 1, pp. 208-223.
11. Ntibanyurwa, A. The Income and Employment Multiplier Effects of Tourism: The Case of Rwanda. (2008). Department of Economics. University of the Western Cape. pages 71-97.
12. Siew, H. Y., Wooi, L. O. and Koon, P. (2015). Income and Employment Multiplier Effects of the Malaysian Higher Education Sector. The Journal of Applied Economic Research. National Council of Applied Economic Research, vol. 9(1), pages 61-91.
13. Siti Rahmah, S. O. and Nurul Naqiah, M. (2015). Foreign Employment Multiplier in Malaysia, An InputOutput Analysis. Department of Statistics Malaysia.

A multi-factor modelling for retail demand forecasting: An empirical analysis of restaurant visitors prediction

Yutaka Kuroki[1]; Takayuki Shiohama[2]
[1] Graduate School of Engineering, Tokyo University of Science
[2] Department of Information and Computer Technology, Tokyo University of Science

Abstract
Analyzing cross-sectional and time-series retail sales data is important for multi store retail managements, especially in service related and retail businesses. This paper presents a use of factor model for numbers of customers forecasting in retail business and tests the validity of the proposed model. The factors are constructed by means of fundamental factors which are common tools for analyzing asset pricing models in financial market analysis. Data analysis using Japanese restaurants data are illustrated and showed that the effectiveness of the multi-factor modeling with high forecasting performances.

Keywords
Marketing; factor model; panel data econometrics; structural time series analysis.

1. Introduction

There has been an enormous growth in needs for big-data analytics in marketing science. Point-of-Sales (POS) data can be helpful to provide accurate demand forecast of a retail shop and be used to analyze consumer buying behavior. Big data analysis makes one-to-one marketing possible, which improves management effectiveness and accurate decision making in their supply chain. Forecasting demand in multi store sales is especially important for effective managements such as franchise chains. Since demand not for a single store, but for whole stores is dominated by calendar effects, which can be considered as an undiversifiable risk called a “systematic risk”. On the other hand, demand forecasting for a single store is not enough explained by such a common effect. We need to model and manage an “idiosyncratic risk” which arise in single store retail businesses by using appropriate statistical approaches.

In this study, we propose the factor models for number of customers of restaurants in Japan. We use seasonal and calendar effects as dominant factors, other factors are also proposed using the similar idea of analyzing the Capital Asset Pricing Model (CAPM) in financial econometrics. These factors include market, size, and volatility factors, which are considered as anomalies of the dynamics of the cross-sectional restaurant visitors time series. The

Capital Asset Pricing Model (CAPM, Sharpe; 1964 and Lintner; 1965) is the most broadly applied asset pricing model in finance and among researchers and practitioners. CAPM describes the relationship between systematic market-portfolio risk and expected excess return of assets. The deficiency of the use of CAPM models including the validity of its assumptions can be found in, for example, Bank (1981), Basu (1983), Bhandari (1988) and Fama & French (1995).

The market risk premium is the difference between the expected return of the market and the risk-free rate. The Fama-French three-factor-model (Fama & French, 1996) expands CAPM by adding size risk factor and value risk factor to the market risk factor. The size risk called "*SMB* (Small Minus Big)" measures excess return of small-cap companies over big-cap companies, and the value risk called "*HML* (High Minus Low)" measures excess return of high book-to-market ratio (value companies) over companies with a low book-to-market ratio (growth companies). Then the three-factor models are defined as

$$r_i - r_f = \alpha_i + \beta_{i1}(r_m - r_f) + \beta_{i2}(SMB) + \beta_{i3}(HML) + \varepsilon_i,$$

where r_f is risk free rate, r_i is the expected return of i-th stock, β are factor coefficients and $(r_m - r_f)$ is the market risk premium. The size factor, SMB is the difference between average return on the Small-firm portfolios and the average return on the Big-firm portfolios, that's why the factor can be an alternative variable for latent capitalization risk premium. As well as SMB, HML is the difference between average return on the High value portfolios and Low value portfolios.

We investigate similar characteristics in retail demand time-series modelling by introducing fundamental factors constructed from the store-specific information. In retail marketing, it is well-known that there are apparent seasonal effects in daily demand time series: weekly and yearly cycles and holiday effects. We consider this seasonal pattern as a market portfolio, since weekly and yearly cycles and holiday effects seem to be "systematic risk" that appears overall demand fluctuation in some retail business. Similarly, SMB like factor is constructed from the portfolio of stores that has large or small number of customers. From the viewpoint of prediction, many time series models explicitly including these effects have been suggested (Harvey & Shephard 1993, Hyndman et al. 2002, Taylor & Letham 2018), but there has not discussed about the characteristics of the retail demand return and risk structure anymore.

In this study, we investigate the risk and return structure of retail stores using number of customers in restaurants in Japan. We introduce a factor model for restaurants demand in a similar way of traditional financial factor models. First, we show that daily demand for whole restaurants is almost dominated by calendar effect like market portfolio. Second, we construct a factor derived from sizes of restaurants, and estimate factor-model by The

Generalized Method of Moment (GMM). GMM has become an important estimation procedure in applied economics and finance since Hansen (1982) introduce. Finally, we confirm the validity of our model and proposed fundamental factors in retail demand series.

The rest of paper is organized as follows. Section 2 describes data and our proposed model. The proposed fundamental factors in restaurant visitor data are also introduced in Section 2. Section 3 shows estimation results of the multiple regression models together with the results of the tests of the model assumptions. Section 4 provides summary and discussions of our results.

2. Data and Fundamental Factors in Retail Demand

The number of customers for Japanese restaurants were recorded by using AirREGI systems of Recruit Holdings, where AirREGI provides a free POS cash-register service. The data were available from Kaggle Recruit restaurant visitor forecasting competitions[1]. The data consists of number of visitors for 795 restaurants in Japan from the period 1st January 2016 to 22nd April 2017. Area covered all major cities in Japan, including Sapporo, Tokyo, Osaka, Fukuoka, and so on. The genres or styles of the restaurants are divided into 14 categories, e.g., Japanese food, Italian/French, and Café/Sweets, and so on. Figure 1 shows the mean number of daily customers of whole restaurants in the observed period. Seasonal fluctuations along with large gaps around the year end and beginning are apparent. The time-series structures of the mean number of the customers are highly correlated and annual trend with weekly seasonal patterns is also observed. The data includes the records for closed days, which are indicated by 0 records. The patterns of the frequency of 0 records are random and depends on situations for each restaurant. In this study, we propose the factor models for the prediction of the number of customers on each restaurant, we need to pay attention for 0 records of the data.

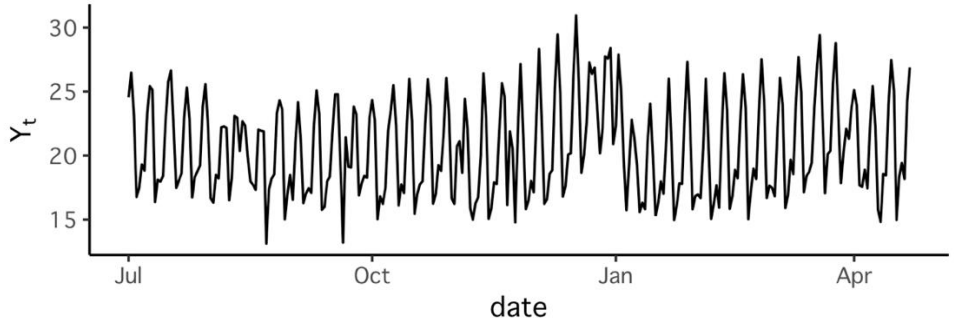


Figure 1. Time series plots for the mean number of daily customers of whole restaurants

[1] https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting

Let Y_{it} denote the number of customers for the i-th restaurant at time t. The first step for data screening is to impute 0 records with time series forecast of the following regression model:

$$\tilde{Y}_{i,t} = \begin{cases} Y_{i,t} & \text{if } Y_{i,t} \neq 0, \\ \hat{Y}_{i,t} & \text{if } Y_{i,t} = 0, \end{cases}$$

where

$$\tilde{Y}_{i,t} = \hat{\beta}_{i,F} F + \sum_{w=1}^{7} \hat{\beta}_{i,w} D_{w,t} + \hat{\beta}_{i,H} H_t.$$

Where F is m-dimensional factor vector explained later, $D_{w,t}$ is dummy variable for each day of week, H_t is a dummy variable for holidays and $\beta_{i,F}, \beta_{i,w}, \beta_{i,H}$ are the OLS estimates of the regression coefficients.

Using these complete panel data Y, the following multiple factor models are defined as follows,

$$y_{i,t} = \sum_{w=1}^{7} \beta_{i,w} D_{w,t} + \beta_{i,H} H_t + \sum_{f=1}^{m} \beta_{i,f} F_{f,t} + \varepsilon_{i,t}, \quad y_{i,t} = \frac{y_{i,t} - \overline{y_i}}{\sqrt{\left(y_{i,t} - \overline{y_{i,t}}\right)^2}}.$$

The fundamental retail demand factors we use in this study is Market (*MKT*), Small minus Big (*SMB*), and Safe minus Risky (*SMR*) factors. Similar to the factor models used for asset returns modeling, we estimated *SMB* and *SMR* by calculating the difference of returns between two portfolios based on corresponding features of restaurants. To construct big restaurants portfolio and small restaurants portfolio, we used mean customer counts for each restaurant and classified each restaurant into "big" and "small" categories. Then the *SMB* (Small minus Big) factor can be obtained by subtracting these normalized small and big portfolios. Similarly, *SMR* (Safe minus Risky) factor can be obtained as follows. We calculate the coefficient of variation for each store, that is $CV_i = \frac{sd(y_{i,t})}{avg(y_{i,t})}$, then we classified each restaurant into "Safe" and "Risky "categories, whose CV_i falls into 1st and 3rd quantiles of the samples, respectively. The *MKT* factor is estimated by removing seasonality from $Y_{I,t}$ with SARIMA model, since *MKT* factor should be constructed to be uncorrelated with seasonal patterns of the observed series.

Figure 3 shows constructed three factors. According to these plots, we can see that the *MKT* factors have no apparent seasonal patterns whereas this factor explains some sort of overall trend of the number of customers. The *SMB* factor becomes large around the year end, which indicates the large profitable opportunities increases for larger stores. On the other hand, the *SMR* factor becomes small around year end, which indicates that stores have

larger exposures on *SMR* factors tends to decrease visitors around the year end.

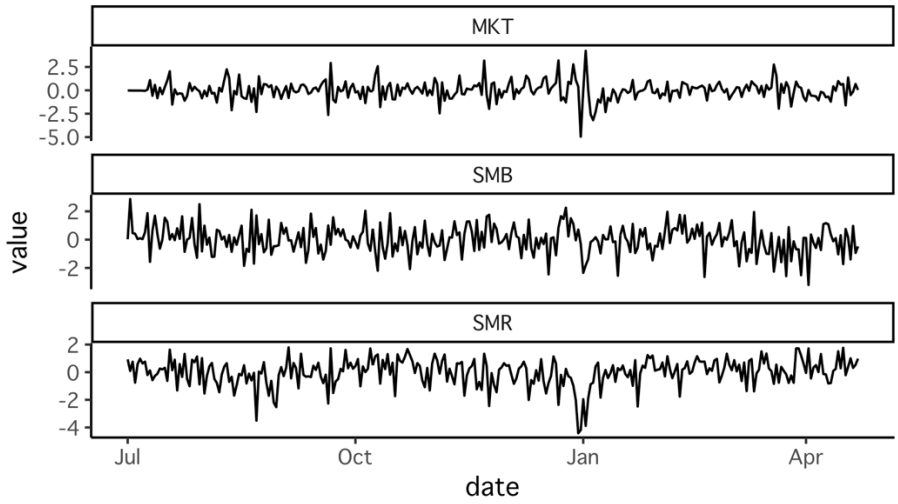


Figure 2. Time series plots for the retail demand fundamental factors.

Next, we investigate the relationship between the fundamental factors and the Principal Components (PC). Table 1 reports the correlation coefficients among these factors and seven principal components. According to this table, 1st PC is the weekly seasonal patterns that explains 93% of the overall variations. The *MKT* explains other seasonal patterns which could not be captured by PC1. The *SMB* correlated with PCs3 and 4 indicates those components are the source of size variations. Similarly, the *SMB* is correlated with PCs3, 5 and 6, and the corresponding principal components indicates the variations arise in unusual variations in restaurants visitors.

Table 1. Correlation matrix for the observed series, factors, and PCs.

	$Y_{I,t}$	*MKT*	*SMB*	*SMR*	PC1	PC2	PC3	PC4	PC5	PC6	PC7
$Y_{I,t}$	1	0.45	-0.07	0.01	0.93	0.28	0.13	0.05	0.13	-0.02	0.01
MKT	0.45	1	-0.13	0.19	0.27	0.36	0.12	0.05	0.17	-0.18	0
SMB	-0.07	-0.13	1	-0.23	0.01	0.21	-0.43	-0.12	-0.22	0.06	-0.17
SMR	0.01	0.19	-0.23	1	0.06	0.14	0.41	0	-0.37	-0.49	0.04

2. Cross-Sectional Regression and its Performance of Prediction

In this Section, we will present the followings:

- **Models and Assumptions on the multifactor models.**
- **Model selection together with tests for model assumptions are investigated.**
- **Interpretations for estimated model parameters are presented.**

- **Some predicted time-series plots together with predicting performance measures are shown**

For the first purpose, recall that we propose the following model:

$$y_{i,t} = \sum_{w=1}^{7} \beta_{i,w} D_{w,t} + \beta_{i,H} H_t + \sum_{f=1}^{m} \beta_{i,f} F_{f,t} + \varepsilon_{i,t}.$$

In order to obtain unbiased estimates of the OLS regression, it must be assumed that error terms and regressors including factors must be uncorrelated in both times-series and cross-sectional direction. This assumption is known to the strictly exogenous assumptions, which seems too strong to be hold. We need to perform some validity tests of our proposed models are adequate or not. The following Fama-MacBeth regression, together with GMM estimations are the powerful tools for panel time-series analysis. Cochrane (2005) recommends keeping the number of test portfolios to less than 10% of the number of observations in the GMM. Since we have 296 daily observations, we constructed 14 test portfolios based on their genre.

Figure 3 is the histograms of model adjusted R-squared values for all restaurants. It shows *MKT*, *SMB* and *SMR* weakly improve the interpretability overall.

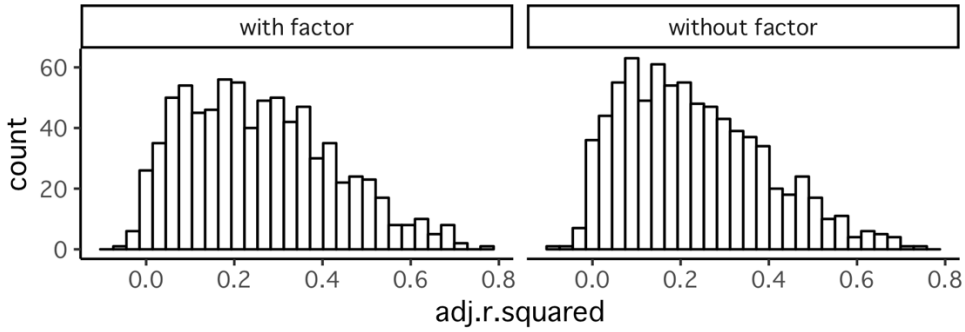


Figure 3. Histograms for the adjusted R-squared values for all restaurants. (left: our model, right: a model without *MKT*, *SMB* and *SMR*)

To see the validity of our model, we use Fama & MacBeth (1973) procedure, which is an alternative procedure for validating how factors describe portfolio or asset returns, and for producing standard errors and test statistics. Fama-MacBeth procedure is two-step regression. First, estimate coefficients with a time-series regression.

$$y_{it} = \Sigma_j \beta_{ij} x_{jt} + \varepsilon_{it}, \quad t = 1, \dots, T.$$

Second, estimate cross-sectional regression at each time period as below.

$$y_{it} = \sum_j \hat{\beta}_{ij} x_{jt} + \alpha_{it}, \quad i = 1, \dots, N.$$

We could estimate the covariance matrix of the sample errors by

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\alpha}}_t, \quad \widehat{\boldsymbol{\alpha}}_t = (\hat{\alpha}_{it}, \dots, \hat{\alpha}_{Nt})', \quad cov(\widehat{\boldsymbol{\alpha}}) = \frac{1}{T^2}\sum_{t=1}^{T}(\widehat{\boldsymbol{\alpha}}_t - \widehat{\boldsymbol{\alpha}})\,(\widehat{\boldsymbol{\alpha}}_t - \widehat{\boldsymbol{\alpha}})',$$

and then use sampling theory to test whether all the errors are jointly zero. See Cochrane (2005).

$$\widehat{\boldsymbol{\alpha}}' cov(\widehat{\boldsymbol{\alpha}})^{-1}\widehat{\boldsymbol{\alpha}} \sim \chi^2_{N-1}.$$

In the case of this study, the p-value of the above test statistic is 0.534 and it shows the errors are not significantly different from zero.

Table 2 shows estimated coefficients of each factor and tests. GMM estimator's asymptotic normality let us construct confidence bands for the estimator and conduct different tests. In the case of this data, *MKT* likely explains the numbers of customers. However, *SMB* seems to be a redundant factor, but *SMR* estimator is significant in café/sweets, dining bar, karaoke/party and "other"**.**

3. Discussion and Conclusion

We investigated he factor models for number of customers of restaurants in Japan. As well as finance, risk analysis is a useful tool for identifying and assessing the risks of retail demands. In particular, the demand for retail stores is strongly affected by calendar effect and it cannot be dispersed because it is a systematic risk. As a result, we showed there are another systematic risk of demand for restaurants besides the calendar effect and suggested a probable factor model based on it.

Table 2. Results of GMM estimate

restaurant genre	*MKT*	*SMB*	*SMR*
asian	0.064	-0.03	0.046
bar/cocktail	0.103*	0.046*	-0.019
café/sweets	0.081*	-0.009	0.06*
creative cuisine	0.132*	-0.008	0.049
dining bar	0.147*	0.012	-0.046*
international cuisine	0.155*	-0.023	0.047
italian/french	0.149*	0.028	0.013
izakaya	0.178*	-0.001	-0.053
Korean food	0.145*	0.01	-0.019
karaoke/party	0.153*	-0.07	-0.197*
okonomiyaki/monja/teppanyaki	0.113*	0.072	0.011
western food	0.118*	0.007	0.025
yakiniku/Korean food	0.182*	0.031	-0.052
other genre	0.098*	0.028	0.052*

* p-value significant at 5%

References

1. Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, ***9***, 3–18.
2. Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, ***47***, 13–37.
3. Basu, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks. Further evidence. *Journal of Financial Economics*, **12**, 129–156.
4. Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: empirical evidence. *The Journal of Finance*, **43**, 507–58.
5. Cochrane, J. (2005). Asset Pricing. Revised ed, Princeton: Princeton University Press.
6. Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3–56.
7. Fama, E. F., & French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The Journal of Finance*, **50**, 131–155.
8. Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: empirical tests. *Journal of Political Economy*, **81**, 607–636.
9. Hansen, L. (2012). Proofs for large sample properties of generalized method of moments estimators. *Journal of Econometric*, **170**, 325–330.
10. Harvey, A. C., & Shephard, N. (1993). 10 Structural time series models. *Handbook of Statistics*, **11**, 261–302.
11. Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, **18**, 439–454.
12. Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance*, **19**, 425–442.
13. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *American Statistician*, **72**, 37-45.

Multilevel time series modeling of mobility trends in the Netherlands for small domains

Sumonkanti Das [1], Harm Jan Boonstra[2], Jan van den Brakel[1]
[1] Maastricht University
[2] Statistics Netherlands

Abstract
The purpose of the Dutch Travel Survey is to produce reliable figures about mobility of the Dutch population. In this paper, multilevel time-series models have been developed to estimate reliable mobility trends at several aggregation levels, accounting for discontinuities induced by two different redesigns, and outliers due to less reliable outcomes in one particular year. The model is fitted to annual input series of direct estimates and standard errors at the most detailed breakdown into 504 domains defined by the combination of sex, age-class, motive and mode for the period 1999-2017. The standard errors of the direct estimates are smoothed through Generalized Variance Function (GVF) method. The model is fitted in a hierarchical Bayesian framework using Markov Chain Monte Carlo (MCMC) simulations. Global-local priors are considered for regularization purposes. Predictions for higher aggregation levels are obtained by aggregation of the most detailed domain predictions, resulting in numerically consistent set of trend estimates.

Keywords
Generalized variance function; Global-local priors; Hierarchical Bayesian approach; MCMC simulation; Small area estimation; Survey redesigns

1. Introduction

The purpose of the Dutch Travel Survey (DTS) is to produce reliable prediction of mobility trends (such as average distance per journey) of the Dutch population. The target variable in this paper is the average number of journey parts per person per day (pppd). A journey with a specific motive (e.g. traveling to work) can be made by more than one transportation mode. In such case, journey parts are defined as the breakdowns of the journey for a specific motive into separate sections made by the transportation modes. Thus journey parts are characterized by journey motive and transportation mode. In this study, the target parameter is estimated at the most detailed level based on the cross-classification of sex (male, female), age-class (0-5, 6-11, 12-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70+), motive (work, shopping, education, other), and mode (car driver, car passenger, train, BTM (bus/tram/metro), cycling, walking, other). Annual direct estimates and their standard errors for

the period 1999-2017 serve as input for developing multilevel time series models to predict more smooth and robust trend series.

Over the period of 1999-2017, there were two redesigns in the DTS: one in 2004 when the survey data collection was transferred to another agency and another one in 2010 when Statistics Netherlands restarted data collection. The period 1999-2003 is referred to as OVG, 2004-2009 as MON and 2010-2017 as OViN. These are abbreviations of Dutch names for the DTS used in the different periods. Discontinuities due to redesigns are sometimes masked by the volatility of the point estimates at detailed level. Some discontinuities are clearly visible at aggregate levels and need to be accounted for in the model development for trend estimation.

The direct estimates with their standard errors are available for all 504 domains, however there are some structural zeroes (such as motive work and mode car driver for children) for which the estimates are identically zero. Additionally, there are many other domains with zero direct estimates, but they are zero only coincidentally due to no observations of journeys in such domains in a particular year. Since the sample sizes of some domains are too small to produce reliable direct point estimates, the variance estimates are also unstable. In addition, for some domains point estimates are found unreliable for the year 2009 due to issues with the field work in the last MON year. These estimates behave as outliers in the time series data of 1999-2017. To obtain sufficiently stable variance approximations for the point estimates, required for time series modeling, the direct variance estimates are modeled by a GVF following Wolter (2007) to construct stable variance approximations for the direct variance estimates.

In Bollineni-Balabay et al. (2017) structural time series models and multilevel time series models were used to estimate trends for mobility (average distance traveled pppd) by journey motive and transportation mode. They found the differences between the two modeling frameworks were generally small. However, Boonstra and van den Brakel (2016) found multilevel time series models in a hierarchical Bayesian formulation have some advantages in terms of flexibility and computational efficiency. Therefore, time series multilevel model has been considered in this study to borrow strength over time and space while accounting for discontinuities and outliers.

2. Methodology

2.1 Input estimates

The direct estimates are obtained by the generalized regression estimator (Särndal et al., 1992). Standard errors of the direct estimates are approximated by Taylor linearization, and account for weighting and unequal inclusion probabilities. Let $\hat{Y}_{it}$ denote the direct estimate of the average number of journey parts pppd for year t, t = 1999,...,2017 and domain i, i = 1,...,504.

The estimated direct standard errors are reasonable for large domains, but not always for small domains with only a few observed journey parts (even zero in some cases). In order to obtain more reliable $se(\hat{Y}_{it})$, a GVF model has been developed based on the strong relationship between $se(\hat{Y}_{it})$ and $\hat{Y}_{it}$ as $se(\hat{Y}_{it}) \approx \frac{\hat{Y}_{it}}{\sqrt{m_{it}+1}}$, where m_{it} is the number of households contributing to this particular domain i in year t. However, since the estimates with small m_{it} (including the cases with $\hat{Y}_{it} = 0$) are not trustable, $\hat{Y}_{it}$ are replaced by simple smoothed estimates $\tilde{Y}_{it} = \lambda_{it}\hat{Y}_{it} + (1 - \lambda_{it})\hat{Y}_{i0}$ with $\lambda_{it} = \frac{m_{it}}{m_{it}+1}$ where $\hat{Y}_{i0}$ denotes the mean number of journey parts pppd over the years and sexes.

The direct estimates and smoothed standard errors are used as input for developing the multilevel time series models to obtain more smooth and robust trend series. To improve the model fit as well the convergence of the MCMC simulations, three different transformations of the input series have been considered. Of these transformations, the SQRT transformation works best in terms of model convergence and prediction of trend series. A Taylor linearization yields approximated standard errors as $se(\hat{Y}_{it}) \rightarrow se(\hat{Y}_{it})/(2\sqrt{\hat{Y}_{it}})$. These standard errors are undefined for the domains with no observed journeys (zero point estimate and standard error), but they were imputed using the GVF smoothing model discussed in the previous paragraph. In this case, the GVF model is applied to the transformed $se(\hat{Y}_{it})$ instead of original.

2.2 Multilevel time series model

Multilevel time series models for small area prediction are extensions of the basic area level model proposed by Fay and Herriot (1979). Here time series models are defined at the most detailed level constructed as the cross-classification of sex, age-class, motive, mode and year. For the description of the multilevel time-series model, the transformed initial estimates $\hat{Y}_{it}^{\text{sqrt}}$ are combined into a vector $\hat{Y}^{\text{sqrt}} = (\hat{Y}_{11}^{\text{sqrt}}, \ldots \hat{Y}_{M_d1}^{\text{sqrt}}, \ldots \hat{Y}_{1T}^{\text{sqrt}}, \ldots \hat{Y}_{M_dT}^{\text{sqrt}})'$, where M_d = 504 and T = 19.

Thus $\hat{Y}^{\text{sqrt}}$ is a $M = M_dT$ dimensional vector. Structural zero domains are not modeled, and hence the number of modeled initial estimates is reduced from $M = M_dT$ = 504 × 19 = 9576 to a total of 8720.

The multilevel models considered in this study can be expressed as a general linear additive form

$$\hat{Y}^{sqrt} = X\beta + \sum_{a} Z^{(a)}Z^{(a)} + e, \qquad (1)$$

where X is a $M \times p$ design matrix for a p-vector of fixed effects β, and the $Z^{(\alpha)}$ are $M \times q^{(\alpha)}$ design matrices for $q^{(\alpha)}$-dimensional random effect vectors $v^{(\alpha)}$. Here the sum over α runs over several possible random effect terms at different levels, such as transportation mode and motive smooth trends, white noise at the most detailed level of the M domains, etc. The sampling errors

$e = (e_{11},\ldots,e_{M_d1}\ldots e_{M_dT})'$ are taken to be normally distributed as $e \sim N(0,\Sigma)$ where $\Sigma = \Phi = \oplus_{t=1}^{T}\Phi_t$ with Φ_t the covariance matrix for the transformed direct estimates observed in year *t*. Here covariances between direct domain estimates are ignored and so Φ_t is assumed to be diagonal. Equations (1) with the distribution of *e* define the likelihood function as

$$p(\hat{Y}^{sqrt}|\eta,\Sigma) = \mathcal{N}(\hat{Y}^{sqrt}|\eta,\Sigma), \quad (2)$$

where $\eta = X\beta + \sum_a Z^{(a)}v^{(a)}$, called the linear predictor. The vector β of fixed effects is assigned a normal prior $p(\beta) = \mathcal{N}(0,100I)$, which is very weakly informative relative to the scales of the transformed direct estimates and the covariates. The random effect vectors $v^{(a)}$ for different α are assumed to be independent, but the components within a vector $v^{(a)}$ are possibly correlated to accommodate temporal or cross-sectional correlation. The superscript α is suppressed in what follows for notational convenience. Each random effects vector *v* is assumed to be distributed as

$$v \sim \mathcal{N}(0, A \otimes V), \quad (3)$$

where *V* and *A* are *d* × *d* and *l* × *l* covariance matrices, respectively, and $A \otimes V$ denotes the Kronecker product of *A* with *V*. The total length of *v* is $q = dl$, and these coefficients may be thought of as corresponding to *d* effects allowed to vary over *l* levels of a factor variable.The covariance matrix *A* describes the covariance structure between the levels of the factor variable, and is assumed to be known. Instead of covariance matrices, precision matrices $Q_A = A^{-1}$ are actually used for computational efficiency. The covariance matrix *V* is allowed to be parameterized as (i) fully parameterized (unstructured) covariance matrix, (ii) a diagonal matrix with different elements (diagonal), and (iii) a diagonal matrix with equal elements (scalar). A generalisation of (3) to non-normal distributions of random effects are considered assuming a Student-t distribution, horseshoe prior, or Laplace distribution.

The models are fitted using MCMC sampling, in particular the Gibbs sampler (Gelfand and Smith, 1990). See Boonstra and van den Brakel (2018) for a specification of the full conditional distributions. Model selection is based on WAIC (Watanabe, 2010) and DIC (Spiegelhalter et al., 2002) criteria. The model, a longer run of 1000 burn-in plus 10000 iterations of which the draws of every fifth iteration are stored, giving $3 * 2000 = 6000$ draws to compute estimates and standard errors.

3. Result

Some additional covariates are constructed in order to model the MON level break in 2004 (br mon taking values 1 for 2004-2009 years), the OViN level break in 2010 (br ovin taking values 1 for 2010-2017 years), and the influence of some lesser quality 2009 input estimates (as a dummy variable

denoted by dummy 2009). The year variable is also used quantitatively to define linear time trends by using its scaled and centered version (denoted as yr.c). The final time series model has been developed incorporating fixed and random effects of these covariates along with the sex, ageclass, motive and mode variables. A summary of the random effect terms included in the selected model is shown in Table 1. A very brief summary of the model building process is given below:

1. In the final selected model the following fixed effects components are included, where the term like $sex * ageclass$ includes both main and interaction effects:
   $$sex * ageclass + motive * mode + (ageclass + motive + mode) * (br\_ovin + yr.c) \quad (4)$$
2. Higher order fixed effect terms are modeled with random effects terms, which included level break effects, random intercepts, and random linear time trends. Full covariance among these effects improved the model. The resulting model term is named "V BR" in Table 1.
3. Time trend components were added at different levels, starting with smooth common trends at an overall level. In the end, using time trends at the two aggregation levels motive × mode and ageclass × motive × mode were found to work best. The terms are named RW2MM and RW2AMM respectively in Table 1. Best results have been obtained with diagonal variance for RW2MM and a scalar variance for the more detailed RW2AMM component.
4. To capture effects of the most influential 2009 outliers, random effects of dummy _2009 have been included at the domain level, and the resulting model term is named "V 2009".
5. White noise was added to capture unstructured dependence over all levels of all factors. The white noise term is named WN in Table 1.
6. Non-normal prior distributions have been tried for most random effect terms. A Laplace prior distribution for the random effects term "V BR" and a horseshoe prior for the outlier term "V 2009" were found to improve the model performance.

The means over the posterior draws are used as trend estimates, whereas the standard deviations serve as standard error estimates. The trend estimates based on the selected model, the direct input estimates and the model fitted values are shown in Figure 1 at overall level. The black lines are series of direct estimates, the red lines are the model fit based on all model components, and the green lines are trend series estimates benchmarked to the OViN level (including white noise). The trend lines, model fit and the direct estimates are compared at different level of aggregation including the most detailed level

of sex-ageclass-motive-mode. For illustration, only one figure has been shown to illustrate the model performance. The plots at the first row of Figure 2 show examples of structural zeros, the plots of second draw indicate how the "V 2009" component captures the 2009 outliers, the plots of third row show the effect of br ovin break variable and those of fourth row indicate combination of these two effects. It is also noted that the horseshoe prior for "V 2009" provides very small effects for most domains but very large for some domains.

4. Discussion and Conclusion

The purpose of the paper was to develop a suitable time series model for predicting the average number of journey parts pppd based on the time-series data of 1999-2017 accounting for the two redesigns and the influence of the 2009 outliers along with the problem of unstable standard errors of direct estimates. The GVF model has been developed to obtain smooth estimates of the standard errors, used as input in time series model development. The final time series model consists of fixed effects as well as several random components which account for the discontinuities, 2009 outlier effects, smooth trend components at two lower aggregation levels, and white noise at the most detailed level. In addition, global-local priors have been incorporated in the distribution of the "V 2009" and "V BR" random components. By construction, the fitted model provides numerically consistent predictions at all aggregation levels. The study shows the fitted model at the most detailed level can be used to produce reliable estimates at the higher aggregation levels. Though the input estimates are assumed independent in this study, their correlations will be incorporated in the model development in further studies.

Model Component	Formula **V**	Variance Structure	Factor **A**	Prior	Number of Effects
V 2009	*dummy* 2009	scalar	*sex* ∗ *ageclass*∗ *motive* ∗ *mode*	horseshoe	464
V BR	1 + *yr.c*+ *br mon* + *br ovin*	unstructured	*sex* ∗ *ageclass*∗ *motive* ∗ *mode*	Laplace	1856
RW2AMM	*ageclass* ∗ *motive* ∗ *mode*	scalar	RW2(yr)	normal	4360
RW2MM	*motive* ∗ *mode*	diagonal	RW2(yr)	normal	532
WN	1	scalar	*sex* ∗ *ageclass*∗ *motive* ∗ *mode* ∗ *yr*	normal	8720

Table 1: Summary of the Random Effect components for the Final Time Series Model for the period 1999-2017. The second and third columns refer to the varying effects

with covariance matrix V in (3), whereas the third and fourth columns refer to the factor variable the effects are varying over, associated with A in (3). The last column contains the number of random effects for each term.

References

1. Bollineni-Balabay, O., J. van den Brakel, F. Palm, and H. J. Boonstra (2017). Multilevel hierarchical bayesian versus state space approach in time series small area estimation: the dutch travel survey. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 1281–1308.
2. Boonstra, H. J. (2018). mcmcsae: MCMC Small Area Estimation. R package version 0.9.
3. Boonstra, H. J. and J. van den Brakel (2016). Estimation of level and change for unemployment using multilevel and structural time series models. Technical Report 201610, https://www.cbs.nl/nl-nl/achtergrond/2016/37/ estimation-of-level-and-change-for-unemployment, Statistics Netherlands.
4. Boonstra, H. J. and J. van den Brakel (2018). Hierarchical bayesian time series multilevel models for consistent small area estimates at different frequencies and regional levels. Statistics Netherlands discussion paper, December 4, 2018.
5. Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of james-stein procedures to census data. Journal of the American Statistical Association 74(366), 269–277.
6. Gelfand, A. and A. Smith (1990). Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.
7. S¨arndal, C.-E., B. Swensson, and J. Wretman (1992). Model Assisted Survey Sampling. Springer.
8. Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64(4), 583–639.
9. Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.
10. Wolter, K. (2007). Introduction to Variance Estimation. Springer.

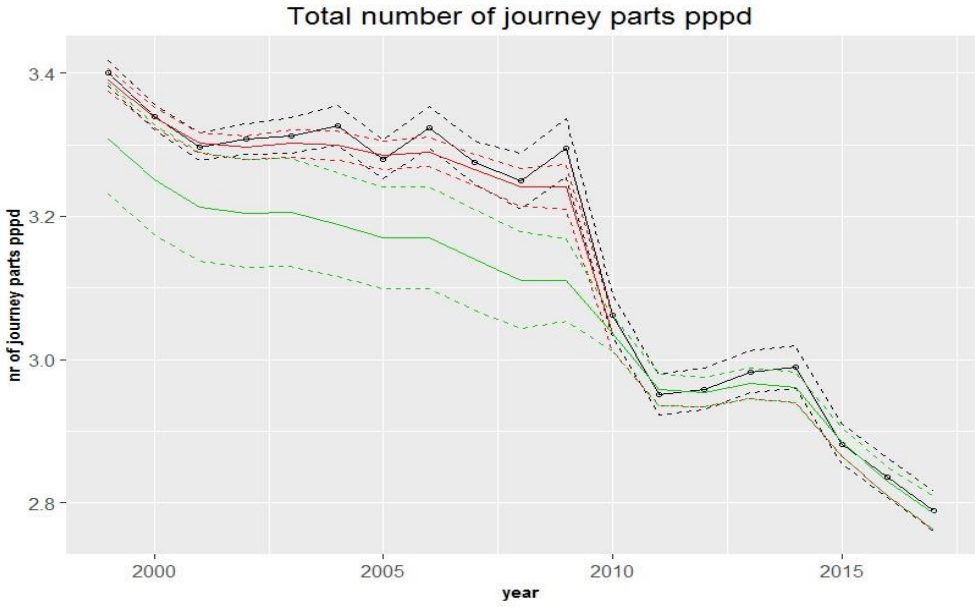


Figure 1: Direct estimates (black), model fit (red) and trend estimates (green) with 95% intervals.

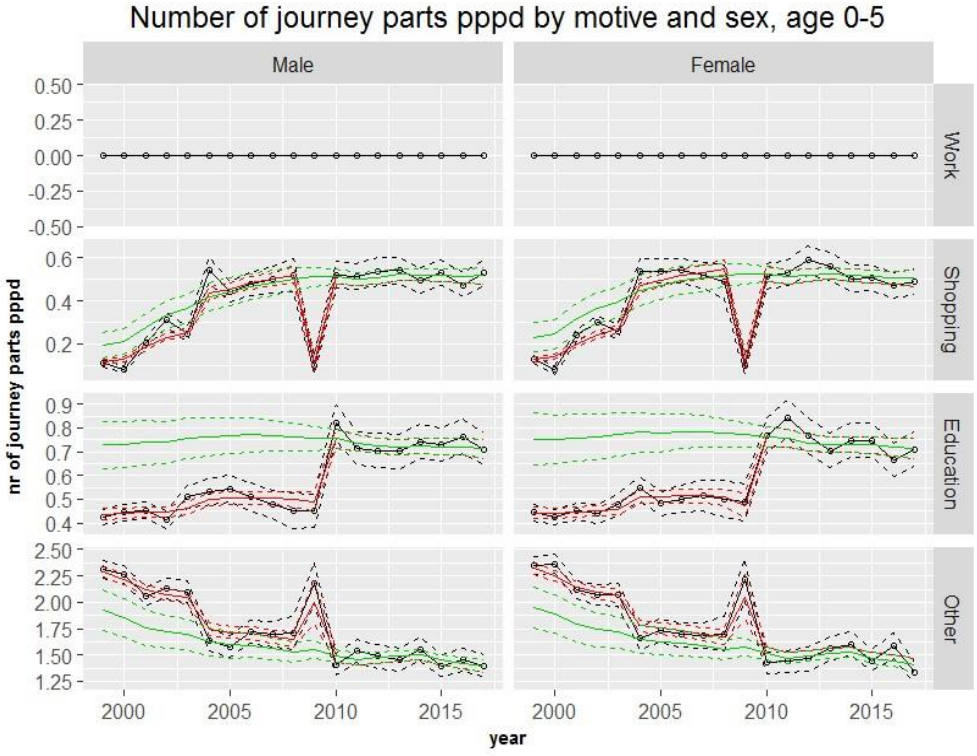


Figure 2: Direct estimates (black), model fit (red) and trend estimates (green) with 95% intervals.

Asymmetry of international trade statistics: a focus on Sarawak's liquefied natural gas (lng) trade with Japan

Jee, Hui-Siang Brenda, Kiew, Leh-Yieng, Omar, Surhardi, Yahya, Roslawati
Department of Statistics Malaysia, Sarawak

Abstract

In the era of globalisation and trade liberalisation, international trade is perceived as a major driver of economic growth. The 2030 Agenda for Sustainable Development also recognises international trade as an engine for inclusive economic growth and poverty reduction, and an important means to achieve the Sustainable Development Goals (SDGs). In relation to this, the compilation of quality data needs to be fulfilled. However, in international trade statistics, it has faced challenges of inconsistency that can be examined through mirror analysis. In this paper, we intend to discuss the asymmetry trade statistics of liquefied natural gas (LNG) between Sarawak and Japan from the year 2008 to 2017. By applying bilateral trade discrepancy index, the results of the study showed that data discrepancy of Sarawak's LNG exports to Japan was positive. This indicates that the reported exports by Sarawak were less than the reported imports by Japan. However, the degree of discrepancy was relatively low, in the range of 3.0% to 14.0% that were influenced by trade valuation, exchange rate and other factors. Although it is difficult to eradicate these issues in international trade, it is however important to measure and identify the cause of asymmetry in future studies for the purpose of compiling better quality statistics.

Keywords

Mirror Analysis; Exports; Trade Discrepancy.

1. Introduction

International trade is the exchange of capital, goods and services across international borders or territories[1]. In the era of globalisation and trade liberalisation, international trade is perceived as a major driver of economic growth. As stated in United Nations Conference on Trade and Development (2018), "The 2030 Agenda for Sustainable Development recognises international trade as an engine for inclusive economic growth and poverty reduction, and an important means to achieve the Sustainable Development Goals (SDGs)".

In relation to this, the compilation of quality data needs to fulfil several dimensions such as data completeness, consistency, accuracy, validity and

[1] Definition of international trade is referred from Wikipedia (2018)

timeliness. However, international merchandise trade statistics has faced the challenges of inconsistency that can be examined through mirror analysis[2] and further reconciliation study as encouraged in International Merchandised Trade Statistics (IMTS), Revision 3, paragraph 9.18[3].

A considerable amount of literature had been published on mirror analysis of merchandise trade statistics, such as Day (2015) conducted a mirror study by comparing China's merchandise trade statistics with its major trading partners with the purpose to assess the accuracy of trade data; Javorsek (2016) analysed asymmetries in bilateral trade between selected Asia-Pacific countries at regional and further details at countries with higher asymmetries; and Markhonko (2014) that studied the issues of international trade statistics asymmetries in constructing inter-country input-output model.

In extension, some researchers had focused their study in certain commodity sections. For example, Ferrantino and Wang (2007) proposed an index in terms of the average asymmetry to measure the discrepancies across trade sectors between United States with China and Hong Kong; Guo (2010) focused on international trade statistics in manufacturing goods and found that asymmetric pattern exist between China and its top five trading partners; and Fisher et al. (2014) reported the asymmetry on external trade statistics between Palestine and Switzerland based on 6-digit tariff heading.

Studies of reconciliation such as that conducted by the Joint Commission on Commerce and Trade Statistics Working Group (2012) and China-Canada Joint Working Group on Trade Statistics Reconciliation (2018) had discussed asymmetries factors in bilateral trade and demonstrated adjustment which more closely aligns the two sets of data.

So far, however, there have been fewer studies about mirror analysis that focuses in detail on Harmonized Commodity Description and Coding System (HS). In addition, no research has been conducted on Sarawak trade asymmetric. Thus, this paper was prepared to study asymmetry trade statistics between Sarawak and Japan by focusing on a specific product which is liquefied natural gas (LNG). The focus on a single product had portrayed an additional source of reference to measure the quality trade statistics with major trading partner and product that have greater influences on that region.

Sarawak, the largest state in Malaysia that covers 124,451 km2 land area, is located in northwest Borneo Island. Abundant in natural resources such as

[2] Mirror analysis refers to the comparison of imports and exports data between the trade partners. The inconsistency exists when the reported exports from country A to country B do not match the reported imports of country B from country A, which term as bilateral asymmetries in United Nation International Trade Statistics.

[3] "Countries are encouraged, therefore, to periodically conduct bilateral and multilateral reconciliation studies or implement data exchange so that their statistics can be more accurate and useful for both national purpose and for international comparisons".

LNG, crude petroleum, timber products and etc. make exports an important activity in Sarawak. In 2017, Sarawak's exports was valued at RM97.6 billion. The major exported products of Sarawak were LNG that contributed 43.0% of total Sarawak's export, followed by palm oil (11.3%), crude petroleum (10.0%), aluminium (6.7%), condensate and other petroleum oil (5.1%) and other products (23.9%). As the Sarawak's top exports product, LNG was exported more than half to Japan (57.6%), and other destinations such as Republic of Korea (13.9%), People's Republic of China (13.5%) and Taiwan (11.5%). Due to the above reasons, it was significance to study Sarawak-Japan trade statistics asymmetry by focusing on LNG product.

The main objective of this paper is to analyse the asymmetry trade statistics between Sarawak and Japan using an empirical method. Particularly to identify the degree of data discrepancy between Sarawak-Japan LNG trade statistics and also assess the quality of data.

The following was the organisation of this paper: Section 2 described the methodology used in this study. Section 3 showed the empirical results and discussion. Finally, Section 4 presented the conclusion of this study.

2. Methodology

In conducting this study, time series annual data spanning from 2008 to 2017 were adopted. The variables used in this study includes Sarawak's exports of LNG to Japan that was obtained from various issues of Sarawak External Trade Statistics published by Department of Statistics Malaysia, and Japan's imports of LNG from Sarawak that was compiled from Trade Statistics of Japan published by Ministry of Finance Japan. All the variables were expressed in US Dollar by converting the local currency to US Dollar based on the exchange rate taken from International Financial Statistics published by International Monetary Fund.

By referring to Guo (2010), this study applied a bilateral trade discrepancy index to gauge trade asymmetry between two trade partners. The formula is as below:

$$DIF^{AB} = \frac{M^{AB} - E^{AB}}{M^{AB}}$$

where M^{AB} is the imports value reported by B from A (country B is the reporting country and A is the partner country) and E^{AB} represents the exports value reported by A to B (country A is the reporting country and B is the partner country).

In the case of country A as an exporter, it measures the difference between the imports reported by B from A (M^{AB}) and the exports reported by A to B (E^{AB}) as a proportion of imports reported by B from A.

The positive or negative value of index from this formula is interpreted as:

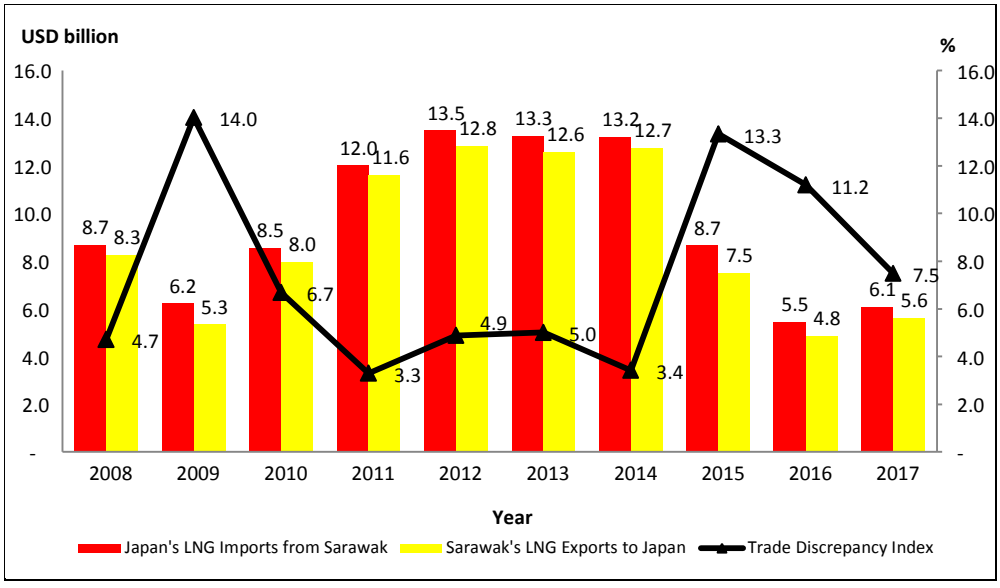
(i) If reported exports is smaller than reported imports (positive value), A is under-reporting exports to B or B is over-reporting imports from A;
(ii) If reported exports is greater than reported imports (negative value), A is over-reporting exports to B or B is under-reporting imports from A.

3. Result

3.1 Results

Figure 1 presented the results for bilateral trade discrepancy index. The figure showed that there was a gap between exports and imports data of LNG reported by Sarawak and Japan. From the year 2008 to 2017, Sarawak as the exporter of LNG to Japan had resulted in positive value of indexes. In other words, the reported export by Sarawak is less than the reported imports by Japan. This implies whether Sarawak was under-reporting its exports to Japan or Japan was over-reporting its imports from Sarawak.

Figure 1: Sarawak-Japan LNG Trade and Trade Discrepancy Index, 2008-2017



Source: Department of Statistics Malaysia, Ministry of Finance Japan and International Monetary Fund.

Throughout the analysis, trade discrepancy indexes were in the range of 3.0% to 14.0%. The index for the year 2008, 2011, 2012, 2013 and 2014 showed relatively low trade discrepancies, which was 5.0% or lower. Meanwhile, in the year 2010 and 2017, the discrepancy index was moderate at the range of 6.0% to 8.0%. The discrepancy index was relatively large in year 2009, 2015 and

2016, in between 10.0% to 14.0%. However, according to Gehlhar (1996), it is regarded as accurate match if the discrepancy index is less than 20.0%. Thus, this result mirroring the quality of trade data between Sarawak and Japan is quite accurate. Although discrepancy did exist, it was due to some reasons that are to be discussed.

3.2. Discussion

From the results discussed above, although both Sarawak and Japan followed the United Nations guidelines on merchandise trade statistics compilation (refer Table 1), it does not mean that the corresponding import/ export data will match. There are several aspects of the guidelines, such as valuation and partner country attribution, that when followed, created bilateral discrepancies.

Table 1: Comparison of Statistical Concepts and Definitions

Region/ Country	Sarawak	Japan
Partner countries	Exports: Country of last known destination Imports: Country of origin	Exports: Country of last known destination Imports: Country of origin
Valuation methods	Exports: FOB; Imports: CIF	Exports: FOB; Imports: CIF
Trade system	General	General
Commodity codes	Goods are classified based on the HS which consist of 6-digit HS code and 4-digit domestic code.	Goods are classified based on the HS which consist of 6-digit HS code and 3-digit domestic code.

Source: Department of Statistics Malaysia and Ministry of Finance Japan.

As stated in the United Nations Trade Statistics official website, the three main and well-known reasons for asymmetries in bilateral merchandise trade are:

(i) The application of different criteria of partner attribution in import and export statistics;
(ii) The use of Cost, Insurance and Freight (CIF) type of valuation in import statistics and Free on Board (FOB) type of valuation in export statistics; and
(iii) Application of different trade systems in data compilation.

Based on the valuation methods, import statistics in Japan included international freight and insurance charges, thus valuing on a CIF basis, while Malaysia excluded these charges in their exports statistics, valuing on FOB

basis. As a result, Japan's value of imports was conceptually higher than Sarawak's value of exports. This study proved the same results, which where import values of LNG by Japan was always greater than exports values of LNG by Sarawak throughout the period of ten years.

To make them more comparable internationally, the asymmetries caused by differences in valuation had to be focused on. As pointed out in IMTS, Revision 3, paragraph 4.8(b) and 4.9, while re-emphasizing the recommendation to compile CIF-type value of import statistics, countries are encouraged "to compile FOB-type value of imported goods as supplementary information" and/ or "to compile separate data for freight and insurance, at the most detailed commodity and partner level possible". Ideally, the FOB-type of imports value should be compiled by gathering transaction-level data on freight and insurance. However, in practice, this method may be not suitable for most countries.

In addition, the time lag between exports and imports is another possible reason for trade discrepancies between trading partners. According to China-Canada Joint Working Group on Trade Statistics Reconciliation (2018), the shipment time lag difference referred to the difference in bilateral statistics that generally results from long-distance ocean shipping, whereby shipment of commodities departs from exporting country at the end of the year and arrive importing country in the following year. For example, goods departing from Sarawak in late December 2016 might only arrive Japan in early January 2017. In our study, the trade discrepancy index for a total ten-year period (2008 to 2017) was only 6.5% which was at moderate discrepancy rate. This revealed that time lag influenced the asymmetry in annual bilateral trade data.

Another factor of asymmetry is the currency exchange rate. Valuation of LNG imports was at currency rate during arrival in Japan while valuation of exports was at currency rate during the time of exportation. Exchange rate fluctuations may, therefore, lead to statistical differences. This was clearly shown in the year 2009 where there was major oil price volatility and a global financial crisis, the trade discrepancy index was the highest throughout the period. In 2015, again there was a sudden rise of trade discrepancy index, which was due to depreciation in Malaysia currency against US Dollar, as well as oil and gas price plunge.

4. Conclusion

In conclusion, this paper analysed the asymmetry trade statistics between Sarawak's exports of LNG to Japan using the empirical method. This was in contrary to most empirical analyses that were less focused on specific products by country. Particularly, we intended to identify the degree of data discrepancy and assess the quality of data between both regions.

By analysing the asymmetry relationship using the bilateral trade index, the following results were observed: First, Sarawak as the exporter of LNG showed positive value of indexes against Japan that was an importer. Second, the trade discrepancy index was relatively low at less than 15.0% throughout the study period. Thus, this result mirroring the quality of trade data between Sarawak and Japan had been much accurate.

From the empirical investigation, the issues on asymmetry trade statistics were noted. They were influenced by trade valuation, exchange rate and other factors. Although it is difficult to eradicate these issues in international trade, it is however important to measure and identify the cause of asymmetry for the mean of compiling better quality statistics. For these reasons, it is suggested to further analyse the cause of asymmetry trade statistics in future studies.

References

1. China-Canada Joint Working Group on Trade Statistics Reconciliation. (2018). *Comparing Canada's and China's bilateral trade data*. Retrieved from https://www150.statcan.gc.ca/n1/pub/13-605x/2018001/article/54962-eng.htm
2. Day, I. (2015). *Assessing China's merchandise trade data using mirror statistics*. Retrieved from https://www.rba.gov.au/publications/bulletin/2015/dec/pdf/bu-1215-3.pdf
3. Department of Statistics Malaysia. *Sarawak External Trade Statistics*, various issues. Sarawak: Department of Statistics Malaysia.
4. Ferrantino, M.J. and Wang, Z. (2007). *Accounting for discrepancies in bilateral trade: The case of China, Hong Kong, and the United States*. United States International Trade Commission, Office of Economics Working Paper No. 2007-04-A.
5. Fisher, N., Pfammatter, M., Carnal, G., Khalifa, H., Almasri, M., Khalil, A., and Abu Bakar, F. (2014, August). *Asymmetry study on external trade statistics between Palestine and Switzerland*. Retrieved from http://www.pcbs.gov.ps/Downloads/book2118.pdf
6. Gehlhar, M.J. (1996). *Reconciling bilateral trade data for use in GTAP*. Retrieved from https://www.gtap.agecon.purdue.edu/resources/download/38.pdf
7. Guo, D. (2009). *Mirror statistics of international trade in manufacturing goods*: The case of China. United Nation Industrial Development Organization (UNIDO), Research and Statistics Branch Working Paper 19/2009.

8. International Monetary Fund. (2018). *International Financial Statistics.* Retrieved from https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505A05A558D9A42&sId=1479331931186
9. Javorsek, M. (2016). *Asymmetries in international merchandise trade statistics*: *A case study of selected countries in Asia-Pacific*. United Nations Economic and Social Commission for Asia and the Pacific (ESCAP), Working Paper Series SD/WP/02/April 2016.
10. Joint Commission on Commerce and Trade Statistics Working Group. (December, 2012). The second phase report on the statistical discrepancy of merchandise trade between the United States and China. Retrieved from https://unstats.un.org/unsd/tradekb/Attachment421.aspx?AttachmentType=1
11. Markhonko, V. (2014). *Asymmetries in official international trade statistics and analysis* of globalization. Paper presented at the International Conference on the Measurement of International Trade and Economic Globalization, Aguascalientes, Mexico.
12. Ministry of Finance Japan. (2018). Trade statistics of Japan. Retrieved from https://www.estat.go.jp/en/statsearch/files?page=1&layout=datalist&toukei=00350300&tstat=000001013141&cycle=1&year20170&month=24101212&tclass1=000001013180&tclass2=000001013182&result_back=1
13. United Nations. (2011). *International Merchandise Trade Statistics*: Concepts and Definitions 2010, Revision 3. New York: United Nation Publisher.
14. United Nations Conference on Trade and Development. (2018). Trade and the Sustainable Development Goals (SDGs). Retrieved from https://unctad.org/en/Pages/DITC/TradeAnalysis/TAB-Trade-and-SDGs.aspx
15. United Nations Trade Statistics. (2018). Bilateral asymmetries. Retrieved from http://unstats.un.org/unsd/tradekb/Knowledgebase/50657/Bilateral-asymmetries
16. Wikipedia. (2018). International trade. Retrieved from https://en.wikipedia.org/wiki/International_trade

Digital transformation on Population and Housing Census of Malaysia 2020

Ezatul Nisha Abdul Rahman, Hamka Ismail, Fatimah Az-Zahra Abdul Shukor, Fatin Nabilah Sabri, Wan Nor Elina Wan Setapa
Population and Demographics Statistics Division, Department of Statistics Malaysia

Abstract
In preparation for Malaysia Population and Housing Census 2020 (MyCensus 2020), Department of Statistics Malaysia (DOSM) aims to modernise and improve the census implementation in 2020 through Malaysia Census Transformation Program (MyCTP). This paper seeks to highlight and explain the modernisation for Census 2020, and to compare and contrast it with similar census transformation program implemented by other countries such as the United Kingdom, United States of America, New Zealand and Australia. MyCTP focuses on six elements of innovation which are modernising address register integrating administrative data from other government agencies (OGA), improving selfresponse rate, modernising census operation, establishing one stop centre for data dissemination through Census Portal and increasing census data security. This paper also discusses the strategy and timeline of implementation for MyCTP, as well as the risks and benefits that will be faced before, during and after MyCensus 2020.

Keywords
Census transformation, census methodology, population, administrative data, integrated database

1. Introduction

Malaysia's population is projected to rise to 33.8 million by 2020, compared to 28.6 million in 2010. Meanwhile, the number of dwellings is projected to reach 9.9 million in 2020 compared to 7.4 million in 2010. The Census 2020 will have ramifications by multiple environmental factors that have the potential to impact its success such as changing of population profile, extensive use of ICT, complex modus operandi and also dynamic and demanding user. Hence, all these factors lead to the needs of Malaysia Census Transformation Programme (MyCTP).

MyCTP is a transformation that encompasses six elements of innovation in the implementation of the Population and Housing Census of Malaysia, 2020 which are modernisation of listing methods, integrating administrative data, increasing self-response rate, modernisation in field work, census portal as a one-stop centre and increasing data security.

People nowadays believe that the government is one of the more trustworthy data collectors and place their confidence in the security of the information regarding their private activities (Rainie and Duggan, 2015). The operational design of MyCTP towards modernization through a new ICTbased framework aligned with Malaysia National Development Strategy (MyNDS) outlined in the 11th Malaysia Plan of delivering high-performance, high impact and cost-effective outcomes. Besides, examining France's experience after the first decade of the rolling census, (Durr and Clanch´e, 2013) note that costs have stabilized across the years but have increased proportionally to the population as the data processing workload.

Furthermore, the role of MyCTP is also to facilitate national and international purposes through the 2030 Agenda Sustainable Development which places increasing demand for expanded data collection. MyCTP also has been set up to meet the best practices as other countries such as United States of America (USA) through 2020 Census Transition Plan, United Kingdom (UK) through Census Transformation Programme and Canada through The 2016 Census Strategy Project.

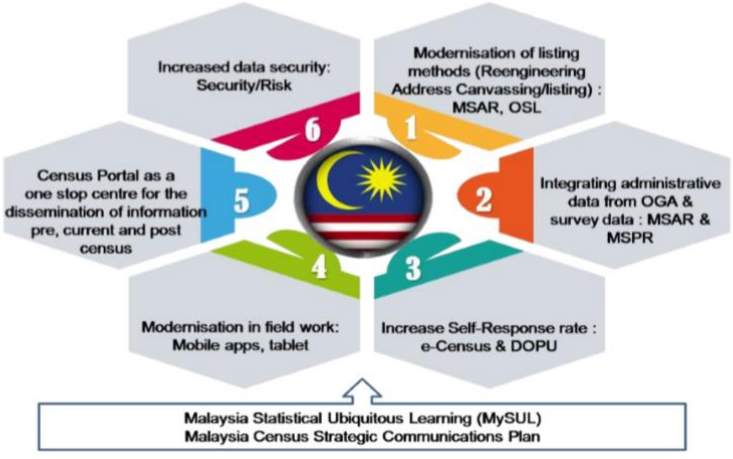
The main objective of MyCTP is to transform the implementation of Census 2020 towards the use of ICT in order to produce quality and timely statistics. It is also to introduce modernisation in census activities to be more efficient and cost-effective as well as to enhance the potential of the Department of Statistics Malaysia through long-term investments. Another objective of MyCTP is to increase the use of administrative data in the Census 2020 by integrating census data, administrative data and surveys into integrated databases. Lastly, it is aimed to provide facilities for a self-response mode of enumeration in order to increase the confidence of respondents in providing information including security of data.

2. Methodology

The six elements in MyCTP will help to overcome the challenges by instituting a transformation program that aspires to ensure a thorough and efficient census management system. Transformation program designed to empower the implementation of Malaysia Population and Housing Census 2020, on top of establishing a basic framework for continuous improvement of population and housing data collection system of Malaysians in the future. Similar census transformation programmes are also implemented by other countries. The scope of transformation implemented varies according to their respective aspirations, as well as social and geopolitical climates.

DOSM developed two systems which are Malaysia Integrated Population Census System (MyIPCS) and Malaysia Statistics Ubiquitous Learning (MySUL) as a part of initiatives to materialise MyCTP. MySUL is a platform for online learning and management of training materials for prior, current and post-

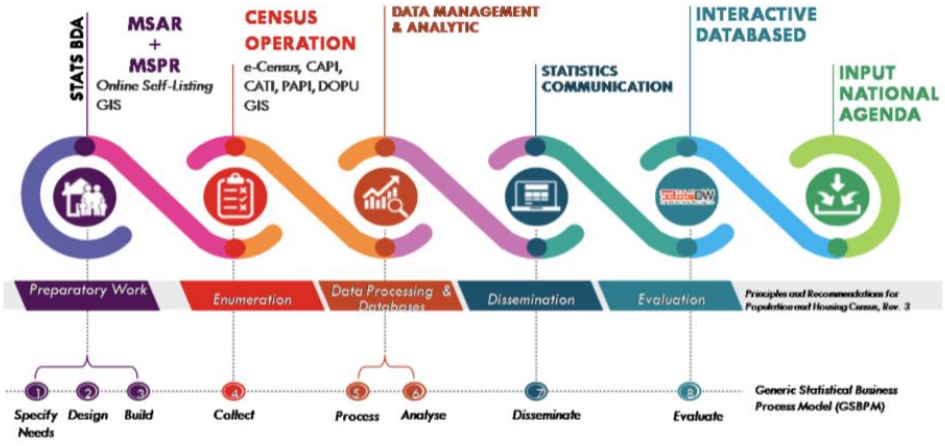
censorship activities. The basis of MySUL is online learning (e-Learning) for delivery of a learning, training or education program by electronic means. It also for involves the use of a computer or electronic device (e.g. a mobile phone) in some way to provide training, educational or learning material. There are four objectives for MySUL, which are implement e-Learning via online, provide online learning management, prepare level for understanding of members, and preparing integrity platforms with DOSM Training Information Management System (DTIMS). Basically, just selected person can only access this system and most of them are from DOSM. All the data collected from each committee that involved in census.



MSAR: Malaysia Statistical Address Register
MSPR: Malaysia Statistical Population Register
OSL: Online Self Listing OGA:
Other Government Agencies
DOPU: Drop-Off Pick Up
Source: Department of Statistics Malaysia

Figure 1: Malaysia Census Transformation Programme (MyCTP)

MyIPCS is the 2020 Census framework developed as a large-scale project planning and monitoring mechanisms. MyIPCS aims to create a centralized and comprehensive census system that provide a uniform and integrated framework covering Pre Census, During Census and Post Census. MyIPCS consists of ten modules which are developed in line with Principles and Recommendations for Population and Housing Census, Revision 3 and Generic Statistical Business Process Model (GSBPM).



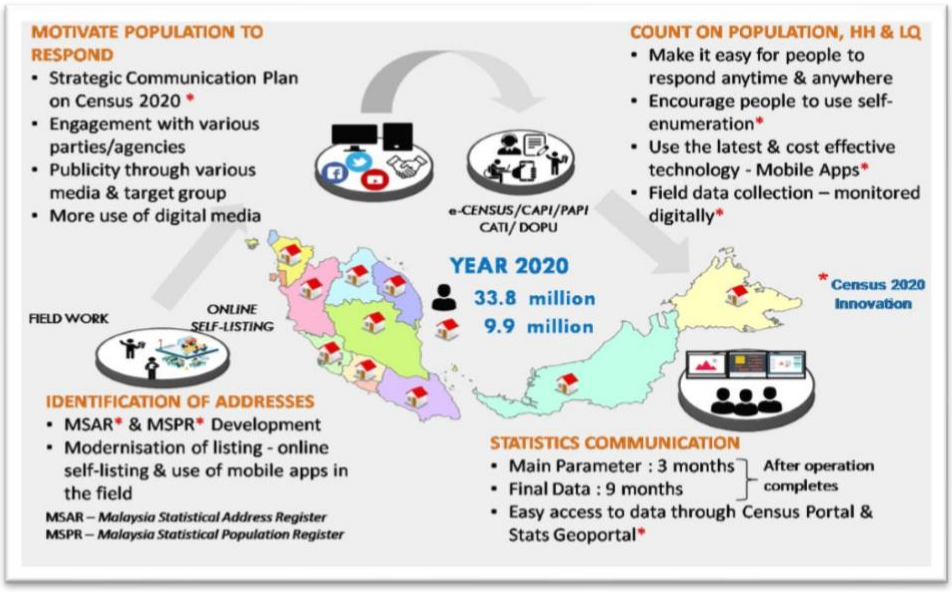
Source: Department of Statistics Malaysia

Figure 2: Malaysia Integrated Population Census System (MyIPCS)

There are four strategies on implementation for MyCensus 2020. The first stage is to identify the addresses by the modernisation of address listing methods. Modernisation can be achieved through the development of Malaysia Statistical Address Register (MSAR) as a complete database of residential addresses in Malaysia. Once the addresses have been identified, the next stage is to motivate the population to respond by engaging with various parties or agencies through media and target group to increase self-response rate. Hence, Strategic Communication Plan on Census 2020 will be implemented in this stage. Next, respondents will fill out the questionnaire forms based on their modes of preference (e-Census, CAPI, CATI, DOPU, PAPI). From the collected data, population, household and living quarters can be counted. The last stage is to produce the final data of MyCensus 2020. After the operation completes, the main parameter will be produced within three months and preparing the final data within nine months.

The five modes of enumeration for Census 2020 namely e-Census, Computer Assisted Personal Interview (CAPI), Computer Assisted Telephone Interview (CATI), Paper Assisted Personal Interview (PAPI) and Drop-Off Pick-Up (DOPU). For e-Census method, respondents will fill in the questionnaire via online which provided with ID Address through Portal Population and Housing Census 2020. This method use to minimize the cost of census-taking and maximize efficiency of available resources (EneMargit Tiit, 2011). Other three methods (CAPI, PAPI, and CATI), enumerators will interview the respondents by visiting their living quarters and all information will be recorded into system (online or offline). For CAPI is a face-to-face interview using tablet or census application, PAPI is on paper and CATI by phone interview. Lastly, is by using DOPU which is enumerators will drop DOPU kit consists of questionnaire, pen, envelope, instruction to complete the questionnaire, instruction to fill in e-

Census and re-visit letter to respondent. Enumerator will pick up the completed questionnaire.



Source: Department of Statistics Malaysia

Figure 3: MyCensus 2020: Strategies and Innovations

3. Result

The implementation of MyCTP and development of MyIPCS will produce comprehensive new statistical information to the smallest geographical area which will become a basis for national planning and development. The transformation will initiate the development of register based statistics (administrative data) which comprises a complete living quarters and population database called MSAR and MSPR. The updated data, especially from MSAR and MSPR, can be shared in various other government agencies thus eventually improving the cooperation among other government agencies. The development of MyIPCS will transform the statistical landscape of Malaysia from paper based to a digital platform which involves the process of Preparatory Work, Enumeration, Data Processing, Dissemination and Evaluation based on Principles and Recommendation for Population and Housing Census, Revision 3. This modernisation will improve the delivery system to be more modern, dynamic and user friendly as well to promote the statistical literacy among the population. The information obtained from MyCensus of 2020 will be the basis of estimates and projections of Population, Living Quarters and Households which will be the population data inputs for developing the next key public policies. The use of the latest technology in MyCTP and MyIPCS is a long-term and value-for-money investment that is not

only for MyCensus 2020 but also for future Censuses and Surveys by using the same integrated architecture.

4. Discussion

DOSM aims to minimize human resources needed, reduce census operation time while rapidly generating current population and demographic data by using the use of latest, more dynamic and flexible ICT infrastructure and application system. Hence, the use of relevant administrative data in MyCensus 2020 is hoped to achieve its objectives effectively.

By using current Open Data Initiative and Big Data Analytics platform in order to analyse the huge data from MyCensus 2020, this transformation will also lead to a more well-organised of pre, during and post census activities. More quality and reliable data will be produced as the primary data can be compared with administrative data from DOSM and other agencies.

The development of MyIPCS is a long term and continuous investment as well as cost effective since it can be used beyond MyCensus 2020. The transformation will lead to the implementation of smaller scale of Census in the future.

5. Conclusion

By providing platforms such as MyCTP and MyIPCS, the effort to ensure MyCensus2020 could be capitalized in data acquisition of the national landscape is fulfilled. The modernisation of Population Census 2020 through MyIPCS introduces new elements and business process to the Department of Statistics Malaysia. As the world changes, these platforms could lead the way in big data and machine learning in the future development of the digital revolution. Thus save valuable resources and financial burdens DOSM faces and increase data processing efficiency. In the future, the study in the area of data integration and data security within the system can be explored.

References

1. Malaysia National Development Strategy (MyNDS), 11th Malaysia Plan - Speech by Prime Minister of presenting 11th Malaysia in Parliament of Malaysia
2. Rainie, Lee, Duggan, M. "Privacy and Information Sharing" Pew Research Center, December 2015.
3. Durr, J. M. and F. Clanch'e (2013). The French Rolling Census: A Decade of Experience. In 59th ISI World Statistics Congress. International Statistical Intitute.
4. MacGibbon, A. (2016). Review of the Events Surrounding the 2016 eCensus: Improving Institutional Cyber Security Culture and Practices Across the Australian Government. Department of the Prime Minister and Cabinet.
5. Census, Stats NZ, 2018, New Zealand Government https://www.stats.govt.nz/
6. 2020 Census Operational Plan, A New Design for the 21st Century, Issued September 2017, Version 3.0, United States Census Bureau, https://www.census.gov/programs-surveys/decennial-census/2020census/planning-management/planning-docs.html
7. Census Transformation Programme,2018-2021,Office for National Statistics, https://www.ons.gov.uk/census/censustransformationprogramme
8. Census Strategy Project, 2011-2016, Statistics Canada, https://www.statcan.gc.ca/eng/consultation/2011/csp-psr-eng
9. E-Census as a New Approach to the Population and Housing Census https://www.stat.go.jp/english/info/meetings/eastasia/pdf/1kpaper.pdf
10. Tiit, E. M. (2011) Population and Housing Census Methodology, Published by Statistics Estonia, Tatari 51, 10134 Tallinn https://www.stat.ee/publication-download-pdf?publication_id=12345

Bias removal through sampling in machine learning models

Luis Sanguiao Sande
Spanish NSI (INE)

Abstract

It is well known that machine learning models have some bias, consequence of the bias-variance tradeoff and the optimization of the square mean error. For aggregates of the output variable(s), a probabilistic sample can be used to correct the bias, but we need this second sample in addition to the training sample. We propose an estimator that uses just one probabilistic sample both for modelling and bias removal. Two examples show that bias is indeed removed. In one of the examples the variance increases notably (this increase is almost exactly compensated by the bias removal) but in the other one, it unexpectedly decreases. This suggests that this kind of methods might be useful to combine with machine learning algorithms when used to estimate aggregates of predicted variables.

Keywords

Machine learning; bias correction; sampling; random forest; bias-variance tradeoff

1. Introduction

Suppose we have a finite population $\mathcal{P} = \{u_{1,2,\dots}, u_N\}$, a set of features for each unit $\mathcal{A} = \{a_{1,\dots}, a_N\}$ and we want to model some variable x. Let $= \{j_{1,\dots}, j_n\}$ be a sample with known sampling design P, and where x is supposed to be known. A machine learning algorithm M maps any sample to a function $\hat{x}_i = M(\mathcal{S})(a_i)$ for each $a_i \in \mathcal{A}$. If the predictors are known, we can estimate the totals of x as

$$X = \sum_{I=1}^{N} xi \cong \sum_{i\epsilon\mathcal{S}} x_i + \sum_{i\notin\mathcal{S}} \hat{x}_i$$

It might be a very good prediction, but it is biased because of the model. If we are reasonably sure that x will not change, we can sample once again the population and obtain an unbiased estimation of the bias, and thus an unbiased estimation of X. But x might have changed or the costs of a second sampling might be too expensive. Another option, closer to our line, would be to use the GREG estimator [6], but it is unbiased only asymptotically.

The method proposed for bias removal, inspired in cross-validation [4], divides the original sample into two subsets (equivalent to training and validation sets) and uses the first one for modeling and the second one for

bias removal. A weighted mean is taken over all possible divisions of the sample, so that the estimator becomes design unbiased. We get the weights from what we will call a two stage decomposition.

Definition 1: Let P_1, P_2 a two stage sampling design, where P_2 depends on the first stage sample denoted by $\mathcal{S}_1$. $(P_{1,2})$ is said to be a two stage decomposition of P if and only if

$$P(\mathcal{S}) = \sum_{\mathcal{S}_1 \subset \mathcal{S}} P_1(S_1) P_2(S \backslash S_1)$$

for any sample $\mathcal{S}$.

Suppose P is just simple random sampling of size n. An example of two stage decomposition is a simple random sampling of size $n - 1$ and a simple random sampling of size 1 on the remaining units. For simplicity, this is the decomposition we are going to use in the examples. Of course, the decomposition is not unique, even if we fix P and the sample size of both P_1 and P_2 . The optimal choice of a decomposition is still an open problem.

The first sample $\mathcal{S}1$ will be used for modeling and the second one $\mathcal{S} \backslash \mathcal{S}_1$ for Horvitz-Thompson estimation [3] of the difference between the model and the target variable.

In the examples, the machine learning algorithm used is random forest [2], because combined with simple random sampling a simpler, approximate version of the estimator can be used [5]. In both examples we extract 10000 samples of the population and the target variable is estimated with and without bias correction. The first population was generated with synthetic data, and unexpectedly the bias removal causes a decrease in the variance. The second one uses real data, but the population is not the real one but a small subsample. This time we have a variance increase, but the increase in the square mean error is barely noticeable.

In both cases the bias removal seems to be useful: in the first one we are at the same time decreasing the variance and in the second one we are eliminating the bias at almost no cost on the square mean error. Of course, some questions arise. Is there an optimum two stage decomposition? When should we expect a variance decrease and when an increase? When should we expect an important increase of the square mean error? We have no definitive answers to these questions yet, but some ideas that might help will be discussed.

2. Methodology

What follows is more extensively explained in [5]. Suppose we have a P_2 based estimator $\hat{X}_{\mathcal{S}1}$ of X. We use the subscript because P_2 (and thus the estimator) depends on the sample $\mathcal{S}_1$.

Definition 2: The second stage based estimator is given by

$$\hat{X}2 = \sum_{\mathcal{S}_1 \subset \mathcal{S}} \hat{X}_{\mathcal{S}1} \frac{P_1(S_1)P_2(\mathcal{S}\backslash\mathcal{S}_1)}{P(\mathcal{S})}$$

The second stage based estimator can be used to build unbiased estimators according to the following proposition.

Proposition A: If the estimator $\hat{X}$ $\mathcal{S}1$ is $P2$ unbiased, then the second stage based estimator is P unbiased.

Proof. We have to proof that $E_p(\hat{X}_2) = X$ so

$$\begin{aligned} E_P(\hat{X}_2) &= \sum_{\mathcal{S}} \sum_{\mathcal{S}_1 \subset \mathcal{S}} \hat{X}_{\mathcal{S}_1} \frac{P_1(\mathcal{S}_1)P_2(\mathcal{S}\backslash\mathcal{S}_1)}{P(\mathcal{S})} P(\mathcal{S}) \\ &= \sum_{\mathcal{S}} \sum_{\mathcal{S}_1 \subset \mathcal{S}} \hat{X}_{\mathcal{S}_1} P_1(\mathcal{S}_1)P_2(\mathcal{S}\backslash\mathcal{S}_1) \\ &= \sum_{\mathcal{S}_1} \sum_{\mathcal{S}_2 \subset \mathcal{P}\backslash\mathcal{S}_1} \hat{X}_{\mathcal{S}_1} P_1(\mathcal{S}_1)P_2(\mathcal{S}_2) \end{aligned}$$

where in the last equality we are just changing the order of the summands, grouping by $\mathcal{S}_1$. But now, we can move the factor $P_1(\mathcal{S}_1)$ outside the second summation, so finally

$$\begin{aligned} E_P(\hat{X}_2) &= \sum_{\mathcal{S}_1} P_1(\mathcal{S}_1) \sum_{\mathcal{S}_2 \subset \mathcal{P}\backslash\mathcal{S}_1} \hat{X}_{\mathcal{S}_1} P_2(\mathcal{S}_2) \\ &= \sum_{\mathcal{S}_1} P_1(\mathcal{S}_1) X = X \end{aligned}$$

since $\hat{X}_{\mathcal{S}1}$ is P_2 unbiased and $\sum_{\mathcal{S}1} P_1(\mathcal{S}_1) = 1$.

Now it is easy to build unbiased estimators. The sample $\mathcal{S}_1$ can be used to build the function $M(\mathcal{S}_1)$ and we can use a second stage Horvitz-Thompson to get an unbiased estimation of the sum of the model errors. Since the predictions are known for the whole population and the target variable is known for $\mathcal{S}_1$ it is trivial to get an unbiased estimator of X.

Under simple random sampling with the decomposition mentioned in the previous section, the expression of the estimator would be

$$\hat{X}_2 = \frac{1}{n} \sum_{i \in \mathcal{S}} \left(\sum_{j \in \mathcal{S}\backslash\{i\}} x_j + \sum_{j \notin \mathcal{S}\backslash\{i\}} M(\mathcal{S}\backslash\{i\})(a_j) + (N - n + 1)\left(x_i - M(\mathcal{S}\backslash\{i\})(a_j)\right) \right)$$

The first summand inside the parenthesis is the sum of the training set, and we have to add it because we want to estimate X and we do not have those elements at second stage. The second summand is the synthetic estimator for the totals of x on the second stage population. The third one is the second stage Horvitz-Thompson estimator of the difference between the totals of x and $\hat{x}$, once again on the second stage population. Therefore, the three summands compound an unbiased second stage estimator of X, and by Proposition A,$\hat{X}_2$ is unbiased for the sampling design P.

Note that we have to fit a lot of models to build these estimators: even if $\mathcal{S}_2$ contains just one element, we will have to fit n models! So second stage estimators are computationally expensive and we lack a fast specific software

implementation. Fortunately, for bootstrap aggregated [1] algorithms (like random forests) and simple sampling designs, we get an approximated version based on the out of bag predictions.

Theorem B: Let $\hat{x}_i = M(\mathcal{S})(a_i)$ the predictions when $i \notin \mathcal{S}$ and the out of bag predictions when $i \in \mathcal{S}$. Let $\hat{e}_i$ be the out of bag errors. Under simple (possibly stratified) design, the estimator

$$\sum_{i=1}^{N} \hat{x}_i + \sum_{i \in \mathcal{S}} \frac{\hat{e}_i}{\pi_i}$$

is an approximation of an unbiased second stage based estimator for X.

Proof. See [5].

This way the estimator is expressed as the sum of a purely model based expression and a sampling based estimation of the bias. Note that this result is also kind of a confirmation that the out of bag errors are a good indicator of the performance of a bootstrap aggregated algorithm. An unbiased estimator for the variance is also known.

Proposition C: Let $\hat{V}_{\mathcal{S}1}$ be a P_2 unbiased estimator for the variance of $\hat{X}_{\mathcal{S}1}$. A P unbiased estimator of the variance of $\hat{X}_2$ is

$$\hat{V}_2 = \sum_{\mathcal{S}_1 \subset \mathcal{S}} \hat{V}_{\mathcal{S}_1} \frac{P_1(\mathcal{S}_1)P_2(\mathcal{S}\backslash\mathcal{S}_1)}{P(\mathcal{S})} - \sum_{\mathcal{S}_1 \subset \mathcal{S}} \left(\hat{X}_{\mathcal{S}_1} - \hat{X}_2\right)^2 \frac{P_1(\mathcal{S}_1)P_2(\mathcal{S}\backslash\mathcal{S}_1)}{P(\mathcal{S})}$$

Proof. See [5].

Note that if we want to build the estimator $\hat{V}_2$ we have to be able to estimate the variance of $\hat{X}_{\mathcal{S}1}$. Thus measurable sampling design is required at second stage in our two stage decomposition. In the examples only one unit is sampled in second stage, so there is no way we can build the estimator. If we wanted to estimate the variance, we should take $n - 2$ elements at stage one and 2 elements at stage two. For random forest, out of bag predictions excluding two elements would be needed, but we do not know any piece of software that provides such predictions.

3. Result

We are comparing the estimator from Theorem B with the pure model based estimator in two very different populations. The first one is based on synthetic data that is constructed to hold a (noisy) equality, and the second one is based on real data (and therefore, it holds no equality). In both cases a big number of samples (10000) are taken, and the estimations are compared to the real (known) aggregated values of target variables. The bias is estimated as the mean of the 10000 estimations minus the real totals of the target variable. The variance is estimated as the variance of the 10000 estimations and the square mean error is estimated as the mean of the square of the difference between each estimation and the real totals.

For the random forest modelling the R package *ranger* was used. The election of this particular package was because it is quite fast (multithreaded) and returns the out of bag predictions that we need to correct the bias.

Synthetic data: The predictors of the population are variables A, B and C. A and B follow uniform distributions (0, 10) and (0, 3) while C is a (0, 1) normal. The target variable is D = AB + C + noise. The noise once again follows a (0, 1) normal. The size of this population is 1000, and the sample size 80. Next we have a summary of the results:

	Variable totals	**Mean of the estimations**	**Estimated bias**	**Estimated variance**	**Estimated SME**
Synthetic estimator	6977.088	6902.723	-74.37	111834.7	117353.8
Bias-corrected estimator	6977.088	6974.815	-2.27	110817.4	110811.5

We see that the bias is effectively removed, with a noticeable impact in the square mean error. But the variance of the estimator is also smaller! In machine learning this might seem sound strange, because of the bias-variance tradeoff. What happens is that there is some additional information we are exploding here: the sampling design, which machine learning algorithms usually ignore.

It could have been thought that the small difference is not significant, but after a rerun with sample size 160 the variances are 32577.72 and 31603.83 being again the bias-corrected the lesser. It seems that the difference increases with sample size.

Real data: This time the population is created from a subset of the SBS survey sample. The predictors are 40 variables from corporate tax and the target variable is Total Personnel Expenses. None of the predictors reflects the target variable concept. This time the population size is 47068 and the sample size is 1000. The summary is

	Variable totals	**Mean of the estimations**	**Estimated bias**	**Estimated variance**	**Estimated SME**
Synthetic estimator	23.73×10^6	23.28×10^6	-0.45×10^6	1.41×10^{11}	3.74×10^{11}
Bias-corrected estimator	23.73×10^6	23.73×10^6	-2.05×10^3	3.88×10^{11}	3.88×10^{11}

Once again, the bias is removed quite well, but in this case there is an important increase in the variance. The increase in the square mean error is smaller though, so it might be better to have unbiased estimations in exchange for a slightly bigger square mean error.

4. Discussion and Conclusion

According to the preceding simulations, the bias correction works, and it might come at small cost or even with an improvement of the square mean error, so it is a tool to consider for machine learning based aggregated

estimations. The increase of the square mean error might be higher though, so it would be a good thing to know when this is going to happen. We do not have the answer to this question, but some ideas that might help.

First of all, we are discarding one element (out of bag) of the sample for each model. This decreases the efficiency of the machine learning algorithm, because the training set is smaller. The higher the sample size the lesser this effect, so this might lead to a faster improvement of the bias-corrected estimator while increasing sample size.

Since estimators for the variance of the second stage based estimator are known, the variance might be estimated. The variance of the synthetic estimator is more difficult to estimate, but a bootstrap approximation might be used. However, it is difficult to say whether a comparison of estimated variances would solve the problem though.

Observe that the first summand in Proposition C gives us an upper bound for the variance, because the subtrahend is always positive. But the expectation of this first summand is the mean of the variances of the difference between the target variable and its predictions (excluding the training set). So, the better the algorithm models, the lesser the variance is. Even if it happens for a small set of samples, overfitting might lead to big variances, while it could be slightly better for the synthetic estimator, because of error cancellation.

About the two stage decomposition it is a difficult choice and further research would be needed. If we want unbiased estimation and the same for variance, the second stage have to have a measurable sampling design. This usually will mean that the second sample will be of at least size two for each stratum. For cluster sampling the decomposition is going to be a little more complicated and probably a minimum of two clusters should be selected at second stage. As it has already been stated, the bigger the second stage sample size, the lesser the effective training set of the model, so we should keep the second stage as smaller as possible.

The decomposition might be chosen so the weights are equal for the second stage based estimator [5], but there is no reason why this has to be better than any other choice (except perhaps for calculations). Note that if we select the weights on sample information, the estimator might become biased.

There exists also the problem of the computational cost, although it would be not so expensive in production. However, if we want to test other algorithms with simulations like these, fast specialized software has to be written before. Fortunately, the problem is embarrassingly parallel so it should be easy to adapt the simulations to run in a cluster for greater speed.

References

1. Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**(2): 123–140.
2. Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.
3. Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**: 663–685.
4. Picard, R. and Cook, D. (1984). Cross-Validation of Regression Models, *Journal of the American Statistical Association* **79** (387): 575–583.
5. Sanguiao (2018). A design unbiased model assisted estimator for finite populations. Unpublished.
6. Sarndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*, Springer, New York.

Malaysia's silver tsunami: Preparing for the impacts of population ageing

Ezatul Nisha Abdul Rahman, Noor Faadlilah Ismail, Wan Hazlin Ezrina Wan Hamat, Filisa Mama, Addri Rahman
Population and Demographics Statistics Division, Department of Statistics Malaysia

Abstract
Malaysia is facing a Silver Tsunami in which is the population is ageing. Population ageing can be defined as the demographic transition due to decreasing birth rate and death rate as the country moves away from the pre-industrial/agricultural-based economy to an industrial economic system. This phenomenon is caused by lower birth rate and decreasing total fertility rate as well as increasing life expectancy. This paper involves the existing statistical data obtained from the Department of Statistics Malaysia (DOSM), UN (United Nations) and other related agencies. Based on Malaysia Population Projection, Malaysia will become an ageing state in 2030, of which 15.3% of the population is 60 years and above. As a result, changes would occur in the social structure and increasing need for long-term healthcare. The paper ends with a short summary and a reflection on the need for further study.

Keywords
Demographic transition, Population pyramid, ageing, life expectancy, fertility

1. Introduction

As a developing nation, Malaysia moving away from an agricultural based economy to an industrial economic system and transiting into an ageing population faster than expected (Samad & Mansor, 2017). The global population aged for those above 60 years old was estimated to 962 million which is more than twice bigger as in 1980 with 382 million older persons worldwide. In addition, by 2050, the global population is projected to be doubled again to reach nearly 2.1 billion, comprising 13 per cent (13%). The global trends of the number of older persons aged above 80 years old are projected to increase more than threefold between 2017 and 2050 in which rising from 137 million to 425 million (United Nations, 2017).

Develop nation such as the USA, Japan, Australia and Singapore also will be experiencing such phenomenon and early preparation had been undertaken to cater the phenomenon (Anderson & Hussey 2000). A country such as Australia now starting to recognised foreign talent and encourage them to migrate under the skills shortage programme (Department of Home Affairs, Australia, 2019).

In Japan, there are declining numbers of working age population (United Nations, 2017), to cater the shortage; Japan has been equipping its nation powerhouse by encouraging the development of modern machinery with IT to substitute human manual labour shortage. By going fully automation integrating modern machinery with IT to substitute human labour in factories, the nation could increase labour productivity and keep production level at the market demand (Otsu & Shibayama 2016). The decreasing working age population will end in decreasing numbers of skilled workers and lessen the amounts of gross wages, thus reducing the number of commodities needed. Another phenomenon in Japan is the senior citizens who own 83% of $14 trillion personal financial assets do not consume as many materials as youths do, but just enjoy saving (Otsu & Shibayama, 2016).

Singapore is one of the most rapidly ageing countries in the world. The proportion of Singaporeans aged 65 and above is projected to more than double from 8% in 2005 to 20% in 2030, and by 2050, 38% of Singaporeans will be aged 60 and above (Kwok 2006). The rapid rate of ageing in Singapore is driven by two demographic trends, a rapidly declining in the Total Fertility Rate (TFR) and an increasing life expectancy (Teo et al., 2006).

Malaysia will be at the stage of aged society in 2030 where elder persons projected to be 15.3% of Malaysia's total population. Malaysia experiences the rapid stage of population shift in which is onlytakes for 30 years to be at staged of aged society from an ageing society in 2010 with seven per cent of elder persons from the total population. Compare to other countries, France taking for 115 years to be at stages of the ageing population. Compare to other countries, France had almost 150 years to adapt to a change from 10% to 20% in the proportion of the population that was older than 60 years, places such as Brazil, China and India will have slightly more than 20 years to make the same adaptation (World Health Organization, 2016).

This rapid demographic transition will have major implications for changing epidemiological patterns in Malaysia which impacting broader economic development and the health workforce (Atun et.al, 2016). Hence, the purpose of this study is to examine the impact of ageing population and how Malaysia preparing the implication of ageing population. Furthermore, this paper carried three objective which is first is to explore the cause of population ageing in Malaysia. Secondly, this study examines the impact of population ageing in economy and lastly to examine the impact of population ageing in healthcare in Malaysia.

2. Methodology

In this study, population dataset from Malaysia and other selected country such as Japan, USA, Australia and Singapore from 1950 to 2050 are obtained from the dataset from the United Nations. In addition, cohort- component projection method is applied for estimating population projection after the year of 2018 (UN Population Projections: Methodology and key Assumptions).

The key areas of consideration that will be taken into account are the population distribution of the selected country by age group of young age from 0 to 14, working age from 15- 59 and aged from the age of 60 years of age till above. The results are presented in population distribution by the group as it will reflect in the population pyramid of Malaysia. This study also will focus on the working group at the age of 14-59 as they are the working group that will affect the contribution in the form of taxation and future policy of the nation.

3. Result

The result is obtained from the projection of demographic data by age group gain from United Nations Dataset. This is to show the trend of demographic transition by age group of Malaysia from post war to 2050. There are also changes in the population trend as the younger generation is going to be slowly dipping, and the older age group is staying health extending their life.

3.1 Age Structure

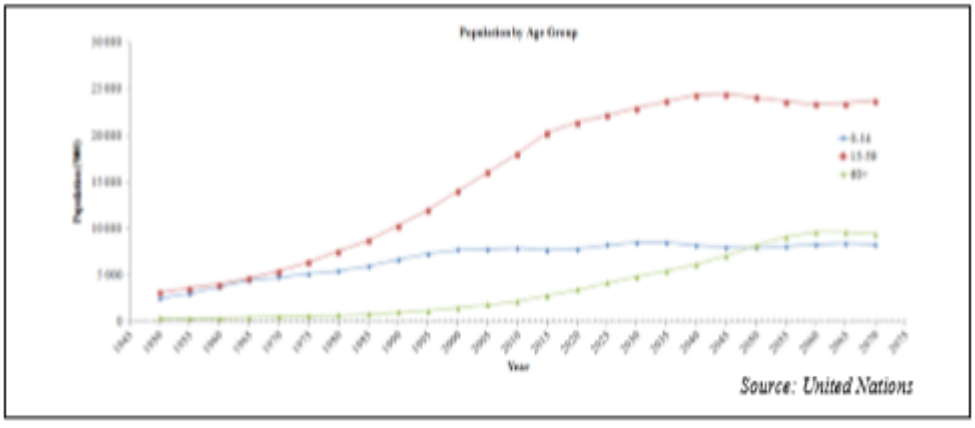


Figure 1: Population by Age Group in Malaysia

The numbers of the ageing population will slowly increase while the young age group (0-14) will slowly decrease and will be surpassed by the old age group by the year 2050 (Figure 1). A country such as Japan, USA, Australia and Singapore had reached that intersection way before Malaysia. Japan in 1990 had reached the point where the population of aged is slowly surpassing the young age group.

Singapore and the USA had reached in 2015 while Australia in 2010, most of this country had programme and policy in place to prepare for the inevitable.

Table 1: Highest Percentage of Age Group from 1950- 2050

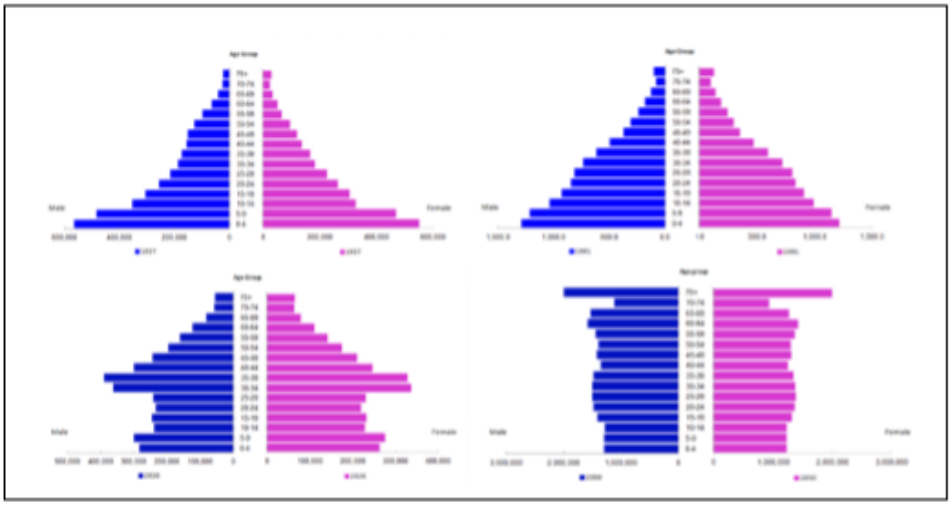
Population by Age Group (%), Highest (Year), United Nations

	0-14	15-59	60 Above
Japan	34% (1955)	64% (1995)	37% (2035)
USA	19% (2030)	60% (2015)	27% (2050)
Australia	19% (2030)	61% (2015)	30% (2050)
Malaysia*	45% (1970)	66% (2013- 2019)	27% (2050)
Singapore	19% (2030)	67% (2015)	33% (2045)

Sources: United Nations and *DOSM

Leading nation in economic, such as Japan had reached the highest percentage growth of the young age group in the year 1955 and 34% of that group contributed from the total population and in 2035, 60 and above age group will be the highest at 37% from the total population that year. Malaysia however, has the highest young age group in the year 1970 at 45% among the highest compared to the age group at 27% in the year of 2050 (Table 1). Malaysia has a stable growth of the working group from the year 2013- 2019.

3.2 Changes in Malaysia Population Pyramid



Sources: United Nations and *DOSM

Figure 2: Changes in Malaysia Population Pyramid in Year 1957, 1991, 2020 & 2050

The changes of the young aged group will reflect the population pyramid structure, with less percentage of young population occupy in the social structure of the society, the numbers of next generation age group will slowly be reduced thus changing the landscape of the population pyramid (Figure 2).

4. Discussion

In this discussion there are three sections, first is the impact of population ageing in economy and healthcare. Then, the discussion on the initiatives to prepare for population ageing that has been implemented both in Malaysia and abroad. The third section is the recommendation that can be applied in preparing for population ageing.

4.1 The Impact on Economy

With the changes in the structure of the population pyramid, the working force is gradually shrinking and less tax could be collected from this age group, thus will lead to lower tax collected by government indirectly will restrict the national spending. In Japan, USA and Singapore the mandatory age of retirement is put at 60 years, 62- 65, and 62 respectively (Kashiwagi 2018). By increasing the retirement age, the active labour forces are at a level where the government could more collect a tax.

Under 11th Malaysia Plan which will be implemented in 2016 to 2020, low-income groups especially in the informal sector, they will be encouraged to participate in voluntary savings and retirement scheme as a guarantee of economic protection. Few more initiatives for the elderly under Budget 2019 such as pensioner who still receives pensions less than RM1000 will get a donation of RM500 from the government. Second, to encourage employment opportunities for over the age of 60, the government proposed that EPF's mandatory contribution reduced from six percent to four percent (Ministry of Finance, 2018).

4.2 The Impact on Healthcare

The health system in Malaysia continues to struggle in catering to the population with comprehensive care. This can be observed through the high number of admissions due to chronic conditions involving a patient with asthma, diabetes mellitus or any other Non-Communicable Diseases (NCDs) (Atun et.al, 2016). Around 15-20 percent of hospital admissions involved conditions that should be effectively managed through ambulatory care which reflect the suboptimal performance of the health system as a whole. These admissions include suboptimal continuity of care between primary, secondary, and tertiary levels. However, the current healthcare system in Malaysia is not well-suited for treating agerelated symptoms (Ministry of Health Report, 2016). In addition, the population ageing drives the implication of the need

for long-term geriatric healthcare. The goal of geriatric care for the elderly is not necessarily to cure but to increase their healthy year of life. However, there is still a lack of access to excellence geriatric healthcare and support system and, the reward for geriatric is not that attractive compare to other specialities (Jacob, 2016).

As the population grows older, the expenditure for healthcare will increase and Malaysia has to revised its expenditure on elderly healthcare. In 1970, healthcare spending was only 2.7 percent of the GDP. Technological change, demographic and epidemiological change, and rising income contribute to the increasing cost and expenditure. This will continue to grow in future and further increasing the national expenditures (Ministry of Health Report, 2016).

4.3 Others Country Initiatives and Practices on Healthcare

In 2002, MIPAA has launched the Second World Assembly on Ageing; The International Plan of Action and Ageing focusing on the changes in policies, attitudes and practices at all level in all sectors. Its aim to ensure persons everywhere are able to age with security and dignity and to continue to participate in their society as citizens with full rights. This plan can be used as guidance for planning and implementing the healthcare system for elder persons in Malaysia.

A progressive initiative such as long-term insurance for all citizens was introduced by Japan. The Health and Welfare Bureau for the Elderly, Ministry of Health Japan launched the Long-Term Care Insurance System (LTCI) in 2016. This plan also aimed to support the independence of elderly people rather than simply providing personal care. The most popular LTCI program provides service rather than cash for care. The most popular service is adult day care, with 1.9 million users, benefiting both older people and their careers. Future, Japan will establish the 'Community-based Integrated Care System' by 2025 which focusing on the provision of health care, prevention, housing, nursing care and livelihood support for elder persons.

Singapore, in 2002 a health care system for elder person comprising of financial support, care services and care-giving which includes the Elder Shield; long-term care insurance scheme targeted at severe disability, especially during old age, Senior Mobility and Enabling Fund (SMF); support caregivers in caring for seniors at home, Elder Fund; new assistance that assisting at severely disabled lower-income age 30 and above who are not eligible for other insurance program and Assistance Programme for The Elderly (IDAPE); is a government scheme providing financial help to needy and disabled elderly Singaporean who are not eligible to join Elder Shield (Agency for Integrated Care Singapore)

In Hong Kong, under Legislative Council Panel on Welfare Services (2018) which concerning the elderly comprising ten initiatives such as enhancing

dementia care, implementing Pilot Scheme on Residential Care Service Voucher for Elderly, implementing the Opportunities for the Elderly Project (OEP) and Elder Academy (EA) Scheme, implementing the Government Public Transport Fare Concession Scheme for the Elderly and Eligible Persons with Disabilities and many more are introduced.

4.4 Malaysia Initiative and Practices on Healthcare

The aging population is inevitable as it experiencing a demographic and epidemiologic transition among the elderly. Hence, there is a need for health policy on ageing elderly (Rose Jacobs, 2016). In Malaysia, the Ministry of Women, Family and Community Development (MWFCD) developed two policy focusing on older persons in Malaysia; The National Policy for Older Persons and Plan of Action for Older Persons which approved by the Government on 2011. These policies are proof of the government's commitment to producing older persons who are independent, honourable and respected by optimising self-potential through healthy ageing, positive, active, productive and supportive according to the national development.

The National Policy for Older Persons aimed to provide effective and efficient services for the individuals, family and society to assure that older persons receive facilitative environment while Health National Policy for Elder Persons helps the initiative to be more effective, coordinate and comprehensive health care. Furthermore, under the 11th Malaysia Plan (2016 - 2020), the government has announced the initiatives for elderly, Strategic B5; to improve the environment of senior citizens. Under this initiative, NGO's will establish cooperation with elderly care centre and social protection for poor elderly will be coordinated and integrated to assure a better quality of life. Apart from that, the awareness programs on senior care also will be strengthened to make senior citizens practice an active and healthy lifestyle.

5. Recommendation and Conclusion

The current initiative can be further be improved by the government for the preparation on facing the population ageing in Malaysia. Creating a less physically demanding job and such as promoting job in the consulting level could encourage the aged population to work even after the age of retirement.

In terms of healthcare, first, while sustainably growing towards a greater light, the elderly population should be prioritized on the elderly to ensure the elderly lives a healthy and fulfilling life. Second, the policies for elderly should be improved in terms of short-term and long-term healthcare and infrastructure for elderly with the consideration of the population dynamics in terms of size, distribution and location, especially in rural areas. Other than that, development programs on the elderly can be done by focusing on monitoring specific needs of the age especially for citizens aged 60 years and

above. For example, promote volunteering among the elderly to improve social engagement, physical health and mental. Furthermore, to have an increase of specialists in geriatric in the healthcare profession to evaluate and manage the unique healthcare needs and treatment preference for elder especially who have the most complicated medical and social problems.

Conclusion, this paper covers the Malaysian scenario of ageing trend and the implication however there more room to explore as to relate the fiscal and economic perspective related to the national.

References

1. Samad, S. A., & Mansor, N. (2017). Population ageing and social protection in Malaysia. Malaysian Journal of Economic Studies, 50(2), 139-156.
2. Otsu, K., & Shibayama, K. (2016). Population Aging and Potential Growth in Asia. Asian Development Review, 33(2), 56-73.
3. United Nations. (2018). Ageing. Retrieved January 17, 2019, from http://www.un.org/en/sections/issues-depth/ageing/
4. Anderson, G. F., & Hussey, P. S. (2000). Population aging: a comparison among industrialized countries. Health affairs, 19(3), 191-203.
5. Department of Jobs and Small Business, A. G. (2018.). National, state and territory skill shortage information. Retrieved January 20, 2019, from https://www.jobs.gov.au/national-state-and-territoryskill-shortage-information
6. World Population Ageing 2017 Highlights. (2018). Statistical Papers - United Nations (Ser. A), Population and Vital Statistics Report.
7. SCHRÖDER-BUTTERFILL, E. (2007). Peggy Teo, Kalyani Mehta, Leng Leng Thang and Angelique Chan, Ageing in Singapore: Service Needs and the State, Routledge, London, 2006, 192 pp., hbk £65.00, ISBN 9780415374873. Ageing and Society, 27(3), 450-452.
8. Agency for Integrated Care Singapore, Ministry of Health Singapore (2002). https://www.moh.gov.sg/cost-financing/healthcare-schemes-subsidies
9. W. (2017, June 12). World report on ageing and health 2015. Retrieved January 20, 2019, from https://www.who.int/ageing/events/world-report-2015-launch/en/
10. Atun, R., W.A. Yap, and K. Shen Lim, MHSR Report on Health Information System: Moving Beyond a Data System to a Health Information and Intelligence System. 2016, Ministry of Health in Malaysia and Harvard T.H. Chan School of Public Health.
11. Dallin Jack, "The Issue of Japan's Aging Population," Law School International Immersion Program Papers, No. 8 (2016)

12. Eleventh Malaysia Plan 2016-2020: Anchoring Growth on People 2015: Economic Planning Unit, Prime Minister's Department, Putrajaya, Malaysia.
13. Jacob, R. (2016). Aging and Current Trends in Malaysia. International Journal of Social Work and Human Services Practice, Volume (4 No. 3), pp. Page(57- 61). Retrieved from http://www.hrpub.org/download/20160830/IJRH1-19206106.pdf
14. Legislative Council Panel on Welfare Services 2018 Policy Address Policy Initiatives of the Home Affairs Bureau, LC Paper No. CB(2)30/18-19(02)
15. Long-Term Care Insurance System of Japan, Health and Welfare Bureau for the Elderly, Ministry of Health, Labour and Welfare (November 2016)
16. Madrid International Plan of Action on Ageing (2002), Second World Assembly for Ageing, Madrid, Spain Declaration, M. P. International Plan of Action on Ageing 2002. Online: http://www. un. org/esa/socdev/ageing/waa/index. html.
17. National Center for Policy Analysis, "Raising Taxes on the Wealthy Would Hurt the Economy" (2015) available at http://www.ncpa.org/sub/dpd/index.php?Article_ID=25285 (last accessed June 15, 2016). 17 Akihiko Kato
18. World Population Ageing 2017 Highlights. (2018). Statistical Papers - United Nations (Ser. A), Population and Vital Statistics Report.
19. Teo, P., Mehta, K., Thang, L. L., & Chan, A. (2006). Ageing in Singapore: Service needs and the state. Routledge.
20. Kwok, Andrew. (2006). Ageing and public policy – A global perspective. Ethos, Issue I, 11-15. Retrieved November 10, 2012, from http://www.cscollege.gov. sg/Knowledge/Ethos/Issue%201%20Oct%202006/Pages/default.aspx
21. Skilled Independent visa. (2018, November 29). Retrieved January 15, 2019, from https://immi.homeaffairs.gov.au/visas/getting-a-visa/visa-listing/skilled-independent-189/pointstested#Overview
22. Malaysia Budget (2019). Secretary-General of the Treasury, Ministry of Finance Malaysia, Putrajaya Malaysia.
23. Kashiwagi, S. (2018, May 07). Japan must abolish mandatory retirement. Retrieved January 20, 2019, from https://asia.nikkei.com/Opinion/Japan-must-abolish-mandatory-retirement2

Economic determinants of import demand for rubber latex products: an econometric analysis for top 4 rubber consuming countries

Aye Aye Khin, Raymond Ling Leh Bin, Yogambigai a/p Rajamoorthy, Evelyn Mok Jo-yee, Chai Jie Si

Faculty of Accountancy & Management (FAM), Universiti Tunku Abdul Rahman (UTAR), Jalan Sungai Long, Bandar Sungai Long, Cheras, 43000 Kajang, Selangor, MALAYSIA

Abstract

The research study is to investigate the economic determinants that are being used to analyze on the import demand for rubber latex products of the top 4 rubber consuming countries such as China, India, USA and Japan. Each import demand model has specifications for world natural rubber (NR) price, exchange rate and domestic NR export price. The models used the econometric analysis such as Vector Error Correction Method (VECM) equation with co-integration analysis and granger causality test. Monthly data from January 2004 to December 2016 were used as an estimation period. The results for VECM are in line with the hypothesis development, where the world NR price, exchange rate and domestic export price have negative relationship with the import demand for China, India, USA and Japan. Furthermore, normality test and heteroskedasticity test are used to examine the model in order to portray an accurate and precise of the models. The findings of this study are able to provide information to fill the gap by determine the import volume on current and potential market such as China, India, USA and Japan and also market participants in their consumptions and financing decisions due to NR is an important commodity for world market.

Keywords

Import Demand; Rubber Latex Products; VECM; Econometric Analysis; Top 4 Rubber Consuming Countries

1. Introduction

The rubber latex is obtained from the bark of the rubber-tree through the trapping process. The transformation of the raw latex is able to manufacture more than 50,000 types of rubber products. The inclusion of the many types of rubber products are such as natural rubber is can be used for making medical gloves, tyres, inner tubes, catheters, footwear, rubber bands, rubber sheets, condoms and so on. After the products has been manufactured and packaged, it will be exported to other countries worldwide (Hnin, 2017).

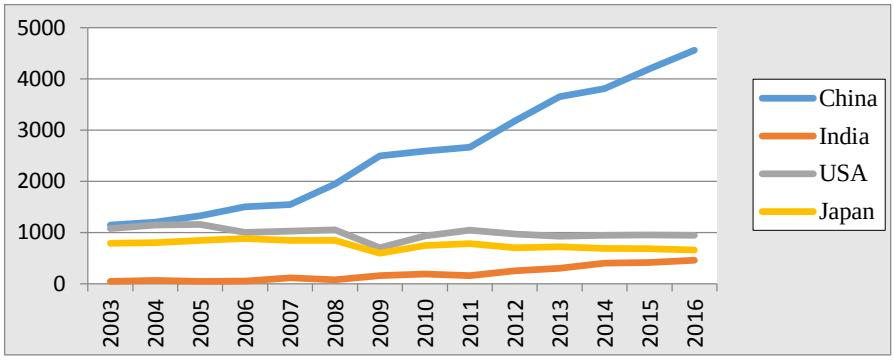
In the past few years, the world NR industry has grown rapidly and changes fundamentally with appearance of many competitors bring latest technology or product into the market. Indian rubber industry has also been growing in

along with the strength and importance, as a part of India's burgeoning role in the global economy. India is the fourth world largest' producer, second largest consumer of NR and also one of the fastest growing economies globally in 2017. There were around 4,600 registered units are manufacturing rubber products with contributing 40 billion Rupees to the government economies. The government of India offered funds to research and development centres to support the rubber industries and end users to improve the quality of rubber products. Government of India also promoted the development of rubber industry by offering manufacturing facilities and technology knowledge (Sabu, 2017).

Demand of natural rubber has grown at a fast pace in past few years in China, both supply and demand will continue to grow in next decade. Global and China NR Industry Report (2017) stated that China was also one of the world's largest producers and consumers of NR, and 77.9% of rubber was used for redial tyres. Production of NR in China was about 764,000 tonnes in 2016, it was around 6.2% of the global output in the market. China NR manufacturers are mainly large-sized agricultural reclamation and rubber groups, represented by Sinochem International, China Hainan Rubber, Guangken Rubber and Yunnan State Farms Group. According to Freedonia Focus Reports Rubber in United States (2017) stated that the demand of rubber in USA forecast to reach 8.9 billion USD in 2021. Main of the demand of NR of USA was needed from tyre manufacturing and re-treading represents, both of the sectors was accounting around three-fifths of domestic rubber consumption. Malaysia Rubber Export Promotion Council has reported that the biggest markets are USA, Germany and Japan for the rubber products from Malaysia. It indicates 28% for USA, 7% for Germany and 6% for Japan, which accounts for more than 40% of total exports of Malaysia's rubber products. China, UK, Brazil, and Australia are also those that impact the Malaysia's export of rubber products (MREPC, 2017).

The trends of the NR import demand of China, India, USA and Japan shows in Figure 1. As seen in the figure, China is the largest NR importer as compared to other three countries. China's NR import demand as been inclining since year 2003 up until year 2016 making them the largest NR importer as compared to India, USA and Japan. India is the lowest NR importer as seen in the graph. According to Kannan (2013), he said that India's share in the production of NR was increasing over the years making them the fourth largest NR producer after Malaysia, Thailand and Indonesia. USA and Japan has a neutral import demand of natural rubber, however, there was a sharp decline in 2009 due to the great recession which affecting the US trade substantially (Ravikumar *et al.*, 2017).

Figure 1: Natural Rubber Import Demand (000' tonnes) of China, India, USA and Japan



Source: International Rubber Study Group (IRSG), 2017

2. Methodology

Incorporation of more variables enables the import demand models for rubber latex products in the current and potential market of the top 4 rubber consuming countries of China, India, USA and Japan and display the VECM econometric models for those countries, which were derived based on the related factors, can be specified as follow:

China

$$\Delta CHnrimp_t = \beta 0 - \beta 1\ \Delta nrstr20_{t-1} - \beta 2\ \Delta exrm_{t-1} - \beta 3\ \Delta nrsmr20_{t-1} - \beta 4\ \Delta CHnrimp_{t-1} + \varepsilon t1 \quad (1)$$

India

$$\Delta INnrimp_t = \beta 5 - \beta 6\ \Delta nrstr20_{t-1} - \beta 7\ \Delta exrm_{t-1} - \beta 8\ \Delta nrsmr20_{t-1} - \beta 9\ \Delta INnrimp_{t-1} + \varepsilon t2 \quad (2)$$

USA

$$\Delta USnrimp_t = \beta 10 - \beta 11\ \Delta nrstr20_{t-1} - \beta 12\ \Delta exrm_{t-1} - \beta 13\ \Delta nrsmr20_{t-1} - \beta 14\ \Delta USnrimp_{t-1} + \varepsilon t3 \quad (3)$$

Japan

$$\Delta JPnrimp_t = \beta 15 - \beta 16\ \Delta nrstr20_{t-1} - \beta 17\ \Delta exrm_{t-1} - \beta 18\ \Delta nrsmr20_{t-1} - \beta 19\ \Delta JPnrimp_{t-1} + \varepsilon t4 \quad (4)$$

where,

$nrimp_t$= NR import latex volume of China, India, USA and Japan ('000 tonnes)

$nrsmr20_{t-1}$= NR export price Standard Malaysia Rubber Grade 20 (SMR20) (USD/ton) deflated by the CPI

$nrstr20_{t-1}$ = World NR price Singapore Commodity Exchange Market (SICOM) (USD/ton) deflated by the CPI

$exrm_{t-1}$= Real average exchange rate of the particular countries currency per USD

T = time trend of 2004 January to 2016 December monthly data

t and εt = time period and error terms respectively

2.1. Hypothesis Development

Ho1: There is no negative relationship between the world NR price and the import demand for rubber latex products of China, India, USA, and Japan.

Ha1: There is a negative relationship between the world NR price and the import demand for rubber latex products of China, India, USA, and Japan.

Ho2: There is no negative relationship between the exchange rate and the import demand for rubber latex products of China, India, USA, and Japan.

Ha2: There is a negative relationship between the exchange rate and the import demand for rubber latex products of China, India, USA, and Japan.

Ho3: There is no negative relationship between the domestic NR export price and the import demand for rubber latex products of China, India, USA, and Japan.

Ha3: There is a negative relationship between the domestic NR export price and the import demand for rubber latex products of China, India, USA, and Japan.

2.2 Data Collection and Sources of Data

The data collection period is ranged from January 2004 to December 2016. Data will be collected from Malaysian Rubber Board (MRB), International Rubber Study Group (IRSG), Malaysian Rubber Export Promotion Council (MREPC), Association of Natural Rubber Producing Countries (ANRPC) and Department of Statistics in Malaysia.

2.3 Unit-Root Test

According to (Studenmund, 2017), the unit-root test is used to check for stationary of the data series. The series variables are non-stationary, with mean and variance non constant (unit root). The null hypothesis Ho shows that the time series data is unit root (nonstationary) while alternative hypothesis Ha shows that the time series data is no unit root (stationary). There are two common unit root tests which are Augmented Dickey-Fuller (ADF) test and Phillip-Perron (PP) test. ADF test is used to check for random walk components in the residuals. PP test specifies the number of periods of serial correlation to include. Based on the unit-root test, all the data are 1^{st} difference stationary at the integrated in order 1 at ADF and PP test, i.e I (1) is at stationary.

2.4 Vector Error Correction Method (VECM) and Co-integration Test

A vector error correction method (VECM Model) is a restricted vector autoregression (VAR) designed for use with non-stationary series that is cointegrated. A VECM model includes a cointegration equation and VECM equations. The cointegration equation is built into the specification in order to restrict the long-term behaviour of the endogenous variables to converge to their cointegrating relationship. The VECM equations on the other hand are all endogenous variables while allowing for short-term adjustment dynamics (Studenmund, 2017).

2.5 Granger Causality

Granger causality is a circumstance in which one time series variable consistently and predictably changes before another variable. Granger causality is important as it allows us to analyse which variable precedes or leads the other variable where those leading variables are extremely useful in order to perform forecasting purposes (Studenmund, 2017). For example, a time series X is said to Granger-cause Y (X Granger-cause Y) when the F-test of X to Y p-value is significantly less than 0.05 level. Granger causality is called stationary and linear combination. It is also called cointegrated and long-term equilibrium relationship among the variables (Studenmund, 2017). Cointegration Analysis happens when there is a stable long-term relationship between two variables even though individually, each variable is nonstationary.

3. Result

3.1 Cointegration Equation

$$-0.1253CHnrimp_{t-2} - 0.6891nrstr20_{t-2} - 0.2000exrm_{t-2} - 0.6195nrsmr20_{t-2} = 0 \quad (5)$$
t-stat = [-14.3594***] [3.1980**] [0.3894 ns] [3.2024**]

$$-0.6985INnrimp_{t-1} - 1.2665nrstr20_{t-1} - 0.0003exrm_{t-1} - 0.9505nrsmr20_{t-1} = 0 \quad (6)$$
t-stat = [-12.2258***] [0.1212 ns] [-4.8910**] [-1.8443*]

$$-0.6065USnrimp_{t-1} - 0.2172nrstr20_{t-1} - 0.0010exrm_{t-1} - 0.3336nrsmr20_{t-1} = 0 \quad (7)$$
t-stat = [-20.1140***] [2.4838**] [-1.4806ns] [1.7022*]

$$-0.0555JPnrimp_{t-2} - 1.2228nrstr20_{t-2} - 0.0071exrm_{t-2} - 0.5598nrsmr20_{t-2} = 0 \quad (8)$$
t-stat = [-14.0862***] [-2.2752**] [-1.8146*] [-3.0182**]

3.2 Vector Error Correction Method (VECM)

$$\Delta CHnrimp_t = -0.3592 - 0.4255\Delta nrstr20_{t-1} - 0.7670\Delta exrm_{t-1} - 0.1628\Delta nrsmr20_{t-1}$$
t-stat = [-6.3237***] [-6.6072***] [-7.5822***] (9)
$$-1.0867\Delta CHnrimp_{t-1} + 1.0390\varepsilon t$$
[-6.6531***]

$R^2 = 0.8725$ Adjusted $R^2 = 0.8644$

$$\Delta INnrimp_t = -0.02237 - 0.01255\Delta nrstr20_{t-1} - 0.1901\Delta exrm_{t-1} - 0.0044\Delta nrsmr20_{t-1}$$
t-stat = [-5.4984***] [-4.2230**] [-6.0891***] (10)
$$- 0.1520\Delta INnrimp_{t-1} + 0.3293\varepsilon t$$
[-1.8520*]

$R^2 = 0.7180$ Adjusted $R^2 = 0.6405$

$$\Delta USnrimp_t = 0.5761 - 0.0691\Delta nrstr20_{t-1} - 0.9635\Delta exrm_{t-1} - 0.01749\Delta nrsmr20_{t-1}$$
t-stat = [-5.6585***] [-1.4576ns] [4.7721**] (11)
$$- 0.5234\Delta USnrimp_{t-1} + 0.4924\varepsilon t$$
[-7.5024***]

$R^2 = 0.9017$ Adjusted $R^2 = 0.8984$

$$\Delta JPnrimp_t = -0.01957 - 0.00463\Delta nrstr20_{t-1} - 0.2841\Delta exrm_{t-1} - 0.00216\Delta nrsmr20_{t-1}$$
t-stat = [-1.7153*] [-8.0355***] [2.0430**] (12)
$$- 1.2035\Delta JPnrimp_{t-1} + 0.6152\varepsilon t$$
[-7.6397***]

$R^2 = 0.8373$ Adjusted $R^2 = 0.8269$

3.3 Granger Causality

Table 1: Results of Granger Causality Analysis for China

Null Hypothesis	F-statistics (p-value)	Decision
nrstr20 granger cause nrimp	2.4675** (0.0475)	Supported
nrimp granger cause nrstr20	3.5149*** (0.0091)	Supported
exrm granger cause nrimp	2.2425* (0.0674)	Supported
nrimp granger cause exrm	3.6254*** (0.0076)	Supported

Table 2: Results of Granger Causality Analysis for India

Null Hypothesis	F-statistics (p-value)	Decision
exrm granger cause nrimp	2.7960** (0.0424)	Supported

Table 3: Results of Granger Causality Analysis for USA

Null Hypothesis	F-statistics (p-value)	Decision
nrimp granger cause nrstr20	2.1106* (0.0825)	Supported
nrimp granger cause nrsmr20	2.2915* (0.0625)	Supported

Table 4: Results of Granger Causality Analysis for Japan

Null Hypothesis	F-statistics (p-value)	Decision
nrimp granger cause nrstr20	2.7396* (0.0679)	Supported
exrm granger cause nrimp	2.7261* (0.0687)	Supported
nrimp granger cause nrsmr20	4.0854** (0.0188)	Supported

Source: Own Data Calculation

Note: *, ** and * denotes significance at 1%, 5% and 10% significant level respectively.**

4. Discussion and Conclusion

As seen the equations above, Equation (5), (6), (7) & (8) shows China, India, USA and Japan NR import demand ($nrimp_t$) cointegration equation. The variables of NR import ($nrimp_{t-2}$), SICOM price ($nrstr20_{t-2}$) and SMR20 price ($nrsmr20_{t-2}$), are cointegrated and has a long-term relationship with statistically significance at α 0.01, 0.05 and 0.10 level respectively. However, the exchange rate ($exrm_{t-2}$) is not significant in China and USA at α 0.05 level.

Looking at Equation (9), it shows China's natural rubber import demand VECM model results, the explanatory variables accounted for about 87.25 percent of the variation in the China natural rubber import demand equation. Estimations reveal that the explanatory variables, namely the world NR price ($nrstr20_{t-1}$), China's exchange rate ($exrm_{t-1}$), NR SMR20 price ($nrsmr20_{t-1}$) and the lag variable of China's natural rubber import demand ($nrimp_{t-1}$), were the most important explanatory variable with statistically significance at the 0.01 level.

For Equation (10), it explains India's natural rubber import demand VECM model results, the explanatory variables accounted for about 71.80 percent of the variation in the India natural rubber import demand equation. Estimations reveal that the explanatory variables, namely the world NR price ($nrstr20_{t-1}$), India's exchange rate ($exrm_{t-1}$), NR SMR20 price ($nrsmr20_{t-1}$) and the lag variable of India's natural rubber import demand ($nrimp_{t-1}$), were the most important explanatory variable with statistically significance at the 0.01, 0.05 and 0.10 level, respectively.

Looking at Equation (11), it explains on USA's natural rubber import demand VECM model results, the explanatory variables accounted for about 90.17 percent of the variation in the USA natural rubber import demand equation. Estimations reveal that the explanatory variables, namely the world NR price ($nrstr20_{t-1}$), NR SMR20 price ($nrsmr20_{t-1}$) and the lag variable of USA's natural rubber import demand ($nrimp_{t-1}$), were the most important explanatory variable with statistically significance at the 0.01 and 0.05 level, respectively.

Last but not least, looking at Equation (12), it explains on Japan's natural rubber import demand VECM model results, the explanatory variables accounted for about 83.73 percent of the variation in the Japan natural rubber import demand equation. Estimations reveal that the explanatory variables, namely the world NR price ($nrstr20_{t-1}$), Japan's exchange rate ($exrm_{t-1}$), NR SMR20 price ($nrsmr20_{t-1}$) and the lag variable of Japan's natural rubber import demand ($nrimp_{t-1}$), were the most important explanatory variable with statistically significance at the 0.01, 0.05 and 0.10 level respectively.

Moreover, Table 4.1 above shows China's granger causality analysis results. In the Engle-Granger test, F-statistics of the two variables of world NR price (nrstr20) and import demand (nrimp); exchange rate (exrm) and import demand (nrimp) are only significant at α 0.01, 0.05 & 0.10 level. Therefore, there are world NR price (nrstr20) "Granger causes" to import demand (nrimp) and exchange rate (exrm) "Granger causes" to import demand (nrimp). Moreover, their granger causality relationships are bidirectional. Then, they are cointegrated and also a long-run equilibrium relationships between each two variables. Table 4.2 above shows India's granger causality analysis results. In the Engle-Granger test, F-statistics of the two variables of exchange rate (exrm) and import demand (nrimp) is only significant at α 0.05 level. Therefore, there is a variable exchange rate (exrm) "Granger causes" a variable import demand (nrimp) and the direction of the granger causality relationship is unidirection. Then, there is cointegrated and also a long-run equilibrium relationships between two variables.

Table 4.3 above shows USA's granger causality analysis results. In the Engle-Granger test, import demand (nrimp) to world NR price (nrstr20); and import demand (nrimp) to SMR20 price (nrsmr20) are only significant at α 0.10 level. Therefore, their directions of granger causality relationship are unidirection. Then,

they are cointegrated and also a long-run equilibrium relationship between the two variables. Table 4.4 shows Japan's granger causality analysis results. In the Engle-Granger test, F-statistics of the two variables of import demand (nrimp) to world NR price (nrstr20) is significant at α 0.10 level. Therefore, there is a variable import demand (nrimp) "Granger causes" a variable world NR price (nrstr20) and the direction of granger causality relationship is unidirection. Then, there is a cointegrated and long-run equilibrium relationship between the two variables of nrimp and nrstr20. In the Engle-Granger test, F-statistics of the two variables of exchange rate (exrm) to import demand (nrimp) is significant at α 0.10 level. Therefore, there is a variable exrm "Granger causes" a variable nrimp and the direction of the granger causality relationship is unidirection. Then, there is cointegrated and also a long-run equilibrium relationship between the two variables exrm and nrimp. In the Engle-Granger test, F-statistics of the two variables of import demand (nrimp) to SMR20 price (nrsmr20) is significant at α 0.10 level. Therefore, there is a variable import demand (nrimp) "Granger causes" a variable SMR20 price (nrsmr20) and the direction of granger causality relationship is unidirection. Then, there is a cointegrated and long-run equilibrium relationship between the two variables of nrimp and nrsmr20.

From the result of this study, the independent variables that include world price, exchange rate and export price has significant impact on import model of natural rubber. Since the supplier and consumer unable to control the price of NR, because the price of NR was affect by the demand and supply of the market (Teh, 2015; Tulasombat *et al.,* 2015; & Khin *et al.*, 2017). It is also important to local farmers to develop a market strategy to improve the production of natural rubber when the prediction of import natural rubber by others countries will increase in near future (Sakan, 2012). Recommendations for future study, the findings of this study able to provide information to fill the gap by determine the import volume on current and potential market such as China, India, USA and Japan and also market participants in their consumptions and financing decisions due to NR is an important commodity for world market.

References

1. Hnin, E. (2017). *Economic Importance of Rubber in Thailand.* Retrieved from http://ap.fftc.agnet.org/ap_db.php?id=819
2. IRSG (2017). Latest Rubber Statistical Bulletin. International Rubber Study Group, 111 North Bridge Road, #23-01/06 Peninsula Plaza, Singapore 179098.
3. Kannan, M. (2013). The Determinants of Production and Export of Natural Rubber in India. *IOSR Journal of Economics and Finance*, 1(5), 41-45.
4. Khin, Aye Aye, Chau, Wong Hong, Yean, Ung Leng, Keong, Ooi Chee, Bin, Raymond Ling Leh. (2017). Examining between Exchange Rate Volatility and Natural Rubber Prices: Engle-Granger Causality Test, International Journal of Economics and Financial Issues (SCOPUS), 7 (6), 33-40.
5. MREPC. (2017). The Malaysian Rubber Export Promotion Council, http://www.mrepc.com/
6. Ravikumar, B., Shao, L., & Sposi, M. (2017, November 29). Why Was the Decline in U.S. Trade Larger This Time? A Global View. Retrieved April 08, 2018, from https://www.stlouisfed.org/ Publications/Regional-Economist/October-2013/Why-Was-the Decline-in-US-Trade-Larger-This-Time-A-Global-View.
9. Sakan, I. (2012). Forecasting the price of natural rubber in Malaysia. *Thesis, California State University, Sacramento.*
10. Studenmund, A.H, (2017), Using Econometrics: A Practical Guide, 6th Edition, Pearson, Prentice Hall. ISBN-10: 1292021276.
11. Sabu, T. (2017). Natural Rubber Based Research and Developments in Indian Rubber Industry. *International Rubber Conference*, 90(3), 72-76.
12. Teh, S. (2015). *The impact of commodites price on volatility on Ringgit Malaysia.* Retrieved from http://eprints.utar.edu.my/id/eprint/2424
13. Tulasombat, S., Bunchapattanasakda, C., & Ratanakomut, S. (2015). The Effect of Exchange Rates on Agricultural Goods for Export: A Case of Thailand . *Information Management and Business Review* , 7(1), 1-11.

Analytical likelihood derivatives for state space forecasting models

Jonathan Hosking, Ramesh Natarajan
Amazon.com, New York, U.S.A.

Abstract

State space models are a flexible and widely used family of statistical models for time series analysis and forecasting. Fitting of the models to historical data is greatly facilitated by the availability of analytical derivatives of the log-likelihood function. We have obtained a new expression for these derivatives in terms of quantities routinely computed in Kalman filtering and smoothing. This result makes it straightforward to construct an optimization method based on gradient descent using analytical log-likelihood derivatives. We present the derivation and give some examples of the gain in speed of parameter estimation when analytical derivatives are used.

Keywords

Time series; maximum likelihood; computation

1. State space models

A state space model for time series data treats the observed data vector yt as a noisy observation of an unobserved state vector αt that evolves according to a Markov process. We consider the linear gaussian state space model, which we write using the notation of Durbin and Koopman (2012, eq. (4.12)):

$$y_t = Z_t\alpha_t + \varepsilon_t \qquad \text{(observation equation)}, \qquad (1)$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t \qquad \text{(state transition equation)}. \qquad (2)$$

Vectors y_t, α_t, and η_t have respective dimensions p, m, and r, with r≤m. The observation noise ε_t, state disturbance η_t, and initial state α_1 have normal distributions:

$$\varepsilon_t \sim N(0,H_t), \qquad \eta_t \sim N(0,Q_t), \qquad \alpha_1 \sim N(a_1,P_1). \qquad (3)$$

The state space model provides a flexible framework for specification of forecasting models. Special cases include ARIMA models (Hamilton, 1994), exponential smoothing (Hyndman et al., 2002), and models involving trend, seasonality, regression, and noise components, variously known as dynamic linear models (West and Harrison, 1997; Petris et al., 2009) structural models (Harvey, 1989). Dynamic linear models, because they decompose a time series into interpretable components, make particularly effective and understandable

forecasting models. They are suitable for modeling and forecasting many kinds of time series data.

Inference for state space models commonly uses Kalman filtering and smoothing iterations, which provide conditional distributions of the state vector αt given observations up to time t or all observations in t =1,...,n, and can also be used to forecast future observations. When used in practice, state space models typically include model parameters that must be estimated from observed data. Maximum-likelihood estimation is commonly used, with the likelihood being maximized by optimization procedures that use numerical derivatives of the likelihood function. The optimization step can be greatly accelerated by the use of analytical derivatives.

In this paper we present a new analytical expression for the log-likelihood derivative, eq. (14) below. This expression can be computed using quantities that are routinely computed during the Kalman filtering and smoothing iterations, and involve no additional iterations. To the best of our knowledge all previously published expressions either require further iterations or do not cover all possibilities of the state space model Our result makes it straightforward to construct an optimization method based on gradient descent: in Sections 5 and 6 we outline the construction of such a method and illustrate its performance.

2. Filtering and smoothing

Inferential procedures for state space models commonly include filtering and smoothing. Let Y_t denote the information available at time t, i.e., the data $\{y_1,...,y_t\}$. The Kalman filter computes estimates of the states at time t based on data up to time $t-1$. The estimates are $a_t \equiv \mathrm{E}(\alpha_t | Y_{t-1})$, $P_t \equiv var(\alpha_t | Y_{t-1})$, and starting with a_1 and P_1 are computed for $t = 1,2,...$ from the filtering equations Durbin and Koopman (2012, eq. (4.24))

$$v_t = y_t - Z_t a_t, \qquad F_t = Z_t P_t Z_t^{\mathrm{T}} + H_t, \tag{4}$$

$$a_{t+1} = T_t a_t + K_t v_t, \qquad P_{t+1} = T_t P_t L_t^{\mathrm{T}} + R_t Q_t R_t^{\mathrm{T}}, \tag{5}$$

where

$$K_t = T_t P_t Z_t^{\mathrm{T}} F_t^{-1}, \qquad L_t = T_t - K_t Z_t = T_t - T_t P_t Z_t^{\mathrm{T}} F_t^{-1} Z_t. \tag{6}$$

State smoothing provides estimates of the mean and variance of the state vector at each time point, given a data set of length n. The estimates are $\hat{\alpha}_t \equiv E(\alpha_t | Y_n)$ and $V_t \equiv var(\alpha_t | Y_n)$. The iterations Durbin and Koopman (2012, eq. (4.44)) start with $r_n = 0$ and $N_n = 0$ and proceed backwards in time: for t =n,n−1,...,1, compute

$$r_{t-1} = Z_t^{\mathrm{T}} F_t^{-1} v_t + L_t^{\mathrm{T}} r_t, \qquad N_{t-1} = Z_t^{\mathrm{T}} F_t^{-1} Z_t + L_t^{\mathrm{T}} N_t L_t, \tag{7}$$

$$\hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad V_t = P_t - P_t N_{t-1} P_t. \tag{8}$$

3. Model parameters and estimation

It is common for a state space model specification to involve unknown parameters, particularly in the variance matrices H_t and Q_t but also in the transition matrices Z_t, T_t and R_t. Estimation of these parameters is then required for the model to be used in forecasting.

A commonly used estimation method is maximum likelihood. The log-likelihood of an observed data set Y_n is Durbin and Koopman (2012, eq. (7.2))

$$\log L(Y_n) = -\tfrac{1}{2}np\log(2\pi) - \tfrac{1}{2}\sum_{t=1}^{n}\log|F_t| - \tfrac{1}{2}\sum_{t=1}^{n}v_t^{\mathrm{T}}F_t^{-1}v_t\,. \tag{9}$$

Estimation then requires the maximization of (9) with respect to the model parameters. Numerical methods are generally necessary. Many numerical optimization methods, including the popular BFGS method (Wiipedia, 2015) and its variants, use information about the derivatives of the objective function. Numerical derivatives, obtained by finite differences, can be used, but if analytical expressions are available for the derivatives, the optimization can often be greatly accelerated.

Log-likelihood derivatives for state space models can be obtained in several ways. For some classes of model, such as ARMA models in state space form, derivatives can be computed using recursions adapted for the particular model class (Ansley and Kohn, 1985; Melard, 1985). Consideration of a complete-data likelihood, as though the states αt were observable, yields explicit derivatives (Segal and Weinstein 1988, 1989; Koopman and Shephard, 1992). but in a simple form only for a restricted set of models for which the matrices H_t, Q_t, and R_t are invertible. Direct differentiation of (9) yields a recursive procedure whose effect is to increase the number of quantities that must be carried through the Kalman filter recursions. This approach is described in more detail in Section 4.

What has hitherto been lacking is a single explicit expression for the log-likelihood derivative that applies to all linear gaussian state space models in the form (1)–(3), without any restrictions on the system matrices. We have derived such an expression, by extending previous authors' treatments of the direct log-likelihood derivative. Our result is included in Section 4 below, as eq. (14).

4. Log-likelihood derivatives for the state space model.

Let θ be a scalar model parameter. We use the notation $\dot{X}$ to denote $\partial X/\partial\theta$. For an invertible matrix X, we note that and $\partial X^{-1}/\partial\theta = -X^{-1}\dot{X}X^{-1}$. Differentiating (9) with respect to θ gives the derivative of the log-likelihood:

$$\frac{\partial}{\partial\theta}\log L(Y_n) = -\tfrac{1}{2}\sum_{t=1}^{n}\mathrm{tr}(F_t^{-1}\dot{F}_t) + \tfrac{1}{2}\sum_{t=1}^{n} v_t^{\mathrm{T}}F_t^{-1}\dot{F}_tF_t^{-1}v_t - \sum_{t=1}^{n} v_t^{\mathrm{T}}F_t^{-1}\dot{v}_t. \tag{10}$$

Differentiating v_t and F_t in (4) gives

$$\dot{v}_t = -\dot{Z}_t a_t - Z_t\dot{a}_t, \qquad \dot{F}_t = \dot{Z}_tP_tZ_t^{\mathrm{T}} + Z_t\dot{P}_tZ_t^{\mathrm{T}} + Z_tP_t\dot{Z}_t^{\mathrm{T}} + \dot{H}_t. \tag{11}$$

Substituting (6) into (5), we can show after some simplification that

$$\begin{aligned}\dot{a}_{t+1} &= L_t\dot{a}_t + L_t\dot{P}_tZ_t^{\mathrm{T}}F_t^{-1}v_t - K_t\dot{H}_tF_t^{-1}v_t + (\dot{T}_t - K_t\dot{Z}_t)(a_t + P_tZ_t^{\mathrm{T}}F_t^{-1}v_t)\\ &\quad + L_tP_t\dot{Z}_t^{\mathrm{T}}F_t^{-1}v_t,\end{aligned} \tag{12}$$

$$\begin{aligned}\dot{P}_{t+1} &= L_t\dot{P}_tL_t^{\mathrm{T}} + K_t\dot{H}_tK_t^{\mathrm{T}} + (\dot{T}_t - K_t\dot{Z}_t)P_tL_t^{\mathrm{T}} + L_tP_t(\dot{T}_t - K_t\dot{Z}_t)^{\mathrm{T}}\\ &\quad + R_t\dot{Q}_tR_t^{\mathrm{T}} + \dot{R}_tQ_tR_t^{\mathrm{T}} + R_tQ_t\dot{R}_t^{\mathrm{T}}.\end{aligned} \tag{13}$$

Substituting (11) into (10), and evaluating the at and Pt occurring in (11) by iterative application of (12)–(13), yields a set of recursions from which the derivative of the log-likelihood can be computed iteratively. Detailed expressions have been given for example by (Zadrozny, 1989; Anderson et al., 1995; Nagakura, 2013). However, by investigating the derivatives in more detail we have obtained explicit expressions for the likelihood derivative in terms of quantities routinely computed in Kalman filtering and smoothing. Details are omitted here, but are available from the authors. The final result is

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L(Y_n) &= \tfrac{1}{2}\sum_{t=1}^{n}\mathrm{tr}\{\dot{H}_t(u_tu_t^{\mathrm{T}} - D_t)\} + \sum_{t=1}^{n}\mathrm{tr}\{\dot{Z}_t(\hat{\alpha}_tu_t^{\mathrm{T}} - P_tZ_t^{\mathrm{T}}F_t^{-1} + P_tL_t^{\mathrm{T}}N_tK_t)\}\\ &\quad + \tfrac{1}{2}\sum_{t=1}^{n-1}\mathrm{tr}\{\dot{Q}_tR_t^{\mathrm{T}}(r_tr_t^{\mathrm{T}} - N_t)R_t\} + \sum_{t=1}^{n-1}\mathrm{tr}\{\dot{R}_tQ_tR_t^{\mathrm{T}}(r_tr_t^{\mathrm{T}} - N_t)\}\\ &\quad + \sum_{t=1}^{n-1}\mathrm{tr}\{\dot{T}_t(\hat{\alpha}_tr_t^{\mathrm{T}} - P_tL_t^{\mathrm{T}}N_t)\} + \tfrac{1}{2}\mathrm{tr}\{\dot{P}_1(r_0r_0^{\mathrm{T}} - N_0)\} + r_0^{\mathrm{T}}\dot{a}_1\end{aligned} \tag{14}$$

where

$$u_t = F_t^{-1}v_t - K_t^{\mathrm{T}}r_t, \qquad D_t = F_t^{-1} + K_t^{\mathrm{T}}N_tK_t. \tag{15}$$

5. Practical implementation

Practical use of likelihood derivatives in model fitting requires an efficient procedure for computing the likelihood and its derivatives. with respect to the model's parametrs. For each system matrix M, a practical way of computing the contribution to the likelihood derivative is the elementwise product. of the matrices $\partial\mathrm{logL}/\partial\mathrm{M}$ and M. For example, for a parameter occurring only in

matrices T_t and P1 the log-likelihood derivative with respect to θ can be computed by the chain rule as

$$\frac{\partial \log \mathcal{L}}{\partial \theta} = \sum_t \sum_i \sum_j \left(\frac{\partial \log \mathcal{L}}{\partial T_t} \right)_{ij} (\dot{T}_t)_{ij} + \sum_i \sum_j \left(\frac{\partial \log \mathcal{L}}{\partial P_1} \right)_{ij} (\dot{P}_1)_{ij} . \tag{16}$$

From (14) it is straightforward to obtain derivatives with respect to any of the system matrices. For. a matrix X for which $\partial logL/\partial\theta = tr(XA)$, we have $\partial logL/\partial X = A^T$; thus, for example,

$$\frac{\partial \log \mathcal{L}}{\partial T_t} = r_t \hat{\alpha}_t^{\mathrm{T}} - N_t L_t P_t , \qquad \frac{\partial \log \mathcal{L}}{\partial P_1} = r_0 r_0^{\mathrm{T}} - N_t . \tag{17}$$

Using expressions (16) and (17), and their analogs for other system matrices, we can construct a practical procedure for parameter estimation. Computation of the log-likelihood derivative itself involves the following steps.

1. Run the Kalman filter, at each time point t =1,...,n accumulating the current term of the likelihood from v_t and F_t, and saving the quantities at, F_t^{-1}, K_t, P_t , and L_t for use in step 3.
2. Run the smoother. At each time point t =n,n−1,...,1 compute r_t^{-1} and N_t^{-1}; also compute u_t, $\hat{\alpha}_t$, and D_t if needed for likelihood derivatives; accumulate the likelihood derivatives with respect to the system matrices.
3. Compute likelihood derivatives with respect to each parameter, using (16) and its equivalentsfor other system matrices.

These operations can be implemented in R code, extending similar functions already available in packages such as dlm or KFAS; implementation in other languages such as Matlab or Python is reasonably straightforward.

6. Example

Hosking et al. (2013) used a state space model to forecast residential electricity usage. The model has independent components including a smoothly varying daily load curve based on B spline basis functions (with B=12), stochastically varying level that also captures seasonal variation, stochastically varying day-of-week effects, and static regression on temperature and price variables. The observations are $y_t,\ t = 1, \ldots, n$, where y_t is a vector containing demand for the m disjoint intervals of a single day (e.g., m=24 for hourly data). We represent y_t by the observation equation

$$y_t = Z\alpha_t + \varepsilon_t \tag{18}$$

where Z is an $m \times B$ matrix of basis function values, its (i,b) element being the value of the b^{th} basis function during the i^{th} interval. The B-vector αt contains the coefficients of each basis function for day t, and itself evolves according to a state space model that contains stochastically varying level and day-of-week effects, and static regression on temperature and price variables. There are two

temperature variables, representing heating and cooling degree days, and two price variables, representing elasticity of demand within the period spanned by the basis function in which the price changes, and cross-elasticity between basis-function periods. There is an additive noise component ε_t in (18), and noise components in the evolution equation for α_t, the stochastic level component, and the stochastic day-of-week effects. A full description of the model and its state-space form are given in Hosking et al. (2013). The model has observation vector length m, state vector length $9B+4$, and 4 distinct model parameters, one for each variance component.

The model for hourly demand (m=24) has observation vector length 24 and state vector length 112. Model parameters were estimated using the L-BFGS-B method as implemented in R function optim. Parameter estimation on 11 months of data took 274 sec using numerical derivatives. With the new procedure using analytical derivatives, estimation time is reduced to 92 sec, faster by a factor of 3.0.

The model can be extended by allowing each diagonal element in the observation and state noise variance matrices to vary independently. This extended model has $60\ (=m+3B)$ model parameters in the variance matrices. Starting with parameter values corresponding to the fitted 4-parameter model, estimation with numerical derivatives took 3246 sec; with analytical derivatives, 81 sec. Using analytical derivatives here delivers the same solution as numerical derivatives, but 40 times faster.

The same model can be used for demand measured over 15-minute intervals, now with m=96. The state vector is the same as for the model for hourly data: it has 112 elements, and their physical interpretations are the same as for the states in the hourly model. Like the hourly model, in its basic form it has 4 distinct model parameters, one for each variance component, in the variance matrices. Using analytical derivatives speeds up the estimation of these parameters by a factor of approximately 2.9. When the 15-minute model is extended by allowing each diagonal element in the observation and state noise variance matrices to vary independently, it has $132\ (=m+3B)$ model parameters in the variance matrices. Using analytical derivatives speeds up the estimation of these parameters by a factor of 84. The computational results are summmarized in Table 1.

7. Conclusions

We have derived an explicit expression for the derivative of the log-likelihood function for a state space time series model. It enables numerical procedures for parameter estimation to compute

Table 1. Computation times for the models discussed in Section 6. "Analytical" computations use analytical derivatives; "Numerical" computations use numerical derivatives; "Ratio" is the ratio of the computation times for numerical and analytical derivatives.

Data	Model type	Number of Parameters	Computation time		Ratio
			Analytical	Numerical	
Hourly	Basic	4	92	274	3.0
Hourly	Extended	60	81	3246	40.0
15-minute	Basic	4	281	809	2.9
15-minute	Extended	132	356	30069	84.5

derivatives using ana anlaytical expression rather than a finite-difference approximation. This speeds up the computations by a factor that in practical applications can be between 3 and 80.

References

1. Anderson, E. W., McGrattan, E. R., Hansen, L. P., and Sargent, T. J. (1995). On the mechanics of forming and estimating dynamic linear economies. In Handbook of Computational Economics, vol. 1, eds. H. M. Amman, D. A. Kendrick, and J. Rust, pp. 171–252.
2. Ansley, C. F. and Kohn, R. (1985). A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. J. Statist. Comput. Simul., 21, 135–169.
3. Durbin, J., and Koopman, S. J. (2012). Time Series Analysis by State Space Methods. Oxford University
4. Press.
5. Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.
6. Harvey, A. C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.
7. Hosking, J. R. M., Natarajan, R., Ghosh, S., Subramanian, S., and Zhang, X. (2013). Short-term forecasting of the daily load curve for residential electricity usage in the Smart Grid. Appl. Stoch. Models Bus. Ind., 29, 604–620.
8. Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecasting, 18, 439-454.
9. Koopman, S. J., and Shephard, N. (1992). Exact score for time series models in state space form. Biometrika, 79, 823–826.
10. Melard, G. (1985). Exact derivatives of the likelihood of ARMA processes. In´ Proceedings of the Statistical Computing Section, pp. 187–192. American Statistical Association, Washington, DC.

11. Nagakura, D. (2013). Exact gradient vector of loglikelihood for linear Gaussian state space models. Available at http://dx.doi.org/10.2139/ssrn.1634552.
12. Petris, G., Petrone, S., and Campagnoli, P. (2009). Dynamic Linear Models with R. Springer, New York.
13. Segal, M., and Weinstein, E. (1988). A new method for evaluating the log-likelihood gradient (score) of linear dynamic systems. IEEE Trans. Auto. Control 33, 763–766.
14. Segal, M., and Weinstein, E. (1989). A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix for linear dynamic systems. IEEE Trans. Info. Theory 35, 682–687.
15. West, M., and Harrison, P. J. (1997). Bayesian Forecasting and Dynamic Models, 2nd ed. Springer, New York.
16. Wikipedia contributors (2018). Broyden-Fletcher-Goldfarb-Shanno algorithm. https://en.wikipedia.org/wiki/Broyden%E2%80%93Fletcher%E2%80%93Goldfarb%E2%80%93Shanno algorithm.
17. Zadrozny, P. A. (1989). Analytic derivatives for estimation of linear dynamic models. Comput. Math. Appl., 18, 539–553.

Reaction times: A new approach using new advances in distribution theory

Carla Susete G. Francisco[1], José António S. Macias[2], Filipa Neiva C. Ribeiro[1]

[1]Catholic University of Portugal, Lisbon, Portugal
[2]University of Coruna, Coruna, Spain

Abstract:

Reaction times (RTs) have been an important measure in the investigation of the cognitive process. Both psychologists and neurologists have been engaged in the development of studies on the evaluation of models to describe the RTs processes. A very important issue is the study of probability distributions that describe the RTs, which more general probability distribution families can hardly explain. Among the several existing models, the most usual for this type of study is the LATER (Linear Approach to Threshold with Ergodic Rate) of Carpenter (1981). Nakahara et al. (2006) developed an extension of the LATER model, which they call the extended LATER model (ELATER), to account for trial-by-trial variability of both pre- and post-processes together. Moscoso (2008) develops a new theory in which TRs are directly proportional to the difficulty of the task and inversely proportional to the rate at which information becomes available to solve. The possibility of fluctuations might be taken into account, in both the information gain rate (r) and in the resting level to the distance (Δ) of threshold and then TRs will follow a distribution corresponding to the ratio between two normally distributed variables. The LATER model considers that the values of (r) and (Δ) are statistically independent of each other, although new pieces of evidence considered aren´t correct, once several experiments produced the necessity for introducing an additional parameter into the model representing the correlation between (r) and (Δ). This project aims to develop a new model from the Fieller distribution for the ratio of two normally distributed variables. Finally, another issue that could cast doubts on the plausibility of LATER as a model of activity accumulation of the neurons is the constrained linear trajectory of the accumulation of evidence of validity. A reformulation of the LATER in terms of the diffusion process, named LATER-d was analyzed.

Keywords

Drift-diffusion model; Ergodic Rate; LATER; Reaction time; Rise-to-threshold model.

1. Introduction

Reaction time (TR), or latency, is the interval between presenting a stimulus and making a response to it. Saccade is the eye movement we make to look at a target in our field of view. We make two or three saccades every second of our lives. Latency is in the order of 200 ms, see Robinson (1964). With modern equipment based on computational systems, it is possible to obtain very large datasets of saccadic latency measurements and to determine the form of their variability. The result is always a skewed distribution, with a longer tail to the right. This distribution does not fit particularly well in any of the most common standard mathematical distributions (Gaussian, Poisson, Gamma, etc.). Observed variability in reaction time might will be due to a variability in the rate of the underlying process. Looking at the reciprocal of reaction time (1 / *T*) promptness. The distribution of the reciprocal of reaction is not only symmetrical, but actually looks as though it might be Gaussian. If it were, that would not only make for easier mathematical analysis, but would also suggest that we had reached a genuinely fundamental phenomenon. A graphical procedure is designed to convert our histogram into a cumulative histogram. For it, we are using a specially-distorted scale, this time on the vertical, probability axis (a reciprobit plot). In this case, if the distribution is indeed Gaussian, we should get a straight line. This approach provides means of characterizing the behavior of the reaction time using experimental data that is summarized through a very small number of parameters, since it is sufficient to specify a median and the intercept of the main distribution, Burle (2004).

2. Methodology

The idea of analyzing reciprocal latency, results from the treatment of reaction time, due to a process whose rate varies randomly from one experiment to another. Some type of decision signal, starting at an initial level S_0, rises to a constant rate r until it reaches a threshold value S_T, at which point a response is initiated. If r is randomly varied from one test to another, such as a Gaussian with mean μ and variance σ^2, the asymmetry of observed latency distributions is immediately explained. There are four approaches:

- LATER (Linear Threshold Approximation with Ergodic Rate) Model.
- ELATER (Extended LATER) Model.
- Fieller distribution with parameters $\kappa, \lambda_1, \lambda_2, \rho$.
- DDM Model.

The DDM (Drift- Diffusion Model) is the most successful of the model family to simulate growths up, to the threshold level. This model shares many of its features with the LATER model. DDM family models are classified as description of implementation at the same process level,

which the RT theory explains at a higher level. The advantages and disadvantages of the DDM model are:

- The DDM model provides a direct mechanism to account for and predict the probabilities of delays and their latencies.
- Suggestive approach to the behavior of neuronal populations. Presence of auditory fluctuations. Complexity of the DDM model. Doubts about the linearity of the accumulation process: experimental observations suggest following some kind of exponential law.

LATER (*Linear Threshold Approximation with Ergodic Rate*) Model. LATER, is a model originally derived empirically and vulnerable to experimental testing. Over the last decade we have been attempting to verify this functional interpretation by trying to challenge its three elements:

- S_0 represents log prior probability. The change in reaction time is linearly related to the log probability.
- S_T represents a threshold. The main part swivels about a fixed intercept. Reduction in latency is associated with a large increase in the number of early responses.
- μ_r represents the supply of information. The rate of information supply affects the mean rate of rise of the decision signal, and this turn causes the distributions not to swivel, but to be shifted horizontally in a parallel fashion. We consider that r is a Gaussian variable with mean μ and variance δ^2.
- Time between the start and the threshold is:

$$T=(S_T - S_0)/r$$

- Reciprocal of latency, 1/T:

$$1/T=r/(S_T - S_0)$$

- Following that r is a Gaussian random variable, distribution of 1/T is Gaussian with mean $\mu/(S_T - S_0)$ and variance $\sigma^2/(S_T - S_0)^2$.

All biological systems are subject to unpredictable perturbations, technically known as noise. the sensorial noise does not contribute significantly to the variability of reaction time, according with a large number of evidences that strongly suggest that, in most conditions, -randomness (Signal-to-noise ratio in sensory systems). Neurophysiological experiments show that the contribution of randomness to the overall variability of reaction time is insignificant. Suggestions such:

- LATER decision mechanism can be thought of as being preceded by a detection stage, obeying random-walk dynamics.
- The observed randomness of reaction time does not originate in the outside world, it is deliberately injected into the system from within. Agents try to be as unpredictable (random) as they possibly can.

Free will can be defined as if our actions do not seem unpredictable to ourselves because we see them coming to fruition, see Figure 1. However, being aware, between the decision and the resultant action, of what one is about to do is not at all the same thing as actually willing the movement in the first place.

Figure 1: Free will Randomness.

ELATER model is an extension of LATER in order to introduce some variability in distance ΔS and the slope r between experiments. Now, we consider that both variables are independently and normality distributed $\Delta S \sim N(\mu_\Delta, {\sigma_\Delta}^2)$ and $\Delta r \sim N(\mu_r, {\sigma_r}^2)$. Latency distribution is determined by:

$$\frac{1+\mu_1\alpha^2 t}{1+\alpha^2 t^2}\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1+\alpha^2 t^2}}\,exp\left(-\frac{(t-\mu_1)^2}{2\sigma_1^2(1+\alpha^2 t^2)}\right), \quad \text{where, } \mu_1=\frac{\mu_r}{\mu_\Delta},\ \sigma_1=\frac{\sigma_r}{\mu_\Delta} \text{ e } \alpha=\frac{\sigma_\Delta}{\sigma_r}.$$

Recinormal Distribution (Figures 2 and 3)

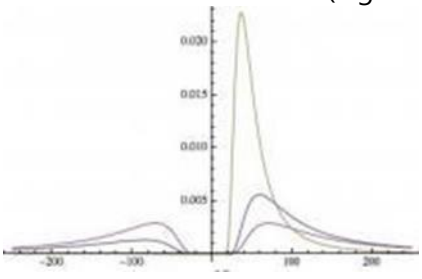


Figure 2: pdf: blue (μ = 0.005, σ = 0.01) red (μ = 0, σ = 0. 01) yellow (μ = 0.02, σ = 0.01)

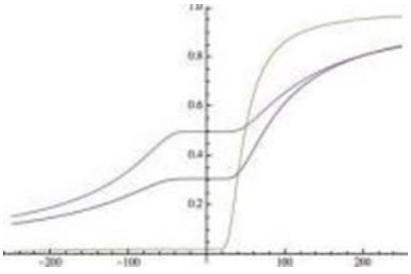


Figure 3: cdf: blue (μ = 0.005, σ = 0.01) red (μ = 0, σ = 0.01) yellow (μ = 0.02, σ = 0.01)

Reciprobit Plot

It's the normal quantile-quantile plot with the axes swapped and a (changed sign) reciprocal transformation on the data.

ELATER Model (Figures 4 to 17)

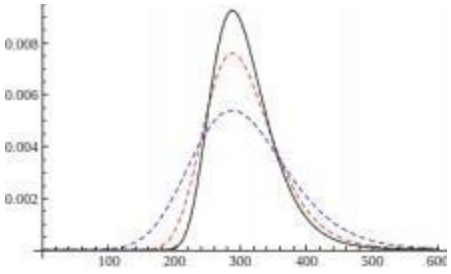


Figure 4: pdf: blue (μ $_r$ = 1 / 30, σ $_r$ = 0. 005, μ_Δ = 10, σ_Δ = 2 - red (μ $_r$ = 1 ⁄ 30, σ_r = 0. 005, μ_Δ = 10, σ_Δ = 1) - green (μ r = 1 ⁄ 30, σ r = 0. 005, μ_Δ 10, σ_Δ = 0. 1) - black (LATER model: μ_r = 1 ⁄ 300, σ_r = 0. 005)

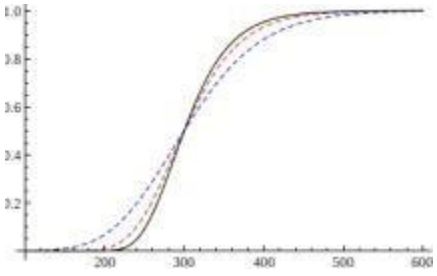


Figure 5: cdf: blue (μ $_r$ = 1 ⁄ 30, σ_r = 0. 005, μ_Δ = 10, σ_Δ = 2) - red (μ $_r$ = 1 ⁄ 30, σ_r = 0. 005, μ_Δ = 10, σ_Δ = 1) - green (μ r = 1 ⁄ 30, σ r = 0. 005, μ Δ 10, σ Δ = 0. 1) - black (LATER model: μ_r = 1 ⁄ 300, σ_r = 0. 005)

Value of λ_1	Value of λ_2	Distribution	Normal QQ-plot
0	0	Dirac(κ)	
any	0 (< .22)	$N(\kappa, (\kappa\lambda_1)^2)$	straight line
0 (< .22)	any	$\text{Recip}N\left(\frac{1}{\kappa}, \left(\frac{\lambda_2}{\kappa}\right)^2\right)$	straight line (on reciprocal plot)
∞ (> .443)	∞ (> .443)	$\text{Cauchy}\left(\rho\kappa\frac{\lambda_2}{\lambda_1}, \frac{\lambda_2}{\lambda_1}\kappa\sqrt{1-\rho^2}\right)$	horizontal line and two vertical lines at edges

Figure 6: Fieller Distribution and LATER model: Special cases of Fieller 's Distribution

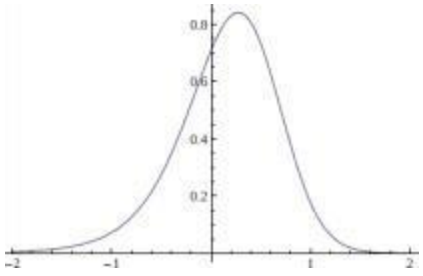


Figure 7: pdf: Fieller Distribution with μ 1 = 2, μ 2 = 10, σ 1 = 5, σ 2 = 2, ρ = 0. 5, κ = 0. 2, λ 1 = 2. 5, λ 2 = 0. 2

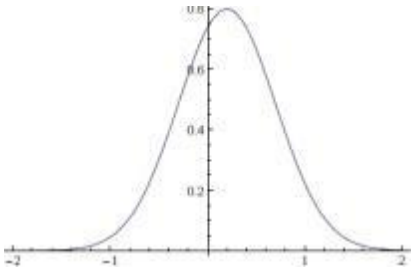


Figure 8: pdf: Normal Distribution with μ = 0. 2 and $\sigma = \lambda\ 1 * \lambda\ 2.\ 5$

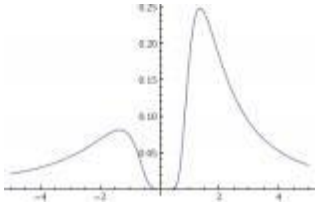


Figure 9: pdf: Fieller Distribution with *μ 1 = 10, μ 2 = 2, σ 1 = 2, σ 2 = 5, ρ = 0. 5, κ = 5, λ 1 = 0. 2, λ 2 = 2. 5*

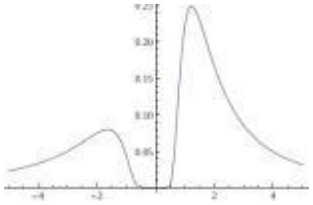


Figure 10: pdf: Normal Distribution with μ = 1/5 and σ = 0.5

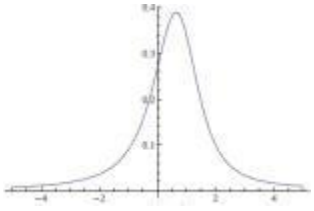


Figure 11: pdf: Fieller Distribution with μ 1 = 2, μ 2 = 4, σ 1 = 5, σ 2 = 3, ρ = 0. 5, κ = 0. 5, λ 1 = 2. 5, λ 2 = 0. 75

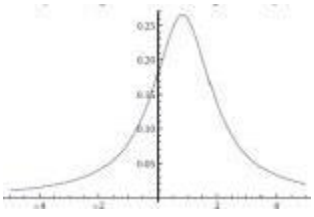


Figure 12: pdf: Cauchy Distribu tion with parameters: 5 / 6, 5 / 3 1 − 0. 5

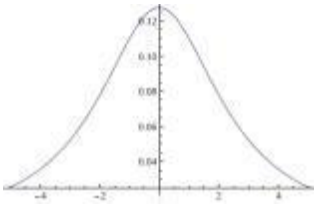


Figure 13: pdf: Fieller Distribution with μ 1 = 2, μ 2 = 4, σ 1 = 5, σ 2 = 3, ρ = 0. 5, κ = 0. 5, λ 1 = 2. 5, λ 2 = 0. 75

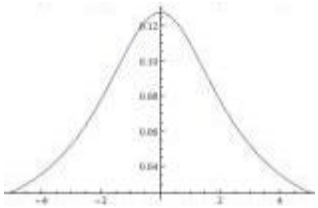


Figure 14: pdf: Cauchy Distribution with parameters: 5 / 6, 5 / 3 1 − 0. 5

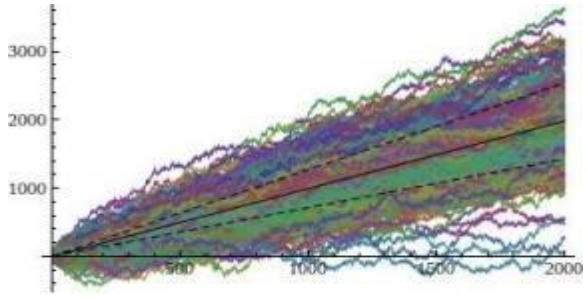


Figure 15: DDM diffusion trajectories Consider a Brownian motion with a drift and infinitesimal variance: We have generated 250 trajectories for an It ˆ o process: $dx\,(t) = dt + 12.02\,dW\,(t)$

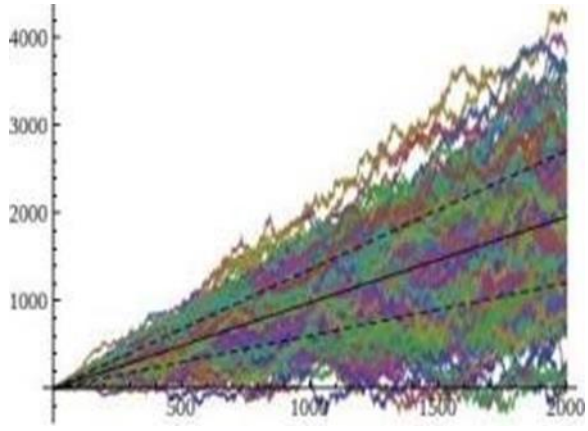


Figure 16: Later-d Trajectories Consider a model with $r = 1$ and $\sigma r = 0.38$. We have generated 250 trajectories for an Brownian motion process: $dx\,(t) = dt + 0.38\,2\,tdW\,(t)$

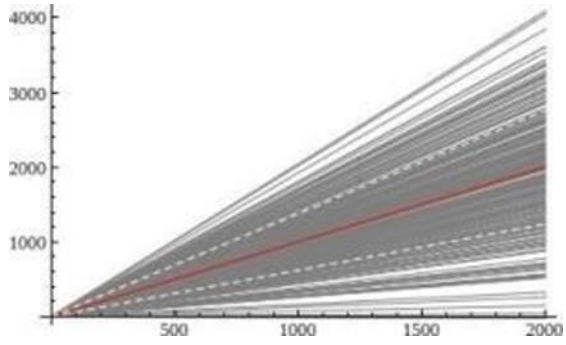


Figure 17: This research w LATER model Consider a LATER model (with $r = 1$ and $\sigma r = 0.38$) with r normally distributed: We have generated 250 trajectories with t between 0 and 2000.

3. Results

Data analysis was performed by distribution analysis methods and by using a data analysis in different categories whenever the data sample is large enough. The analysis techniques used where: multi- variance data

analysis, factor analysis and data adjustment. In order to perform the data analysis, it will be necessary to use the software R and RStudio (R CoreTeam, 2017) and their respective packages for data visualization and factorial analysis. The use of spreadsheet software was also required for some data processing. The research was carried out based on existing theories, through the analysis of a model that collects all the advantages of different approaches. This theoretical model was used as a basis for obtaining simulated data, which was later used to compare the recollected data with the actual experimental design models. In order to obtain real data, two of the following sources where used: Preexisting sources or real data. Pre-existing sources provide us with data of experiments already carried out, such as the one located in the following website: https://www.humanbenchmark.com/. Real data was obtained by conducting some direct experiments, from a website, in order to model and compare both response times, from the simulated data and the actual data. All these experiments have to be done by using small samples, since the results obtained present some problem regarding the size of the datasets used, however possible in later developments, the datasets can be extended to obtain more accurate results.

4. Discussion and Conclusion

The results of this study is try to explain the asymmetry of observed latency distributions. There are several possibilities to make a significant contribution to these studies:

- Adjust the best distribution analysis for the real data:
  - Fieller distribution implies linearity.
  - Model LATER-d implies non-linearity.
- Recinormality assessment: The evaluation of the recinormality of the data sets by item will be provided by the analysis of the Reciprobit graphs.
- Separation of the effect stop-down of the bottom-up: distinguish between the variations of the intercept and the slope being the latter the top-down effect. Such as manipulations of the response prepriest probability to manipulate the difficulty of perceiving a stimulus.
- Zone recinormal: This analysis assumes that the data are normally distributed, that is, they fall into the Recliner zone of the Fieller distribution ($\lambda_1 < 0.22$). But some assays can lead to lambda values between: ($0.22 < \lambda_1 < 0.4$). Therefore, we could determine the proximity to the Recinormal zone.

References

1. Burle, B., F. V. C. T. T. H. (2004). "*Physiological evidence for response inhibition in choice reaction time tasks*". Brain and Cognition, (56):153–164.
2. Carpenter, R.H.S., Williams, M.L.L. (1995). "*Neural computation of log likelihood in control of saccadic eye movements*". Nature, (377):59–62.
3. Carpenter, R. (1981). "Oculomotor procrastination". D.F. Fisher, R. A. Monty J.W. Senders (Eds.), Eye Movements: Cognition and Visual Perception, Hillsadale, New Jersey: Lawrence Erlbaum Associates, 237–246.
4. Carpenter, R. (1999). "*A neural mechanism that randomises behaviour*". Journal of Consciousness, (6):13–22.
5. Carpenter, R. (2000). "*The neural control looking*". Current Biology, (10):R291–R293.
6. Carpenter, R. (2002)."*Neurophysiology*", 4th Ed.. London: Arnolds, London.
7. Fieller, W. (1932). "*The distribution of the index in a normal bivariate population*". Biometrika, (24):428–440.
8. Moscoso del Prado Martín, F. (2008). "*A fully analytical model of the lexical decision task*". B. C. Love V. M. Sloutsky (Eds.) Proceedings of the 30th Annual Conference of the Cognitive Science Society, (30):1035–1040.
9. Nakahara, H., K. N. . O. H. (2006). "*Extended later model can account for trial-by-trial variability of both pre- and post-processes*". Neural networks, (19):1027–1046.
10. R Core Team (2017).R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Ratcliff, R. (1978).
11. Ratcliff, R. (1978)."*A theory of memory retrieval*". Psycohological Review, (85):58–59.
12. Ratcliff, R., R. C.B. R. (2001). "*Putting noise into neurophysiological models of simple decision making*". Nature Neuroscience, (4):336–337.
13. Reddi, B.R. C. (2000). "*The influence of urgency on decision time*". Nature Neuroscience, (3):827–831.
14. Robinson, D. (1964). "*The mechanics of human saccadic eye movements*". Journal of Physiology, (174):254–264.

Robust wavelength selection using input scaling of filter-wrapper methods on near infrared spectral data of oil palm fruit mesocarp

Divo Dharma Silalahi[1], Habshah Midi[2], Jayanthi Arasan[2], Mohd Shafie Mustafa[2], Jean-Pierre Caliman[1]

[1]*SMART Research Institute, PT. SMART TBK, Riau, Indonesia*

[2]*Institute of Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia*

Abstract

In this study, a new robust wavelength selection based on input scaling method is introduced. The method called Filter-Wrapper method that combines the modified Variable Importance Projection (VIP) and modified Monte Carlo Uninformative Variable Eliminations (MCUVE) to scale the wavelength variable as input factor. The modified VIP uses the orthogonal components of PLS in investigating the informative variable in the model by applying the amount of variation both in $\mathbf{X}$ and $\mathbf{y}$ $\{SSX, SSY\}$, simultaneously. The VIP score then is calculated by using the normalized loading $\mathbf{t}$ from the obtained loading $\mathbf{w}$. Using the VIP score as scaling input method of wavelength variable, the modified MCUVE uses the robust tolerance interval to eliminate the most j^{th} uninformative variable in the scaled input wavelength. In the experiment using simulation data and real data, the proposed method offered some advantages such as improved model interpretability, computationally extensive, and increases the model accuracy.

Keywords

Partial Least Squares, Variable Selection, Variable Importance Projection, Uninformative Variable Eliminations

1. Introduction:

In practice, it is considered if difficult to eliminate all the irrelevant variables and it is also noted if a less number of $\mathbf{X}$ variables used in the calibration will result to the over or under fitting. To overcome this, a new procedure in wavelengths selection using input scaling method based on the combination of Orthogonal Projections to Latent Structures (OPLS)-VIP score and modified UVE is proposed. The scaling method called as mod-VIP-MCUVE (also denotes as Filter-Wrapper method) benefits to guarantee all the wavelengths have equal contribution in the model and improve the convergence speed of the algorithm (Kim et al., 2015; Kim et al., 2016). With related to the Near Infrared Spectroscopy (NIRS) spectral data, the method has benefit to highlight the relevant wavelengths and to downgrade the influence of irrelevant wavelengths in the Partial Least Square Regression (PLSR) model. In the recent work, it has been investigated if only auto-scaling method that mostly applied

in the data pre-processing step. To examine the performance, the proposed method was compared with the classical VIP (Oussama et al., 2012) and MCUVE (Cai et al., 2008) using different datasets. These statistical measures use Desirability Indices (Trautmann, 2004) such as Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Bias, and standard Error (SE) by contrasting the actual (measured **y**) with the results given in the model prediction. This study provides a development for process control in the vibrational spectroscopy technique through wavelengths selection method. Particularly in the chemical analysis, this will assist to a better understanding on the main chemical compositions of the target sample.

2. Input Scaling of Filter-Wrapper Method

Following the OPLS-VIP (Galindo-Prieto et al., 2014), the VIP score measures not only the contribution of each jth wavelength in multivariate models based on the projections to PLS components but also include the orthogonal components. Statistically says the VIP score in the OPLS model considers two amounts of variations that are in response variable $\mathbf{y}$ (SSY) and in predictor variable $\mathbf{X}$ (SSX). There are four versions of OPLS-VIP developed by Galindo-Prieto et al. (Galindo-Prieto et al., 2014), the fourth variants is the recommended one due to its interpretative information ability at the wavelengths that are more relevant both in predictive and orthogonal components. In this study, the fourth variants OPLS-VIP score which considers the combinations $\{\mathrm{SSX},\mathrm{SSY}\}$ in the weighting parameters and normalized loadings $\mathbf{v}_g$ is preferred to be used for the scaling. In line with the earlier PLSR theory which sensible only on the predictive components, with related to the OPLS model then the variations in predictor variable $\mathbf{X}$ is also integrated in the formulation. Here, there are two VIP scores proceed separately for the final OPLS-VIP score which are VIP_{pred} (predictive components) and VIP_{ortho} (orthogonal components). Let redefine g as the predictive component and g_o as the orthogonal component, then l stands for total number of predictive components and l_o stands for total number of orthogonal components with m and m_o are the total numbers of variables used in the predictive and orthogonal components, respectively. The calculation for OPLS-VIP score both in predictive and orthogonal can be written as below

$$\mathrm{VIP}_{pred} = \sqrt{\frac{m}{2} \mathrm{x} \left(\frac{\sum_{g=1}^{l} \left(\mathbf{v}_g^2 \mathrm{x}\, \mathrm{SSX}_{comp;g} \right)}{\mathrm{SSX}_{cum}} + \frac{\sum_{g=1}^{l} \left(\mathbf{v}_g^2 \mathrm{x}\, \mathrm{SSY}_{comp;g} \right)}{\mathrm{SSY}_{cum}} \right)} \qquad (1)$$

$$\mathrm{VIP}_{ortho} = \sqrt{\frac{m_o}{2} \mathrm{x} \left(\frac{\sum_{g_o=1}^{l_o} \left(\mathbf{v}_{o_{g_o}}^{2} \mathrm{x}\, \mathrm{SSX}_{comp;g_o} \right)}{\mathrm{SSX}_{cum}} + \frac{\sum_{g_o=1}^{l_o} \left(\mathbf{v}_{o_{g_o}}^{2} \mathrm{x}\, \mathrm{SSY}_{comp;g_o} \right)}{\mathrm{SSY}_{cum}} \right)} \quad (2)$$

the sum of square (SS) both in variable $\mathbf{y}$ and variable $\mathbf{X}$ has subscript $comp;g$ and $comp;g_o$ for the explained SS of g th component in the predictive and g_o th component in the orthogonal, then the SS with subscript cum for the cumulative explained SS over all components in the model. The total OPLS-VIP score (denotes as VIP-total) then is just a sum for both variable importance projection in predictive and in orthogonal components; or VIP_{pred} and VIP_{ortho}

$$\mathrm{VIP}-\mathrm{total} = \sqrt{\frac{M}{2} \mathrm{x} \left(\frac{\sum_{g_o=1}^{l_o} \left(\mathbf{v}_{o_{g_o}}^{2} \mathrm{x}\, \mathrm{SSX}_{comp;g_o} \right)}{\mathrm{SSX}_{cum}} + \frac{\sum_{g=1}^{l} \left(\mathbf{v}_{g}^{2} \mathrm{x}\, \mathrm{SSX}_{comp;g} \right)}{\mathrm{SSX}_{cum}} + \frac{\sum_{g_o=1}^{l_o} \left(\mathbf{v}_{o_{g_o}}^{2} \mathrm{x}\, \mathrm{SSY}_{comp;g_o} \right)}{\mathrm{SSY}_{cum}} + \frac{\sum_{g=1}^{l} \left(\mathbf{v}_{g}^{2} \mathrm{x}\, \mathrm{SSY}_{comp;g} \right)}{\mathrm{SSY}_{cum}} \right)} \quad (3)$$

M is the total number of variables used in the model or can be defined as the sum of variables used both in the predictive and orthogonal components $\left\{ m = M \Big/ \left(\frac{\mathrm{SSX}_{cum;g}}{\mathrm{SSX}_{cum}} + \frac{\mathrm{SSY}_{cum;g}}{\mathrm{SSY}_{cum}} \right) \right\}; \left\{ m_0 = M \Big/ \left(\frac{\mathrm{SSX}_{cum;g_o}}{\mathrm{SSX}_{cum}} + \frac{\mathrm{SSY}_{cum;g_o}}{\mathrm{SSY}_{cum}} \right) \right\}$.

The total OPLS-VIP score is used to scale the original wavelength variables as the new input matrix. Let define $\tilde{\mathbf{X}}$ as the scaled input variable that is constructed by using the total OPLS-VIP score on predictor variable $\mathbf{X}$ which are not scaled, mathematically it can be written as

$$\tilde{\mathbf{X}} = \mathbf{X}\Omega \quad (4)$$

$$\Omega = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_m\right) \quad (5)$$

where $\Omega \in \Re$ is said to be the diagonal weight matrix with size $m \mathrm{x}\, m$, with the i th element λ_j in the diagonal matrix is a non-negative input variable scaling factor for the j th input wavelength. This $\tilde{\mathbf{X}}$ then is used as new input matrix in the elimination process of MCUVE.

In the MCUVE, the drawbacks of the classical cut-off threshold criterion had been discussed by Centner et al. (see Centner et al., 1996). As alternative, the new modified robust cut-off criterion based on a one-sided tolerance interval from Natrella (1963) is proposed for a better stable elimination on the irrelevant wavelengths. The cut-off value is calculated using the median and the Median Absolute Deviation (MAD) of the reliability coefficients obtained from the added artificial uninformative random variable. In addition, it includes the value of k factor as function of the desired proportions, level of error, and number of repetition used in MC random subsample selection. Using the c_{artif}

in MCUVE threshold, then the new proposed cut-off criterion can be defined as

$$cut\text{-}off\ value = median((c_j)_{artif}) + k\,(MAD(c_j)_{artif}) \tag{6}$$

where k can be calculated as

$$k = \frac{z_\gamma + \sqrt{z_\gamma^2 - ab}}{a} \tag{7}$$

with constant parameters $\left\{a = 1 - \frac{z_\alpha^2}{2(r-1)}\right\}$ and $\left\{b = z_\gamma^2 - \frac{z_\alpha^2}{r}\right\}$; r as number of MC random repetition, α as a level of error, and γ as desired proportion. The wavelengths with reliability c_j less than the cut-off threshold criterion in (7) are moved in the deleted set as D and while the rest wavelengths which are the relevant wavelengths are placed in the remaining set as R. Updating the total OPLS-VIP score in (3) only using the remaining set R then the new scaled input variable in (4) for PLSR model just follows.

3. Result

3.1 Simulation Data

The training set uses 150 samples data and the testing set uses 50 samples data that both were generated randomly using uniform distribution with 0.03 of noise was also applied. The number of input variables and output variable is 40 and 1, respectively. The formulation of this illustrative simulation can be defined as follows

$$\begin{aligned}
\mathbf{c}_j &\sim runif\,(n,1,10) && (j = 1, 2, 3, \ldots, 40)\\
\mathbf{e}_j &\sim rnorm\,(n) && (j = 0, 1, 2, \ldots, 40)\\
\mathbf{x}_j &= \mathbf{c}_j + \mathbf{e}_j\\
\mathbf{y} &= \mathbf{c}_1 + 3\mathbf{c}_5 + 0.85\mathbf{c}_7 + 2\mathbf{c}_{15} + 1.75\mathbf{c}_{22} + 0.9\mathbf{c}_{35} + \mathbf{e}_0
\end{aligned} \tag{8}$$

here, $\mathbf{c}_j$ and $\mathbf{e}_j$ are independent each other and are not measured variables while $\mathbf{x}_j$ and $\mathbf{y}$ are illustrated as observable variables. As seen in (8), there were 6 input variables ($\mathbf{x}_1, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_{15}, \mathbf{x}_{22}, \mathbf{x}_{35}$) related to the response variable, while the remaining 34 input variables were not used in the formulation and were assumed as irrelevant variables. The different coefficients value in the formulation (8) shows the contribution level of each relevant variable to the response variable. It should be considered if these relevant variables were manually selected in the formulation, in fact the importance of input variables is generally unknown. All these input variables are represented as $n \times m$ matrix $\mathbf{X}$ and used in the calculation for model construction. In the PLSR model, the number of latent variables (also called as components) is a principal indicator in the modeling since it may always be

subjective. In the study, to select the optimum number of latent variables in the PLSR model, the result of a re-sampling procedure called cross-validation with lowest standard error from overall best model is used.

As the number of latent variables used in the PLS model increases, the mean and standard error of RMSEP would also decreases. The optimum number of latent variable will depend on how well for certain numbers of original variables have contribution to the model. Using the dataset (see Table 1), it is clear to see if the proposed Filter-Wrapper method using mod-VIP-MCUVE need five latent variables to achieve less RMSEP than VIP scaling method and classical PLS method with no input scaling applied. The MCUVE input scaling method uses similar number of latent variables as like in Filter-Wrapper method but with higher RMSEP value. With this less variable used as predictor in the PLS model, the faster computational speed will be attained. Here, the proposed Filter-Wrapper method has succeeded to reduce the RMSEP and improved the accuracy of the PLSR model. The summarization of the prediction results using training and testing dataset can be seen in Table 1.

Table 1. Statistical measures on prediction results using sine function

Dataset	Methods	LV	RMSEP	R^2	RPD	Bias	SE
Training	PLS	9	0.1330	0.9999	82.3860	0.0049	0.1334
	VIP-PLS	9	0.1437	0.9998	76.2685	0.0052	0.1441
	MCUVE-PLS	5	0.1320	0.9999	83.5957	0.0044	0.1324
	mod-VIP-MCUVE	5	0.1266	0.9999	87.1402	0.0041	0.1270
Testing	PLS	9	0.1547	0.9998	75.8295	0.0075	0.1563
	VIP-PLS	9	0.1544	0.9998	75.9917	0.0223	0.1560
	MCUVE-PLS	5	0.1410	0.9999	83.2257	0.0308	0.1424
	mod-VIP-MCUVE	5	0.1311	0.9999	89.4751	0.0155	0.1325

Comparing the SE and RMSEP values (Table 1), the proposed Filter-Wrapper method both in training and testing dataset produced slightly better accuracy than the other methods which are 0.127 and 0.126, respectively. The reliability on these methods also was examined using the RPD value, the proposed method performs the reliable model compared to the others. It can be appreciated by removing some irrelevant variables in the model the ability of trained model on testing dataset at least comparable to the methods with full variables involved. This shows that when the retained variables in the model is too large, the irrelevant variables contained may influence the model hence decrease the model accuracy.

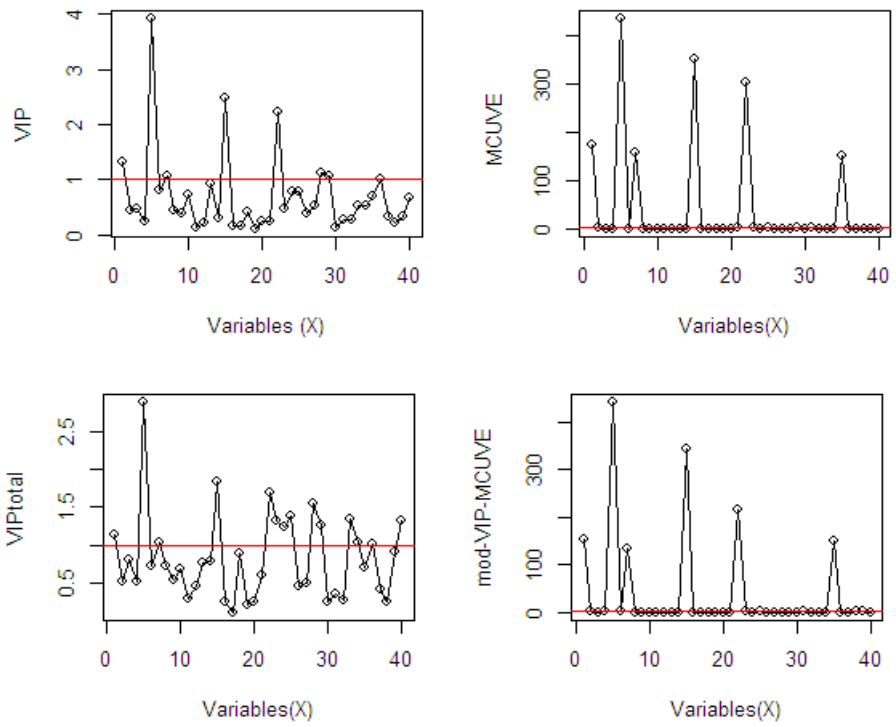


Figure 1. Comparison of the selected relevant variables on sine function data

The most relevant variables selected by the methods in the model were determined based on their cut-off threshold criterion using the score values. This was calculated and afterwards plotted in Figure 1 to evaluate the interpretability of the results. As seen in Figure 1, there were 8 variables ($\mathbf{x}_1, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_{15}, \mathbf{x}_{22}, \mathbf{x}_{28}, \mathbf{x}_{29}, \mathbf{x}_{36}$) with the classical VIP score greater than 1 and considered as the most relevant variables. Comparing with the VIP-total score (denotes as OPLS-VIP), the VIP score vectors both classical VIP and VIP-total provided similar profiles. But in the VIP-total suggested more number of relevant variables (14 variables) than the classical VIP. Using the MCUVE and mod-VIP-MCUVE input scaling method, there were 6 variables ($\mathbf{x}_1, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_{15}, \mathbf{x}_{22}, \mathbf{x}_{35}$) with score values greater than cut-off threshold criterion. Both classical VIP and VIP-total score produced over-selection variables compared to the MCUVE and mod-VIP-MCUVE. The contribution level of the selected relevant variables provided in the MCUVE and mod-VIP-MCUVE input scaling model were also closely comparable to the original formulation as stated in (8). It is clear to claim if using the proposed method in this study, the final subset of selected relevant variables guarantees the best prediction capabilities both in the training and testing dataset.

3.3 Experimental NIRS Dataset

The NIRS spectral data was obtained by scanning the fresh and dried ground fruit mesocarp, right after spectra collection the samples were sent to the laboratory for wet chemistry analysis. The percentage of Oil to Dry Mesocarp

(%ODM) observed in the wet chemistry analysis. The %ODM with range $[56.38, 86.9]$ and standard deviation 5.124 was used as dependent variable in the analysis. In this study, the PLSR analysis was performed with case a single vector of dependent variable $\mathbf{y}$ and processed separately. Total of 960 observations and 488 wavelengths (in the range 550-2500nm) of NIR spectral dataset of fresh mesocarp were used in the analysis. These wavelengths are primarily attributed to the overtone or combination bands of C-H (Fats, Oil, Hydrocarbons), O-H (Water, Alcohol) and N-H (Protein) (Stuart, 2004). This fresh mesocarp sample should be dried and ground before it was sent to the laboratory for conventional soxhlet extraction to get its wet chemistry value such %ODM. In the raw spectra the higher spectral absorbance shows the higher %ODM, while the lower spectral absorbance shows the lower %ODM contained in the fresh fruit mesocarp. Here the importance of the wavelengths was generally unknown and need to be investigated.

Table 2. Statistical measures on prediction results using %ODM data

Dataset	Methods	LV	RMSEP	R^2	RPD	Bias	SE
%ODM	PLS	29	2.967	0.666	1.727	0.180	2.969
	VIP	29	3.011	0.657	1.702	0.241	3.013
	MCUVE	25	3.107	0.633	1.650	0.168	3.108
	mod-VIP-MCUVE	27	3.029	0.654	1.695	0.116	3.021

As seen in Table 2, all the methods provided not slightly different in the statistical measures. With no wavelength selection and no input scaling applied, the conventional PLS method showed a slight better performance compared to the methods with wavelength selection and input scaling applied. The MCUVE used less number of latent variables in the PLS model; this result to the low accuracy in the prediction error of the model since there were many variables most probably removed in the computation (see Figure 2). Oppositely to the proposed mod-VIP-MCUVE, with a different cut-off threshold applied in the wavelength selection and also with less number of latent variables used in the model, the method still provided a slight better performance to the MCUVE-PLS. It was known if as seen in Figure 2, the cut-off threshold in the mod-VIP-MCUVE succeeded to remove only the most irrelevant wavelengths and keep the remaining of relevant variables in the model. This result confirmed the usefulness of the wavelengths selection and input scaling applied in the input variables which attained the faster convergence speed and produced similar accuracy to the conventional PLS.

It can be observed in Figure 2, if all the wavelength selection methods selected the same spectral region which has most relevant contribution to the response variable. But the methods showed different cut-off threshold indicating the irrelevant wavelengths that was not considered informative in the regions. As seen in the selection plot, the VIP, MCUVE method and VIP-total showed many irrelevant wavelengths were removed in the model. As it

assumed earlier if more wavelengths excluded in the model were impact to produce low accuracy in the prediction result. In the fourth plot of mod-VIP-MCUVE, the green line shows the old cut-off threshold using previous MCUVE (threshold = 5.486), while the red line is the new proposed cut-off threshold (threshold = 2.916). The proposed mod-VIP-MCUVE method showed better classification using the modified threshold, since there only less number of wavelengths removed from the model, but the performance was still satisfied and better in interpretability.

The diffuse selected reflectance is very important to identify the relevant wavelengths with related to the %ODM. This exhibit their fundamental attribute to the overtone or combination bands involve the molecular stretching and bending absorption over a wide spectral range. Based on Figure 2, it was possible to observe if the well defined absorption bands are from visible red color (668-684nm), CH_2 of oil and O-H of Water (936-961nm), C-H absorption by stretching-bending (1232-1344nm), first overtone (1404-1444) of C-H stretch O-H of water and C-H of oil and its combinations (1700-1776) of C-O oil and C-H stretching by first overtone, C=O absorption by stretching-bending (1888-2008nm), C-H second overtone of protein and oil (2296-2360nm), and corresponding to absorption which associated with the second overtone in C-H of oil (2364-2496nm).

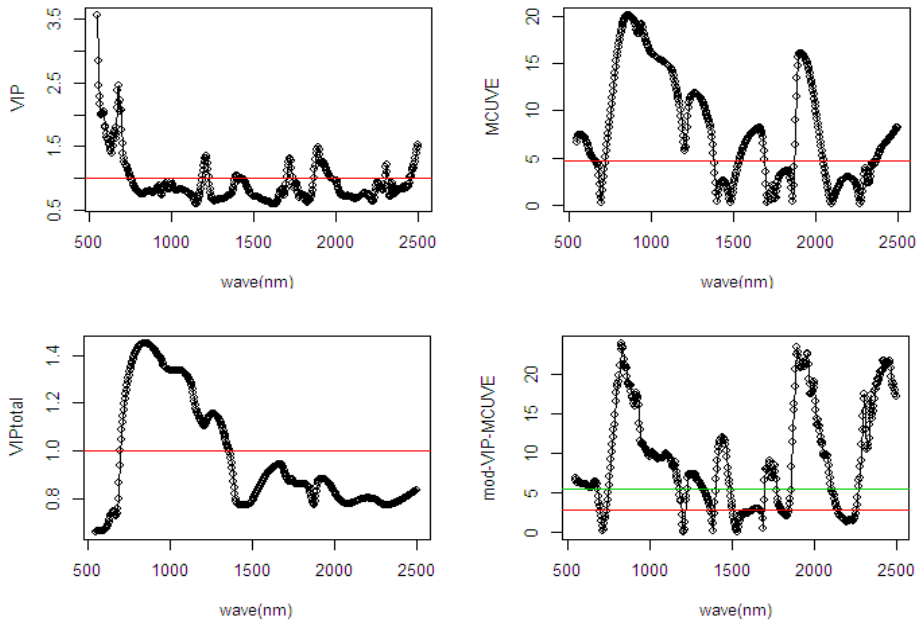


Figure 2. Comparison of the selected wavelengths in the scaled input variables on the NIR spectral data of fresh fruit mesocarp using different methods

4. Conclusion

The study has shown the promising of wavelength selection using input scaling method particularly with application on a high dimension dataset such NIRS spectral data. The proposed method was also robust since it applied a robust measure of central tendency and robust measure of scale in the cut-off threshold calculation. The proposed mod-VIP-MCUVE method has confirmed the superiority to the other reference method such the conventional PLS with no wavelength selection and input scaling applied, the VIP method, and the MCUVE. In the selection of relevant wavelengths, using the modified cut-off threshold the proposed method succeed to remove only the most irrelevant wavelengths in the model, hence it also can still maintained the use of less number of latent variables in the PLS model. Moreover, the proposed method has confirmed the importance of wavelength selection method to reduce the data dimension and to improve the model interpretability particularly to investigate the fundamental attribute of diffuse selected reflectance of NIRS spectral absorption and understanding of the system studied.

References

1. Centner, V., Massart, D. L., de Noord, O. E., de Jong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical chemistry*, *68*(21), 3851-3858.
2. Galindo-Prieto, B., Eriksson, L., & Trygg, J. (2014). Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *Journal of Chemometrics*, *28*(8), 623-632.
3. Kim, J., Kiss, B., & Lee, D. (2016). An adaptive unscented Kalman filtering approach using selective scaling. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 000784-000789). IEEE.
4. Kim, S., Kano, M., Nakagawa, H., & Hasebe, S. (2015). Input variable scaling for statistical modeling. *Computers & Chemical Engineering*, *74*, 59-65.
5. Natrella, M. G. (1963). *Experimental Statistics Handbook 91*. US Government Printing Office.
6. Oussama, A., Elabadi, F., Platikanov, S., Kzaiber, F., & Tauler, R. (2012). Detection of olive oil adulteration using FT-IR spectroscopy and PLS with variable importance of projection (VIP) scores. *Journal of the American Oil Chemists' Society*, *89*(10), 1807-1812.
7. Stuart, B. (2004). Infrared spectroscopy : fundamentals and applications. Wiley: Canada, 78.
8. Trautmann, H. (2004). *The desirability index as an instrument for multivariate process control* (No. 2004, 43). Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen.

Percentile-based approaches for assessing impacts of school day energy expenditure on 18-month change in bmi among elementary school-aged children

Carmen D. Tekwe [1], Gilson Honvoh[2], Roger S. Zoh, Lan Xue [1], Anand Gupta[3], Mark Benden[4]

[1] Department of Epidemiology and Biostatistics Indiana University, Bloomington, IN
[2] Department of Biostatistics, University of North Carolina, Chapel Hill, NC
[3] OhioHealth, Columbus, OH
[4] Department of Environmental and Occupational Health, Texas A&M University, College Station, Texas

Abstract

Childhood obesity is defined based on age- and sex- adjusted body mass indexes (BMI) within upper percentile ranges. Most studies assessing impacts of interventions on BMI rely on traditional linear regression models designed to assess intervention effects on children within "normal" BMI percentile ranges, limiting assessments of how interventions affect children at higher risks for overweight and obesity. Thus, statistical approaches that permit evaluations of intervention effects across the full distribution of BMI are more desirable for determining their impacts on subjects at higher risks for developing overweight or obesity. In this manuscript, we determine the association between energy expenditure obtained at a prior time on subsequent risk or progression to obesity. We describe the use of conditional functional quantile regression models to study the relationship between school day energy expenditure, a function-valued covariate, and BMI. Through empirical comparisons, we present the results from mean regression and quantile regression- based models. The benefits of using quantile regression-based methods in assessing intervention effects in obesity research are also discussed.

Keywords

B-splines, Cluster randomized trial, Functional data analysis, Physical activity, Quantile regression

1. Introduction

About 90% of children diagnosed with type 2 diabetes are either overweight or obesity patients (Liu, et al. 2010). While it is well known that obesity results from a chronic imbalance between energy expenditure and energy intake, as well as from environmental exposures and genetic predisposition, the exact role of energy expenditure in obesity development is unclear (Bandini, et al. 2004). To combat this growing epidemic among children, behavioural researchers are increasingly interested in employing

school-based interventions as targeted interventions designed to reduce sedentary behaviour among children. An example of such behavioural school-based intervention is the activity permissive learning environment (Benden, et al. 2014). Activity permissive learning environments introduce stand-biased desks into classrooms as a means of increasing physical activity among school-aged children. By reducing sedentary behaviour, physical activity behaviour is encouraged during the school day, and devices such as physical activity monitors are used to assess the behavioural patterns of physical activity. These devices provide estimates of school day energy expenditure (SDEE), the total amount of energy or calories expended by the body, to perform physical activity and routine bodily functions during the school day. Overweight and obesity in children are defined based on age- and sex- adjusted body mass indexes (BMI) in the upper percentile ranges. However, most studies assessing impacts of interventions on BMI rely on traditional linear regression models designed to assess intervention effects on children within ``normal'' BMI percentile ranges, limiting assessments of how interventions affect children at higher risks for overweight and obesity. Thus, statistical approaches that permit evaluations of covariates effects across the entire distribution of BMI are preferable for assessing their effects on subjects at higher risks for developing overweight or obesity (Koenker, 1978). Quantile regression is a statistical technique used to estimate effects of predictors on quantile functions of a response. Examples of quantile functions include the median, the 85th and the 95th percentiles of the outcome. A drawback to the use of classical mean regression models in modelling BMI as an outcome is that these methods provide incomplete answers to questions related to BMI values that lie within the tails of its distribution. Additionally, covariates such as SDEE and age may influence the quantile functions differently. Therefore, statistical approaches that allow one to determine covariate effects across the full spectrum of quantile functions of BMI is preferable in obesity studies (Koenker, 1978).

Our current work was motivated by a problem in childhood obesity research. In a recent study, standbiased desks were introduced to three elementary schools in a Texas school district as a means of increasing physical activity. A research question of interest was to determine the impact of SDEE obtained at baseline on subsequent risks for obesity. The recruited children were given BodyMedia SenseWear® armband devices (BodyMedia, Pittsburgh, PA) to measure their energy expenditure during school hours, while sex- and age- adjusted BMI was used as an indicator for obesity. Physical activity monitoring devices are designed to measure the intensity of physical activity. Data from these devices are collected either at the second or minute level over multiple days resulting in high dimensional longitudinal data that appear as curves. Thus, SDEE data are collected over time and can easily be

represented by curves rather than scalar valued summary numbers (Tekwe, et al. 2018). Functional data analysis focuses on the analyses of experimental data collected as curves, functions or images and treats the curves as the unit of statistical analysis (Silverman, et al. 2005).

Parametric regression approaches have been considered in functional data settings (Eubank, et al. 1999). In these settings, the exact forms of the regression curves are assumed known. For example, nonlinear or polynomial mixed effects models can be used to parametrically model the effects of curves on an outcome. However, a limitation of parametric approaches to curve fitting is the requirement of strong parametric assumptions regarding the shapes of the curves. Thus, semi- and non- parametric approaches are standard approaches to analysing functional data. These approaches provide more flexibility for fitting curves to data since they do not require a specific parametric form. Additionally, their abilities to easily accommodate the high dimensionality of functional data is

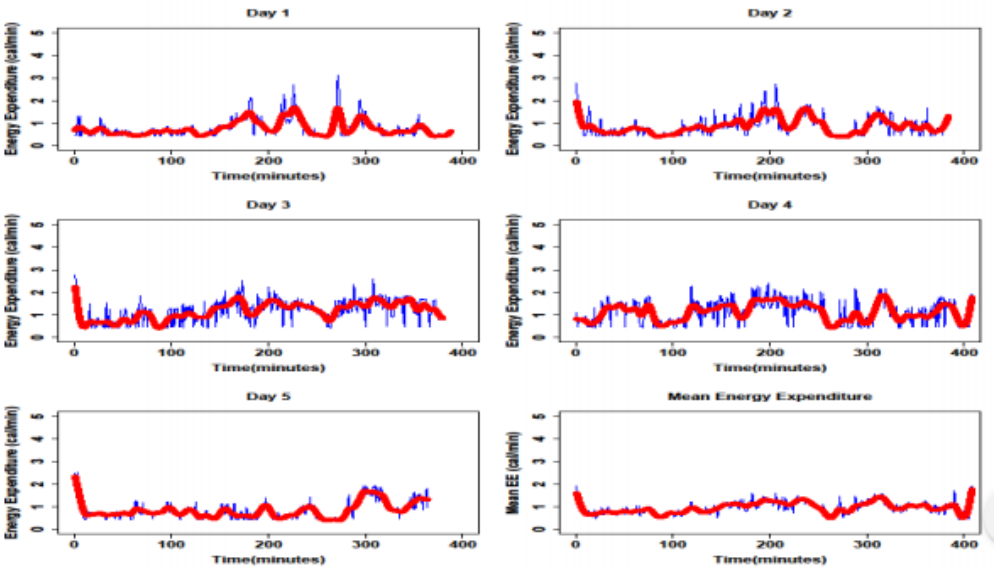


Figure 1. Plot of school day energy expenditure and mean energy expenditure over five days for a randomly selected subject included in the stand-biased desk study.

desirable. As an example, Figure 1 illustrates energy expenditure data gathered about every minute over five school days for a randomly selected student from our motivating example. Data like these are often summarized as a scalar-valued summary statistic such as the mean energy expenditure or the total energy expenditure in their statistical analyses, (see Benden et al., 2014; Wendel, et al. 2016 for examples). Other approaches include summarizing the data from observations taken per minute to hourly mean energy expenditures and subsequently applying standard regression approaches, such as polynomial mixed effect models, (see for example Tekwe,

et al. 2013). However, more complex statistical data reduction techniques such as functional principal components analysis (FPCA) or polynomial basis expansions for approximating the mean of the curves data have also been used (Silverman, et al. 2005). Polynomial basis expansions approximate curves by describing their shapes by a few main features. Thus, an advantage of using polynomial splines is that they summarize the information contained within the curves into basis functions that adequately capture their patterns. Unlike summary statistics, such as the mean, which accounts for only one source of variation in the data, each basis function accounts for a different source of variation in the data. An example of such basis functions includes the B-splines (deBoors, 1978). B-splines do not assume a specific form for the shape of the curves but rather they assume that the individual curves can be approximated by spline functions with random coefficients (Rice, et al. 2001). In Figure 1, nonparametric smoothing was used to approximate the mean of the SDEE. By smoothing the mean, we uncover underlying patterns in the data while also retaining some of its important features (Rice, et al. 2001).

The objectives of this manuscript are two-fold. First, we examine the relationship between SDEE obtained at baseline and future progression towards obesity indicated by measures of body mass indexes at 18 months post-baseline. Secondly, we describe the use of conditional functional quantile regression models to study the relationship between SDEE and BMI, by treating SDEE as a curve or functionvalued covariate after adjusting for relevant socio-demographic variables. Through empirical comparisons, we determine if results obtained from standard approaches used in obesity research such as the multiple linear regression provide notably different results from those obtained from either functional linear regression models or conditional functional quantile regression models. To the best of our knowledge, this is the first comparative analyses focused on determining the usefulness of SDEE as a predictor for subsequent progression towards obesity among elementary school-aged children. The manuscript is organized as follows. In the first section, we briefly describe the data from our motivating example and discuss some limitations of the use of standard regression approaches to assess the association between objective measures of physical activity behaviour and BMI. Next, we provide descriptions of statistical models considered in our applications. We then present the results from our analyses and end with some concluding remarks.

2. Methodology

The stand-biased desks study was conducted from 2012 to 2014 in three elementary schools within the College Station Independent School District (CSISD) (Benden, et al. 2014). The cluster randomized study has been described elsewhere, but briefly, at the beginning of the 2012-2013 academic

year, 24 teachers from eight elementary school were recruited and randomly assigned to the use of either standbiased desks {Stand2learn LLC College Station, TX, USA, stand-biased desk (models S2LK04) and stool (models S2LS04)} or traditional desks (model 2200 FBBK Series by Scholar Craft Products, Birmingham, AL), and chairs (9000 Classic Series, by Virco Inc., Torrance, CA, USA) for in-class activities (Benden, et al. 2014). A total number of 374 students from second through fourth grades were assented and included in the study at baseline. Each student's height and weight were obtained at the start of each semester by trained research assistants to calculate their BMI. The study participants were required to wear calibrated BodyMedia SenseWear® armband devices (BodyMedia, Pittsburgh, PA) during the school hours for a week for each semester from fall 2012 to spring 2014. The devices recorded subject-specific steps counts and caloric energy expenditure per minute while worn. Of the 374 recruited students, 193 students completed the study, while the remaining either graduated from elementary school or their parents retracted their consent from the study. Students with large proportions of missing data were excluded from our analyses. Thus, our final analytic sample size was 157. The study was approved by the Texas A&M IRB. To analyse the data, we considered the linear regression model (LRM), functional linear regression model (FLRM), and conditional functional quantile regression model (CFQRM).

3. Result

The mean BMI at baseline (fall semester of year 1) was 17.14; kg/m2 ; (SD=2.71), while the mean BMI at the end of the study (spring semester of year 2) was 17.55; kg/m2 ; (SD=3.17). The study sample was composed of 77 girls and 80 boys and the average age of the enrolled students at baseline was 7.73; (SD = 0.74) years. About 75.8% of the students were whites, 7% Hispanics, 7% African Americans, and 10.2% Asians/native Americans. In Figure 2, we provide density plots of the log(BMI) at 18 months post-baseline and the residuals obtained after adjusting for baseline covariates and clustering. The skewness and long tails of both distributions indicate a possible violation of the normality assumption from linear regression models. Prior to fitting the mixed effects model considered, we computed the log of BMI at both baseline and 18 months post baseline. Through AIC comparisons, we determined that the log transformation of BMI provided a better fit than the inverse and square root transformations. Next, mixed effects models were used to obtain the baseline and cluster randomized adjusted residuals for BMI. Overall, we did not find a significant impact of age at baseline on the BMI values at 18 months post baseline (p=0.209). However, there were statistically significant differences between African Americans and whites (p= 0.05). The random

intercept for the nested effects of teachers within schools was statistically significant (p=0.04).

Results from LRM

The LRM summarizes the high dimensional measures of SDEE per subject to a scalar-valued measure. This summary scalar-valued measures were obtained by computing the arithmetic mean of all measures of SDEE by subject.

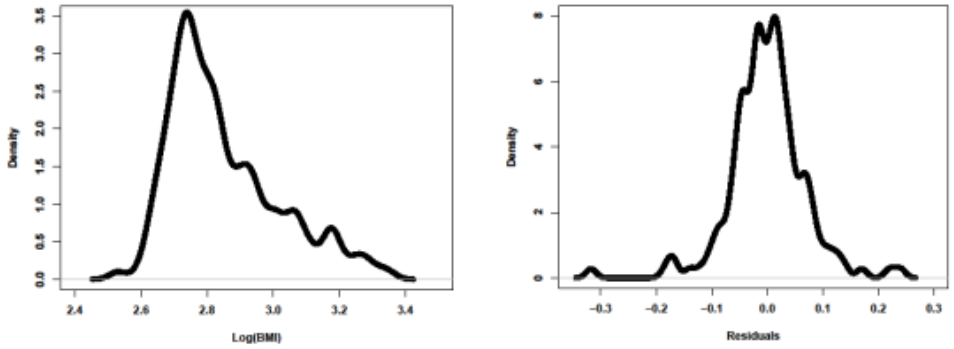


Figure 2. Density plot of BMI distribution 18 months post baseline (Figure 2a) and density plot of the baseline covariates adjusted residuals (Figure 2b). The skewness of the distribution of the BMI outcome and the adjusted residuals are outlined in the two plots.

From the LRM results, we concluded that mean SDEE was not statistically predictive of BMI at 18 months post-baseline ($\widehat{\beta}_{1} = 0.03$, 95% CI: -0.00, 0.06, p=0.064). Thus, application of the LRM indicated that the overall mean energy expenditure obtained at baseline could not be used as a predictor of future values of the conditional mean of BMI after adjusting for the socio-demographic covariates (p=0.064). The LRM produced an AIC of -396.6. While the use of overall mean SDEE to represent patterns of school day physical activity behaviour results in loss of information, functional regression models correct for this loss of information by using the full profile of the function-valued covariate in the estimation process.

Results from FLRM

Energy expenditure measures obtained at baseline were summarized using four basis functions. The final number of basis functions were selected by comparing the AIC values from the FLRM under varying number for the basis functions. The computed AIC values ranged between -392.2 and -384.9, with the lowest value of -392.2 achieved with four basis functions. The basis functions were subsequently used as explanatory variables for SDEE in fitting the FLRM. Once the model was fitted, SDEE was considered statistically significant when all estimated coefficients of the basis functions yielded small

p-values(p0.05 for all of the estimated coefficients). Point-wise bootstrap confidence intervals were obtained at the 95% confidence level. The estimated functional coefficient illustrates the curvilinear patterns of SDEE over time, indicating that the patterns of physical activity is not constant across time. Thus, the FLRM provides more flexibility in the estimations when compared to the LRM in our application. The confidence intervals also support the conclusion that there is insufficient evidence in our data to indicate that SDEE is predictive of BMI values at 18 months post baseline among the children.

Results from CFQRM

At each quantile, measures of energy expenditure were reduced to linear combinations of splines and basis functions. Similar to the FLRM, the final numbers of basis functions were selected by comparing the AIC values computed under varying numbers of basis functions at each quantile. The AIC comparisons led to the choice of Kn=;4 at the 10th, 50th and 85 th quantiles, Kn =6 at the 25th quantile, while Kn=7 was selected at the 95th and 99th quantiles. We did not detect any statistically significant associations between SDEE and the conditional quantile functions of our response across all the quantile regressions considered ($p>0.05$ for all the spline coefficients). Figure 3 provides plots of the estimated functional coefficients on SDEE and their corresponding 95% point-wise confidence intervals. The plots also illustrate the patterns of physical activity behaviour across time under each quantile regression. Varying patterns in the physical activity behaviour were observed under the six quantile functions.

4. Discussion and Conclusion

Three regression-based methods were used to investigate the impact of baseline SDEE on BMI values at 18 months post baseline among elementary school-aged children recruited from a Texas school district. Using the LRM, we assessed the impact of overall mean SDEE on the outcome of interest. A potential disadvantage of this approach is that it does not account for potential diurnal patterns of physical activity behaviour and the focus of the analyses is on assessments of the covariate on the conditional mean of the outcome. Using splines in the FLRM provided more flexibility by modelling the objective measures of SDEE as curves, while the outcome was also the conditional mean of BMI at 18 months post baseline. We note that unlike the spline methodology employed, the use of overall mean SDEE to represent physical activity behaviour at baseline resulted in loss of information. While both the LRM and FLRM enable evaluations of covariate effects on the conditional mean of BMI, the CFQRM enables assessments of covariate effects across the entire distribution of the outcome.

Based on our analyses, none of the fitted models detected statistically significant associations between SDEE at baseline and BMI at 18 months post baseline. By ad hoc comparisons of the AIC values obtained under all the models considered, the AIC value associated with the 50th quantile function of BMI was the lowest (AIC = -416). The AIC values for the LRM and FLRM were -396 and -392, respectively, indicating that the use of polynomial splines to represent SDEE did not necessarily provide a better model fit when the two conditional mean-regression based methods are compared. However, we note that the AIC values associated with the quantile regression-based models varied from -218 to -416. No statistically significant associations were also observed for the impact of SDEE on the conditional quantile functions of BMI under all the quantile functions considered. The AIC values obtained for all the quantile functions except for the 50th were either comparable to the values obtained under the mean regression-based methods or larger. Based on our analyses, the use of the CFQRM at the 50th level provided the best model fits when compared to all the other models considered.

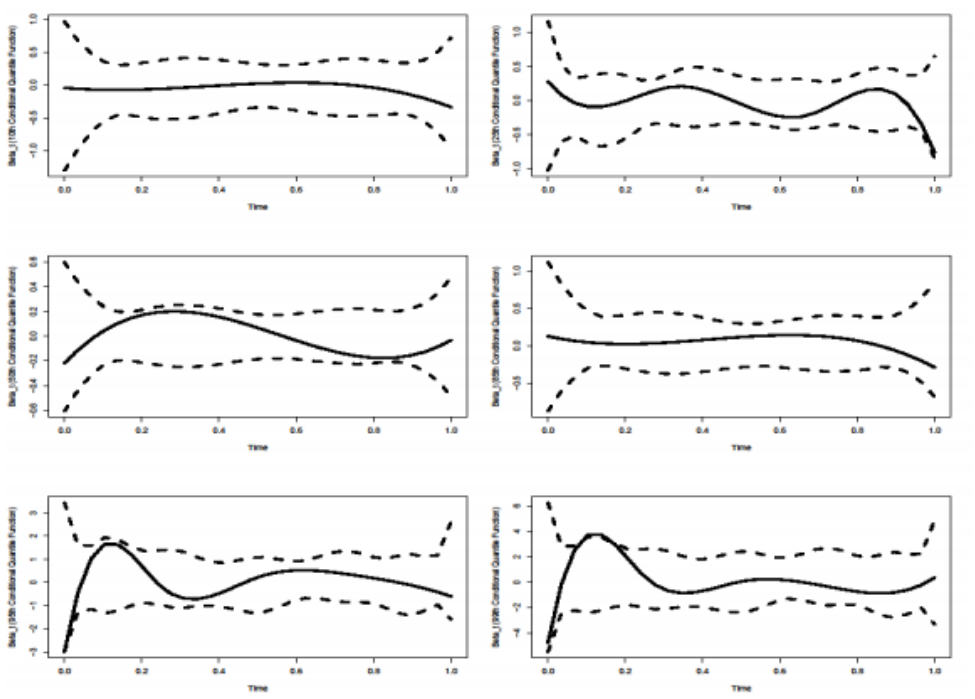


Figure 3. Plot of the estimated functional coefficients and their corresponding 95\% point-wise bootstrap confidence intervals at the 10th, 25th, 50th, 85th, 95th and 99th quantiles. For each plot, the solid line indicates the estimated quantile specific functional coefficient, while the dashed lines represent the upper and lower bounds of the confidence intervals.

References

1. Liu LL, Lawrence JM, Davis C et al. Prevalence of overweight and obesity in youth with diabetes in USA: the search for diabetes in youth study. Pediatric Diabetes 2010; 11(1): 4–11.
2. 2.Bandini LG, Must A, Phillips SM et al. Relation of body mass index and body fatness to energy expenditure: longitudinal changes from preadolescence through adolescence. The American Journal of Clinical Nutrition 2004; 80(5): 1262–1269.
3. Benden ME, Zhao H, Jeffrey CE et al. The evaluation of the impact of a stand-biased desk on energy expenditure and physical activity for elementary school students. International Journal of Environmental Research and Public Health 2014; 11(9): 9361–9375.
4. Koenker R and Bassett G. Regression quantiles. Econometrica: Journal of the Econometric Society 1978; 33–50.
5. Tekwe, C. D., Zoh, R. S., Bazer, F. W., Wu, G., & Carroll, R. J. (2018). Functional multiple indicators, multiple causes measurement error models. Biometrics, 74(1): 127-134.
6. Tekwe CD, Lei J, Yao K et al. Oral administration of interferon tau enhances oxidation of energy substrates and reduces adiposity in zucker diabetic fatty rats. BioFactors 2013; 39(5): 552–563.
7. Silverman B and Ramsay J. Functional Data Analysis. New York, NY: Springer-Verlag, 2005.
8. Eubank RL. Nonparametric regression and spline smoothing. New York, NY: CRC press, 1999.
9. Wendel ML, Benden ME, Zhao H et al. Stand-biased versus seated classrooms and childhood obesity: A randomized experiment in Texas. American Journal of Public Health 2016; 106(10): 1849–1854. 18.
10. de Boor C. A practical guide to splines. New York, NY: Springer-Verlag, 1978.
11. Rice JA and Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 2001; 57(1): 253–259.
12. Trinh A, Campbell M, Ukoumunne OC et al. Physical activity and 3-year BMI change in overweight and obese children. Pediatrics 2013; 131(2): e470–e477.

Determinants of pupils' success in primary schools in Benin

Felicien Donat Edgar Towenan Accrombessy[1], Marc Luc Dagbégnon Akplogan[2]

[1]Senior Statistician Economist (World Bank), facrombessy@worldbank.org, Cotonou, Benin
[2]Marc Luc D. Akplogan Statistician Economist, Consultant, marc.akplogan@yahoo.fr

Abstract

This paper aims to identify the factors of success of children in the primary school (PS) in Benin by using transversal analysis and hierarchical linear model. Main findings reveal that the following factors are correlated with the performance of pupils in 2004 and 2014: (i) repeating students perform less well than non-repeaters, (ii) older students are less successful, (iii) inequalities in performance exist according to the gender of the pupil; (iv) the socio-economic status of families is positively related to students' academic performance, (v) the practice of the language of instruction outside school is a factor of success. (vi) students' performance at the beginning of the school year is positively correlated with the end-of-school scores; (vii) participation in extracurricular work (trade, physical or agricultural work) penalizes the acquisitions of students; (viii) public schools are less efficient than private schools; (ix) a significant number of pupils per class is negatively correlated with their learning; (x) teacher training did not have a significant impact on children's learning but women-led schools have better scores than their counterparts.

Keywords

Hierarchical linear model; Pupils; Primary School; Achievement ratio; PASEC test scores.

1. Introduction

The main issues related to the quality of teaching have remained central to the debates of various international organizations on Education for All (EFA) since Jomtien's call in 1990. It is recognized that, the lessons accumulated by students at primary school, especially in writing, reading and mathematics, have important repercussions when they enter secondary school, higher education or begin a professional life. In addition, the quality of education largely determines a country's ability to cope with future challenges in all fields. Indeed, the countries that have experienced a high level of economic performance in recent decades are those that have been able to put in place efficient and quality education systems (Singapore, Taiwan, Mauritius, Botswana, Rwanda, ...).

In Benin, several efforts have been made to raise the level of education. However, although perceptible progress in terms of access to primary school education, as evidenced by the significant enrollment ratios, the country faces more difficulties in terms of the quality of education (RESEN, 2014). The situation of the teaching of writing, reading and mathematics in Benin reveals gaps in quality: Benin is largely behind other comparable countries. At the beginning of schooling, the national average scores in reading are 458.3 points and 454.7 points in maths falling below the average of the ten countries surveyed in 2014 by the PASEC, set at 500 points. At the end of schooling, the national average scores in reading (523.4 points) and maths (496.9 points) are close to the average of the ten countries.

Nevertheless, these scores remain lower than those of several comparable countries such as Senegal and Cameroon. To help understand this problem and to find out the right policies to achieve the SDGs, this paper aims to analyze the factors that influence school performance in language and mathematics among students in grades 2 and 5 of primary school. The main lessons of this study are to inform policy makers by helping them to identify the reasons for poor pupil's performance in Benin and thus to better define the corrective educational policies to be implemented. Starting from the fact that a quality school is one where quality teaching is provided, the quality of education is measured by the achievement of students quantifiable by their school achievements based on their PASEC14 tests scores. Thus, the present study exploits all the PASEC data for the years 2004 and 2014 applied to Benin, on samples of 1705 pupils of 2nd grade and 1823 pupils of 5th grade of primary school belonging to 273 schools. We grouped the variables available into three categories: individual variables, family variables and those related to the school context.

To consider the hierarchical structure of the data, because of survey sampling design, instead of using OLS (Ordinary Least Square) linear models that have limitations, more advanced multilevel models (Michaelowa (2000) and Bressoux (2010)) are preferred to link inputs from the educational production function and the outputs. The econometric analysis of the explanatory factors of pupils' academic performance will answer the following questions: What are the individual and school context characteristics that contribute positively to the differences in student scores? Are there links between the two levels of the analysis? Are the main explanatory factors of school acquisitions the same over the two periods of analysis (2004 and 2014)? What lessons does the interaction between some key variables of the two levels of analysis on student achievement provide?

[14] PASEC is the CONFEMEN Program for the Analysis of Education Systems. CONFEMEN is the Conference of Ministers of Education of French-Speaking Countries

2. Methodology

First, the primary school achievement trend was plotted and PCA depicted performance profiles by groups of locations and sorted determinants of pupils' performance, and to inform variables included in the school performance determinants regression.

The 2013 population and housing census informed primary school attainment and completion. Primary school completion was calculated using the proportion of the population aged 12 years and plus that completed at least the primary 6 school grade. The ratio is calculated by age group and by birthyear.

Principal components analysis (PCA) is an extremely powerful statistical tool used to synthesize information, rather than performing a graphical representation for each of the primary education indicators (from the administrative source). PCA calculates main components concentrating most of the information in the matrix of original indicators in order to distinguish clearly the school performance of municipalities (communes), and to classify them into shorts groups with common characteristics. The advantage is to provide policy makers with evidence on levers to activate in order to improve the performance of the education system by targeting the intervention based on the evidence provided by each group of communes now recognized by their particular profile that the PCA drew. The statistical units considered to realize the PCA are the 77 communes[15] of Benin. The variables are the indicators of primary education. The following set of 16 variables includes:

- seven (7) ratios related to student access and performance in the education system (Gross Enrollment Ratio (GER), Gross Admission Ratio (GAR), Completion Ratio (TAC), Net Enrollment Rate (NER), Promotion Ratio (PR), Drop-out ratio (DR), CEP admission ratio, the percentage of graduates after the national exam to validate primary school grade 6 completion,
- four (4) ratios (number of pupils per teacher ratio, number of pupils per classroom ratio, number of pupils by instructional group ratio and number of textbooks per pupil ratio), and
- proportions and averages (percentage of qualified teachers, percentage of civil servants (state agents), percentage of pupils sitting

[15] Benin is at administrative level divided in 12 regions (the Departments). Each Department is divided in sub-regions called "Communes", 77 in total. The Communes are composed of Arrondissements, and the important level of geographic administrative decomposition is the Village or City unit (Quartier de ville). In the analysis, we preferred to do the PCA at the Commune level, because it offer significant policy decision incentive particularly regarding the advanced decentralization process in Benin.

on tables-benches of 2 places at most, percentage of pupils sitting on benches and the average number of blackboards per classroom.

Hence, evidence of the descriptive analysis and PCA informed the regression. The theory and empirical literature on the determinants of students' academic school performance is relatively abundant. A number of factors (inputs) combine a series of activities, practices and conditions to produce school output. Academic performance studies identify three categories of variables as factors of the production function in education. The first category of inputs relates to the individual characteristics of pupils: gender, age, cognitive structures (intelligence, motivation, self-perception), etc. Then, the second category of inputs concerns the variables related to the family environment of the pupils such as: the level of education of the parents, the availability of capital goods and educational material within the household (computer, dictionary, textbooks, ...), the language used at home by the family, the size of the family, the child's participation in domestic or rural work, etc. Finally, the variables related to the school context consider the characteristics of the teacher (gender, training, motivation, ...) and those of the school such as class size, equipment, pedagogical practices and organizational characteristics. Sociocultural variables, particularly those related to the family environment, influence the child's academic success (Duru-Bellat 2003, Diallo 2001, Fuchs et al 1999). Estimates show that school results are better for children whose parents are educated. The same conclusions are reached in the case of Haiti. On the other hand, in the case of Morocco, Hijri et al, 1995 show that this relation is not significant. This result can be explained by the fact that women with a high level of education were easily engaged in professional life and entrusted the care and support of their children to housekeepers, often without any level of education. In Benin, based on the 2014 survey data and linear descriptive modeling, PASEC assessed success factors at the end of primary schooling. As a limit, this study does not consider family and other individual variables that may be involved in explaining these differences. For example, the availability of capital goods and teaching materials within the household, the language of the family, the size of the family, the skill level of students entering PS, the amount of time spent for homework, student engagement in learning, etc. are also key factors.

Our analyses focus on PASEC data collected in 2004 and 2014 on student learning outcomes in primary education in Benin according to standards that facilitate comparison between CONFEMEN countries. Since a multiplicity of factors acts simultaneously on school performance, we retain the variables to which the literature attaches great importance and that the PCA put emphasis on, including the three categories of variables mentioned above. The existence of two levels of analysis within the model poses the problem of non-compliance with the two essential assumptions of the ordinary least squares

(OLS) approach, namely the independence of observations and homoscedasticity (Snijders and Bosker, 1999 and Bressoux, 2007). An attempt to estimate this hierarchical model by the OLS method introduces bias in the implementation of the tests. Relationships between learning outcomes and higher-level characteristics can be illusory.

Multilevel models help overcome the shortcomings of the OLS method in remedying the problem of homoscedasticity, by introducing a lower hypothesis, according to which the variance of the residuals can vary as a linear or non-linear function of the explanatory variables. Following Bressoux, 2010, the models with the performance of pupils (Sij), q characterizing variables (Xqij) of the pupil (level 1) who studies in school j (level 2) with its characteristics (Wlij) are:

$$S_{ij} = \beta_{00} + \left(\sum_{q=1}^{Q} \gamma_{q00} X_{qij} + \sum_{l=1}^{L} \beta_{l00} W_{lj} \right) + \left(\sum_{q=1}^{Q} \sum_{l=1}^{L} \gamma_{ql00} W_{lj} X_{qij} \right) + \left(\varepsilon_{ij} + \varepsilon_{0j} \right)$$

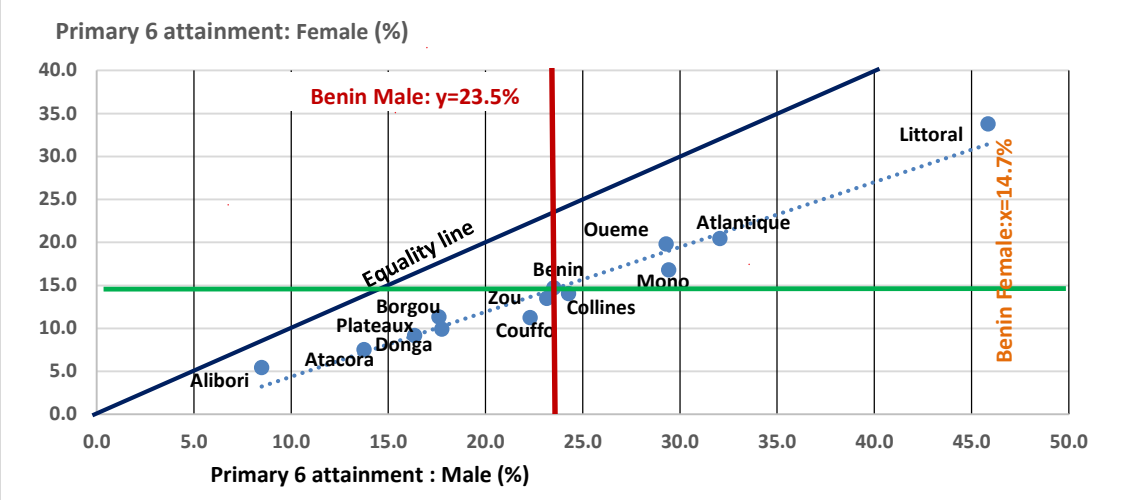
3. Results

a. Primary education achievement trends

According to census data, primary education attainment and completion ratios improved in recent years. While seven people out of ten had no education in 2002, this number decreased to 53 percent in 2013. At the same time, the enrollment rate at the post-primary level doubled from 10.2 percent in 2002 to 24.2 percent in 2013. Now, students have more opportunities than in the past to continue their education studies after grade 6. Nevertheless, beyond this success, several challenges remain from the point of view of equity. Completion rates by gender and region show significant disparities of more than two-fold disadvantage for women and rural areas. The completion rate for men is 23. percent compared to 14.7 percent for women.

Depending on the location, in urban areas, this ratio was 27.2% against 12% in rural areas. Inequalities are less pronounced in regions with low completion rates, most of which are located in the northern part of the country. On the other hand, in the south and the center regions completion rates stood above the national rate. When we consider the primary school grade 6 completion at the national level, four regions substanstially stand other the 19.0 percent, namely Littoral (+20.6 percentage points), Atlantique (+7.2 percentage points), Oueme (+5.4 percentage points) and Mono (+3.9 percentage points).

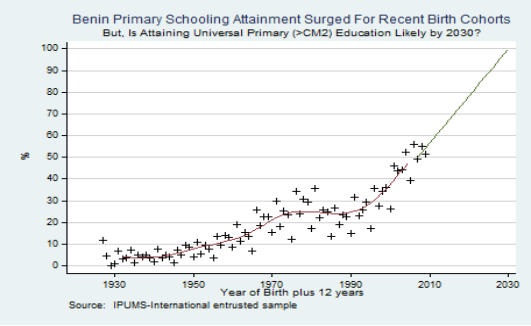
Figure 1: Primary school attainment by sex and department in 2013



Source: Population and housing Census 2013, Author's calculations

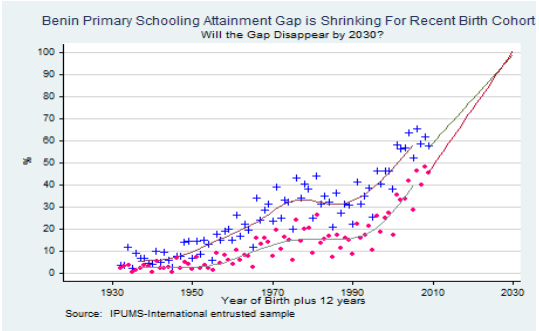
From a gender approach, the remarks are still the same. Male primary school grade 6 completion is higher than the national level (23.5 percent) in Littoral (+22.3 percentage points), Atlantique (+8.6 percentage points), Mono and Oueme (repectively +5.9 and +5.8 percentage points. Similarly, female primary school grade 6 completion is higher than the national level (14.7 percent) in Littoral (+19.1 percentage points), Atlantique (+5.7 percentage points), Oueme and Mono (repectively +5.1 and +2.1 percentage points). To achieve gender equity and reach the completion rate of 100% in 2030, the overall completion rate must increase annually for men by 1.7% and 2.3% for women and the overall annual increase of 2%.

Figure 2a: Benin primary schooling attainment surged for recent birth cohorts



Source: Population and housing Census, 2013

Figure 2b: Benin primary schooling attainment gap is shrinking for recent birth cohort



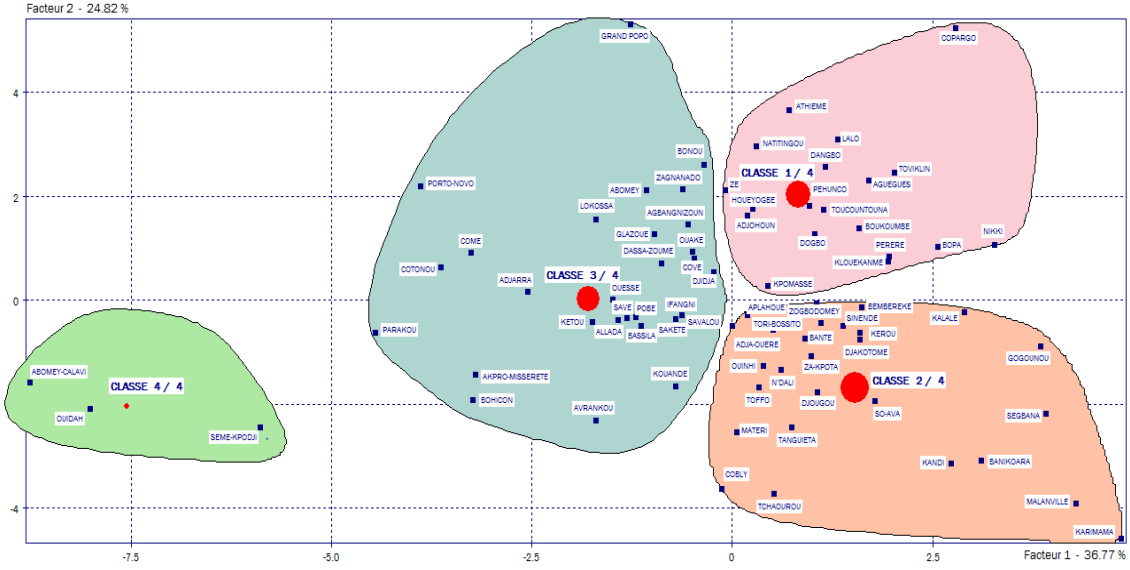
Source: Population and housing Census, 2013

In Benin, the financing of the school is essentially the responsibility of the State and its partners in development. However, households contribute to the education of their children through enrollment in private schools that are growing in urban areas.

b. Primary school performance by location: A principal components analysis

We observe strong positive relationship between the indicators related to access to school and negative relationship between these variables and the dropout ratios and the percentages of qualified teachers. More than three quarters of the variance (75.8 percent) of the total information are explained by the three first axes. The first four axes, with eigenvalues greater than unity, account for 82.2% of the information. We notice that there are positive correlations with the indicators of the quality of teachers and negative correlations with the indicators of school access and success to primary school grade 6 completion exam. The output allowed to categorize the municipalities in four distinct profiles as shown in the figure below, and exploit this information for the regression analysis.

Figure 3: Factorial map of the Communes



Source: Author's calculations, using Ministry of primary education administrative data

c. Model estimation and interpretations

The results of the estimation show that inter-school variance represents 63.9 % of the total variance of the final pre and post-test scores in 2004. This variance is large and significantly different from zero (p-value <0.001). Also,

for the 2014 data, the inter-school variance represents more than half of the total variance of the end-of-school score in reading and mathematics. The models are globally significant and the variables are introduced in a progressive process. The main findings are: Repeaters are less efficient than non-repeaters: Repetition hurts the performance of the education system, in addition to the fact that it costs for the whole system (parents and authorities). Repeating students' progress less quickly than others during the year (PASEC 2004-2005). The results show that pupils who repeat at least once are significantly less effective at the end of their schooling than those who have never repeated according to the output of the model estimated on the 2014 data. Older students are less successful in their learning. The results point out that older students are at a disadvantage in their learning: 2004 and 2014 results indicate that students' ages are significantly and negatively associated with their academic performance in reading and mathematics. The older the students, the less they assimilate the concepts. There are also inequalities of performance by gender of student: The relationship between student gender and school performance varies by the PASEC year considered. According to the 2004-2005 PASEC data, under the control of other explanatory variables, girls are on average less performing than boys in language-reading and mathematics. In 2014, the trend is reversed: inequalities of performance in reading and mathematics are in favor of girls at the end of their schooling. This observation could be explained by the policy in favor of girls set up throughout the country. In addition, girls in urban area outperform than those living in rural area. The fact that girls are taught by women has no significant effect on their acquisitions. Then, the socioeconomic status of families is positively related to student achievement. The standard of living positively affects the level of students at the end of the year. The interaction between the socioeconomic level of families and the location of schools is negatively linked to school attainment. Also, in relation to school performance, the results of the linear hierarchical model estimates show significant differences in scores to the detriment of students who do not speak the language of instruction (French) in their family environment. As other studies of education in Africa show, the practice of the language of instruction enables children to acquire language skills before they enter school and during primary school cycle. More, students' performance at the beginning of the school year is positively correlated with the end-of-school scores: Students with one point of difference in their initial levels are, all other things being equal, with a difference of 0.55 at the end of the year. Since the coefficient is positive and less than unity, this indicates that learners who are most successful in their acquisitions remain those who have a high level at the beginning of the year and that the gap between students is narrowing towards the end of the year. Otherwise, the participation of students in the field work, trade and physical

work affects negatively their learning. These out-of-school activities are often the source of absenteeism and school drop-out for students. In contrary, the fact that the student does domestic work, less penalizing than those mentioned above, positively influences school achievement. Public schools are less efficient than private ones: On average and all things being equal, there is a negative and significant relationship between attendance at a public school and student performance. The coefficients associated with this variable are relatively high and reflect the differences observed between the two types of schools. Students in private schools continue to perform better than those in public schools. This finding could be explained by the organization of private schools, rigorous supervision of pupils and the socio-economic level of the parents. High number of students per class is negatively correlated with learning: Students in high size classes have on average lower scores than students in lower size classes. However, the differences in scores are not significant. A positive role of class and school equipment: The school equipment and resources influence the students' academic performance. Students whose schools are better equipped, and located (an asphalt road, electricity, a hospital, a health center, a police station, a bank, a post office and a cultural center or library) in more developed environment are better performers than their counterparts. In addition, the teacher's level or professional training of more than one year does not improve the scores of their learners. This raises the problem of the quality of teacher basic education level/training. In addition, the use of guides (teacher's book) in maths and french does not have a significant influence on the educational attainment.

4. Discussion and Conclusion

Primary school attainment and completion improved substantially the last three decades, since the 1990s. Both administrative data and censuses/household surveys show this evidence. The indicators of access in the primary education improved among others thanks to the different policy reforms occurred in the primary school system, including free schooling for girls. A typology of communes of Benin according to the indicators of the education system shows the schools of the localities located in the northern part of the country are less efficient than their counterpart of the South.

The analysis in this paper makes a substantial contribution to the literature in two substantive ways. First, capacity building of teacher is a necessity. Second, since repeaters fail to fill their academic gap with non-repeating students, specific follow-up is important for these categories of student to reach the performance of non-repeaters. The analysis in this study generates robust results that consider the specificities of class and school environment and are more informative for policy formulation to attend the Goal 2 of SDGs in Benin in 2030.

References

1. Bressoux Pascal, (2007) L'apport des modèles multiniveaux à la recherche en éducation, Éducation & Didactique, 1, 2, p. 73-78;
2. Duru-Bellat Marie, (2003), Les apprentissages des élèves dans leur contexte : les effets de la composition de l'environnement scolaire, Carrefours de l'éducation 2003/2 (n° 16), p. 182-206;
3. PASEC, (2016). PASEC 2014 - Performances du système éducatif béninois : compétences et facteurs de réussite au primaire », 2016;
4. G. Feeney: Literacy and Gender: Development Success Stories, The Population Council, Inc. Data and perspectives and development review 2014;
6. E. Smith-Greenaway: Educational attainment and adult literacy: A descriptive account of 31 Sub-Saharan Africa countries, 2015 Demographic research Volume 33, article 35, pages 1015–1034, 2015
7. K. Gyimah-Brempong, Education and Economic Development in Africa, African Development Review, Vol. 23, No. 2, 2011.

Logistic model averaging for predicting type of tumor in high dimensional data: A randomize approach

Septian Rahardiantoro, Anang Kurnia
Department of Statistics, IPB University, Indonesia

Abstract

The main idea of model averaging is to combine some predictions of model candidates to be the final prediction using specified weight. It is very commonly used in high dimensional data that number of predictors more than number of observations. In application, the model averaging concept also can be used in the prediction of class of response variable. This research applied the model averaging concept using logistic regression model for predicting the class of patients having different types of tumor: KIRC and LUAD. The data set is a part of the RNA-seq, that contain the collection a random extraction gene expression with dimension 20532 gene belong to 287 patients. The model candidate of logistic regression constructed by selecting randomly the gene with size: 50, 100, and 150; to predict the class of patients. Based on the evaluation criteria, lower value of size gene in the logistic model could reach higher accuration, sensitivity, and spesificity of prediction.

Keywords

classification; high dimensional data; logistic model averaging; model candidate; predictive modelling

1. Introduction

High dimensional data happens when the number of features (p) in data exceeds the number of observations (n). In recent century, high dimensional data is very commonly found in many part of life. It can be found in social media data set, genomic data set, econometric data set, and also in satellite data set. Because of the size is very big, the main challenge when deal with this data is in prediction of responsce context. There are some methods that often to used for handling this case, such as best subset selection, lasso regression, and model averaging.

This research attempts to apply the model averaging approach to handle the prediction case in high dimensional data set. The main principle of model averaging is to construct some model candidate that would be everaged to be the final model [1]. The model that commonly used is linear regression that also based on the scale of response variable. There is an application of model averaging using linear regression to predict the response variable in the genomic data set [2]. Furthermore, the development of this method is well

developed to use the logistic regression in the model averaging process when the categorical scale in response variable [3].

This research foccus on constructing the model candidate of logistic regression in case of prediction class of response variable. The model candidate is cunstructed by selecting the predictor variables randomly to get the prediction of response variable class. This process applied several times to get some prediction and then the prediction would be averaged using the specified weight. In this case, the probability form is used in the prediction of response variable to be avaraged.

The data that used in this research is RNA-seq data set that part of a random extraction gene expression of patients having different types of tumor: KIRC (Kidney Renal Clear-Cell Carcinoma) and LUAD (Lung Adenocarcinoma). This data set contains 20532 gene based on 287 patients [4]. The number of gene selected in model candidate is 50 genes, 100 genes, and 150 genes with number of model candidate contain 50 models. In practices, there are selected about 40% part of patients to be the testing data to evaluate the accuration, sensitivity, and spesificity of prediction.

2. Methodology

In this section would be described the data set that used in this research, the model averaging concept in logistic regression approach, and also the evaluation of the prediction.

2.1 Data

The data set that used in this research is a part of The Cancer Genome Atlas (TCGA) Research to profile and analyze large numbers of human tumors to discover molecular aberrations [4]. This research took the subset of this data on patients having KIRC and LUAD based on their RNA-seq. Therefore, the response variable of this research is class of patients based on their suffered. The number of patients in this data is n = 287 patients with p = 20532 genes to be the predictor variables, which is include in the high dimensional data (p≫n). For data analysis, the class KIRC simbolized by 1 and LUAD is 0.

2.2 Model Averaging

Let $\boldsymbol{X}_{n\times p}$ is high dimensional data with number of observations n and number of predictors p ($p \gg n$), and $\boldsymbol{X}^{*}_{n\times m}$ is the subset of $\boldsymbol{X}$ with number of predictors m ($m < p$). Let $\boldsymbol{y}_{n\times 1}$ is the response variable in the case. Assume the regression model of the subset predictor data is $\boldsymbol{y} = f(\boldsymbol{X}^{*}) + \varepsilon$. The model averaging concept is creating some model candidates or the subset predictor model to combine to be the representative form of final model. The number of model candidates is k which contains m predictors in each model.

Therefore, the prediction of the each model candidate can be formulated below.

$$\hat{\boldsymbol{y}}_i = f(\boldsymbol{X}^*); i = 1,2,\dots,k$$

Then, the averaging of the model candidate is

$$\hat{\boldsymbol{y}}^{MA} = \sum_{i=1}^{k} w_i \hat{\boldsymbol{y}}_i$$

with w_i indicates the weight of i-th model candidate, and $\sum_{i=1}^{k} w_i = 1$. [5]

2.3 Proposed Method

In case of binary response variable, $\boldsymbol{y}_{n\times 1} = [y_i]$; $y_i \in \{0,1\}$, the model candidate constructed by implementing the logistic regression to averaged in model averaging process. The model candidate in this case can be described below.

$$logit(\hat{\boldsymbol{y}}_i) = f(\boldsymbol{X}^*); i = 1,2,\dots,k$$

where $\boldsymbol{X}^*_{n\times m}$ contains m predictor variables that randomly selected from $\boldsymbol{X}$. Then, the next step is averaging of probability prediction each model candidates ($\hat{p}_i$) using the AIC weight,

$$\hat{p} = \sum_{i=1}^{k} w_i \hat{p}_i$$

before it transforms to be the class of response variable. In this research, AIC weight applied to average the prediction each model candidates that is based on the value of AIC in each model candidates. Suppose there are k model candidates, therefore the i – th AIC weight follows

$$w_i = \frac{exp\left(\frac{1}{2}a_i\right)}{\sum_{i=1}^{k} exp\left(\frac{1}{2}a_i\right)}$$

where a_i denotes the value of AIC in the i – th model candidates, and $w_i \geq 0$; $\sum_{i=1}^{k} w_i = 1$ [6].

In practices, the data set separated to be two parts; training data for constructing the model, and testing data for evaluating the prediction. The observation that selected to be the content of testing data selected randomly with size about 40% of observations that is 100 observations, therefore training data has 187 observations. There are three m used in this research, $m = \{50,100,150\}$ with $k = 50$ that to be evaluated by 100 replications in each processes. In detail the randomly process for selecting observation in testing data and for selecting predictor in model candidate are applied in each replictions.

2.4 Evaluation criteria

The evaluation need to be implemented to check the consistency of prediction in each class of response variable. Because of the response variable has binary scale, the concept of confussion matrix to calculate accuration value, sensitifity value, and specificity value are essential to apply. Suppose a 2 × 2 table with notation based on data set [7].

	Reference	
Predicted	KIRC (1)	LUAD (0)
KIRC (1)	A	B
LUAD (0)	C	D

The formulas used here are:

$$Accuracy = \frac{A + D}{A + B + C + D}$$

$$Sensitifity = \frac{A}{A + C}$$

$$Specificity = \frac{D}{B + D}$$

3. Result and discussion

The data set that used in this research almost has balance proportion of class. Figure 1 describes the percentage of each class of tumor, KIRC 51% and LUAD 49%. Therefore the observations having class KIRC is 146 people, and 141 people for class LUAD.

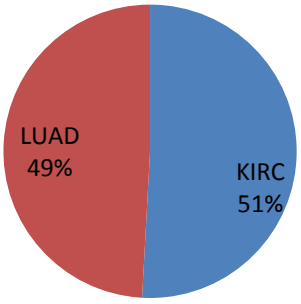


Figure 1 Pie chart of member of class of tumor: KIRC and LUAD

Based on 100 replications, Table 1 shows the mean and standard deviation of each evaluation criteria of prediction based on testing data in each number of predictor variables in model candidate (m). Based on the table, almost all of mean values each evaluation criteria are decreasing when the m size is

bigger. In other hand, the standard deviation for bigger size of m tends to increase. It can be said that smaller size of m in this case can make the prediction very accurate.

Table 1 Mean and standar deviation of evaluation criteria in each m

Evaluation Criteria	Value	m=50	m=100	m=150
Accuracy	mean	0.998	0.998	0.994
	sd	0.007	0.004	0.012
Sensitifity	mean	0.999	0.998	0.992
	sd	0.006	0.008	0.020
Specificity	mean	0.997	0.998	0.996
	sd	0.010	0.005	0.009

In addition, from 100 replications also can be investigated that there are many iteration in the replications can predict the testing data very well in perfect condition that discribed in Figure 2.

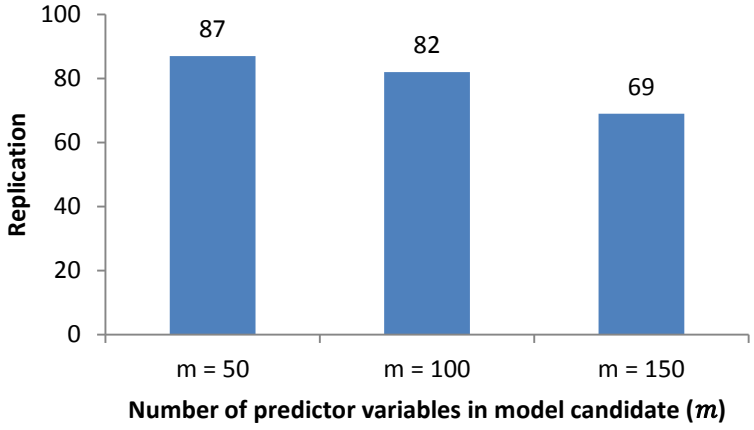


Figure 2 Bar chart of number of replication which has very good prediction based on m

Therefore, when in the model candidate use $m = 50$, 87% of 100 replications has very good performace of prediction. Besides when in the model candidate use $m = 100$, and $m = 150$. Because of that, in this case it can be showed that by using $m = 50$ can give the prediction performance very good in accuracy, sensitifity, and specificity.

4. Conclusion

It can be concluded that the logistic model averaging using randomized approach for cunstructing the model candidate seems to be good alternative in prediction case of high dimensional data of tumor class of patients. Based on the 100 replication of modeling process, the method has good performance when the number of predictor variables in model candidate (m) is 50. It is indicated from the mean, standard deviation, and also number of replication which has very good prediction based on accuracy, sensitivity, and specificity value.

References

1. Perrone MP. 1993. Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization [dissertation]. Providence(US): Brown University.
2. Rahardiantoro S, Sartono B, Kurnia A. 2017. Model Averaging for Predicting the Exposure to Aflatoxin B1 Using DNA Methylation in White Blood Cells of Infants. IOP Conf. Series: Earth and Environmental Science. 58 (2017) 012019.
3. Ghosh D, Yuan Z. 2009. An Improved Model Averaging Scheme for Logistic Regression. J Multivar Anal. 100(8): 1670–1681.
4. The Cancer Genome Atlas Research Network, Weinstein JH, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nature Genetics 45, no. 10 (September 26, 2013): 1113-1120.
5. Ando T, Li KC. 2014. A Model-Averaging Approach for High-Dimensional Regression, Journal of the American Statistical Association. 194: 254-265.
6. Claeskens G, Hjort NL. 2008. Model Selection and Model Averaging. New York (US): Cambridge University Press.
7. Kuhn M. 2008. Building Predictive Models in R Using The caret Package, Journal of Statistical Software.

Clustering of Interval-valued data

Lynne Billard[1], Fei Liu[2]
[1] University of Georgia
[2] Bank of America

Abstract
The concept of symbolic data originates in Diday (1987). We consider cluster methodology for intervals. While there has been a lot of activity in using regression based algorithms to partition a data set into clusters for classical data, no such algorithms have been developed for a set of interval-valued observations. A new algorithm is proposed based on the k-means algorithm of MacQueen (1967) and the dynamical partitioning method of Diday (1973) and Diday and Simon (1976), with the partitioning criteria being based on establishing regression models for each sub-cluster.

Keywords
Partitions, Regressions

1. Introduction

With the advent of the modern computer, there has been an explosion in the size of data sets across all scientific arenas. Analyses of such data sets usually require aggregation in some form driven by the scientific questions underlying these analyses. The aggregation perforce produces symbolic data (such as lists, intervals, histograms, and the like) describing the observations within each aggregated class. Thus, instead of points as for classical observations, observations are now hypercubes or products of Cartesian distributions, in p-dimensional space. Such data were originally introduced by Diday (1987). We consider a dynamic partition of interval data using regression criteria, in Section 2. After briefly describing the basics (in Section 2.1), the k-means and k-regressions algorithms are compared in Section 2.2. The performance of the k-regressions algorithm is then studied on different data set structures, in Section 2.3. We conclude in Section 3.

2. Regression-based Partitions

2.1 Basics

The k-means algorithm was first introduced by MacQucen (1967). Charles (1977) extended the dynamical algorithm of Diday (1973) and Diday and Simon (1976) to build a regression- based algorithm for classical point data.

For interval data, predictor/regression variables take values $X_{ij} = [a_{ij}, b_{ij}]$, with $a_{ij} < b_{ij}$, and the response/dependent variable takes values $Y_i = [c_i, d_i], c_i < d_i, i = 1, \dots, n, j = 1, \dots, p$. The model is

$$Y = X'\beta + \epsilon$$

where $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$, $X' = (1, X_1, \dots, X_p)$ and ϵ is the error interval vector.

The goal is to partition the n observations into K non-overlapping clusters, $= (C_1, \dots, C_K)$, with n_k observations in C_k and $\Sigma_k n_k = n$. Accordingly, we fit the regression line to the n, observations within each C_k, $k = 1, \dots, K$. The least squares estimators of the parameters are $\beta = (X'X)^{-1}X'Y$. The elements of the matrices $X'X$, and likewise for $X'Y$, are functions of the covariance functions between two variables X_i and X_{ij} with realizations $X_{ui} = [a_{ui}, b_{ui}]$ and $X_{uj} = [a_{uj}, b_{uj}]$, $u = 1, \dots, m$, say. This covariance between two interval-valued variables is given by (see Billard, 2008)

$$\begin{aligned} Cov(X_i, X_j) = \frac{1}{6m}\sum_{u=1}^{m}[&2(a_{ui} - \bar{X}_i)(a_{uj} - \bar{X}_j) + (a_{ui} - \bar{X}_i)(b_{uj} - \bar{X}_j) \\ &+ (b_{ui} - \bar{X}_i)(a_{ui} - \bar{X}_i) + 2(b_{ui} - \bar{X}_i)(b_{uj} - \bar{X}_j)], \\ \bar{X}_i = \frac{1}{2m}\sum_{u=1}^{m}&(a_{ui} + b_{ui}) \end{aligned}$$

Thus, we want to find an optimal partition that minimizes the sum of squared residuals (SSR) given K,

$$SSR = argmin_{P;\hat{\beta}_k}\sum_{k=1}^{K}\sum_{i\in C_k} r_{ki}^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k} r_{ki}^2,$$
$$r_{ki} = d(y_i, \hat{y}_i) = d(y_i, x_1'\beta_k)$$

where $d(y_i, \hat{y}_i)$ is a distance between y_i and $\hat{y}_i$. In this work, we use the three distances

1. Center distance: $d_C(X_1, X_2) = \sum_{j=1}^{p} |X_{1j}^c - X_{2j}^c|, X_{1j}^c = (x_{ija} + x_{ijb})/2$;
2. Hausdorff distance: $d_H(X_1, X_2)$: $\sum_{j=1}^{p} \max\{|x_{1ja} - x_{2ja}|, |x_{1jb} - x_{2jb}|\}$
3. City-block distance: $d_{CB}(X_1, X_2) = \sum_{j=1}^{p}[|x_{1ja} - x_{2ja}|, |x_{1jb} - x_{2jb}|]$.

The k-regressions algorithm is:

(i) Initialization: Choose a partition $P^{(0)} = (C_1^{(0)}, \dots, C_K^{(0)})$ randomly from all the possible partitions, or partition the whole data set to K clusters based on some prior knowledge.

(ii) Representation: For $k = 1, \dots, K$, fit regressions $Y_k = X_k'\beta_k + \epsilon$ to the observations in each of the K clusters for partition $P^{(1)} = (C_1^{(1)}, \dots, C_K^{(1)})$ where $l = 0,1, \dots$, denotes the l^{th} iteration.

(iii) Allocation: For observation y_i, $i = 1, \dots, n$, calculate its distance to its prediction $\hat{y}_i$ obtained by its k^{th} regression line $d(y_i, x_1'\hat{\beta}_k), k = 1, \dots, K$, and allocate the observation to its closest line; i.e,

$C_k = \{(x,y)|d(y,\hat{\beta}_k) \leq \forall k \neq k'\}$. The updated partition is now $P^{(l+1)} = (C_1^{(l+1)}, \ldots, C_K^{(l+1)})$.

(iv) Stop: Repeat (ii) and (iii) until the improvement of SSR. is smaller than a predetermined criterion, or the number of iterations reaches a predetermined maximum number.

2.2 Comparison of k-means and k-regressions algorithms

To compare the k-means algorithm with the k-regressions algorithm, consider the data set

(I) which is composed of three clusters that follow the equations:

$$(1): y = 142 + 5x + \epsilon_1$$
$$(2): y = 53 - 3x + \epsilon_2$$
$$(3): y = -43 + 0.6x + \epsilon_3$$

These data are displayed in Figure 1.

Data I with 3 underlying clusters

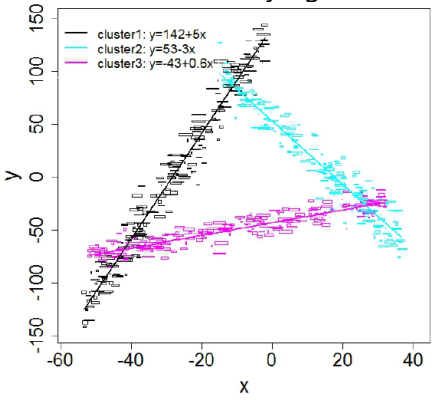


Figure 1

The k-means algorithm produced the partitions of Figures 2(a) and 2(b); Figure 2(a) used the city block distance and Figure 2(b) used the Hausdorff distance. In contrast, Figure 2(c) is the partition obtained by the k-regressions algorithm (after 10 iterations); clearly, this is closer to the original data sect than were those obtained by the k -means method.

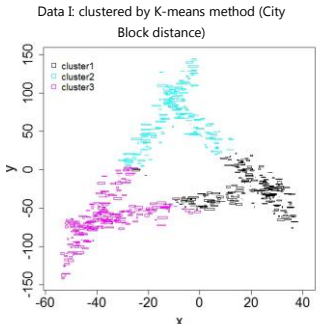


Figure 2(a)

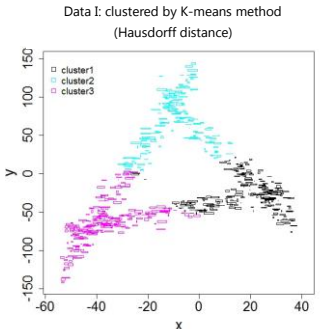


Figure 2(b)

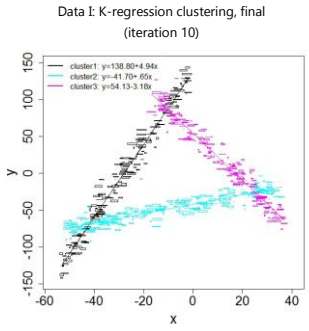


Figure 2(c)

Suppose now we take data set (IT) composed of three clusters that follow the equations:

$$(1)\ : y = 150.5 + 4.5x + \epsilon_1$$
$$(2)\ : y = 53 - 3x + \epsilon_2$$
$$(3)\ : y = -53 + 0.5x + \epsilon_3$$

These are displayed in Figure 3.

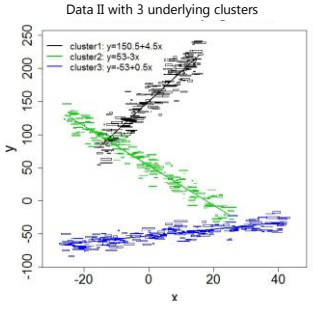


Figure 3

Figure 4(a) and Figure 4(b) show the result when using the k-means algorithm for the city block and Hausdorff distances, respectively. The k -regressions algorithm produced the partitions of Figure 4(c) (after nine iterations), again out-performing the k -means method.

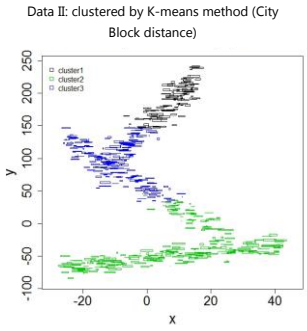


Figure 4(a)

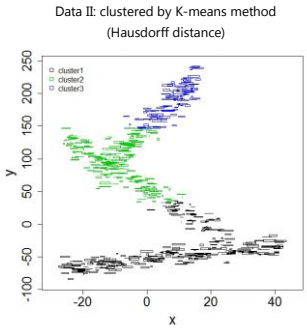


Figure 4(b)

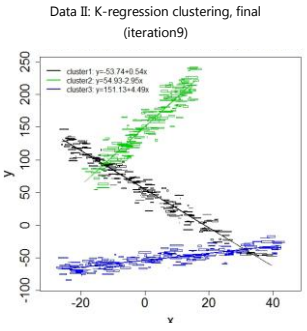


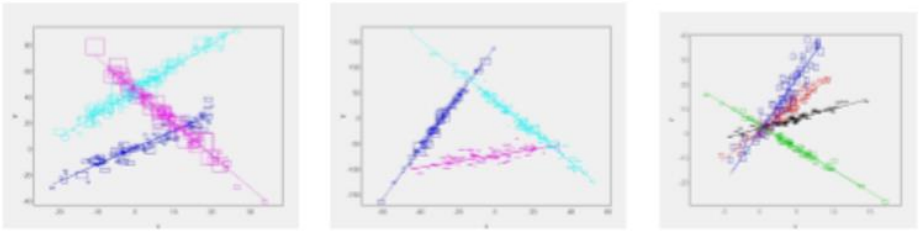
Figure 4(c)

2.3 Different Data Structures

Let us now see how well the k-regressions method performs on the following three data sets with respective structures:

(1) $y = 1.0 + 1.3x$	(1) $y = 142 + 5x$	(1) $y = 2.0 + 0.8x$
(2) $y = 45 + 1.8x$	(2) $y = 33 - 3x$	(2) $y = 1.0 + 2.3x$
(3) $y = 45 - 2.5x$	(3) $y = -73 + 0.6x$	(3) $y = 3.0 - 1.8x$
	(4) $y = 1.0 + 4.3x$	
Data A	Data B	Data C

The plots of these three data sets are as shown in Figure 5.



Data A Data B DataC

Figure 5

The table below shows the mean and standard deviations of the regression parameter estimates based on 100 replications when the *k*-regressions clustering algorithm is applied to the Data C, for each of the center distance, the city-block distance and the Hausdorff distance; also shown are the true parameter values. Clearly, the algorithm works well; likewise, for Data sets *A* and *B*.

		True	center		city-block		Hausdorff	
	β_j	Values	mean	std	mean	std	mean	std
Cluster 1	β_0	2.00	2.06	0.38	3.55	1.76	3.86	2.98
	β_1	0.80	0.81	0.05	0.68	0.17	0.73	0.18
Cluster 2	β_0	1.00	1.32	1.31	3.06	2.74	5.20	4.04
	β_1	2.30	2.36	0.22	2.28	0.50	1.97	0.73
Cluster 3	β_0	3.00	2.90	0.32	3.12	0.34	3.02	0.39
	β_1	-1.80	-1.78	0.04	-1.81	0.05	-1.80	0.05
Cluster 4	β_0	1.00	2.13	1.87	4.29	2.67	4.24	2.82
	β_1	4.30	4.27	0.35	4.04	0.46	4.05	0.48
SSR	-	-	296.25	20.52	777.90	44.84	521.13	27.75

3. Conclusion

We have introduced a new k-regressions algorithm and shown that it works better than does the traditional k -means algorithm on interval valued data. More details can be found in Liu (2016).

References

1. Billard, L. (2008). Sample covariance functions for complex quantitative data. In: Proceedings World Congress, International Association of Statistical Computing (eds. M. Mizuta and J. Nakano) Japanese Society of Computational Statistics, Japan, p. 157-163.
2. Charles, C. (1977). Regression Typologique et Reconnaissance des Formes. These de 3eme cycle, Universite de Paris, Dauphine.
3. Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. International Journal of Computer and Information Sciences 2, 61-88.
4. Diday, E. (1987). Introduction a l'approche symbolique en analyse des donnees. Premier Jouneles Symbolique-Numerique, CEREMADE, Universite Paris - Dauphine, 21-56.
5. Diday E. and Simon, J. C. (1976). Clustering analysis. In: Digital Pattern Recognition (ed. K. S. Fu). Springer, Berlin, 47-94.
6. Liu, F. (2016). Cluster analysis for symbolic interval data using linear regression method. Doctoral Dissertation, University of Georgia.

Fuzzy individual and global assessments, and FANOVA. Application: Fuzzy measure of poverty with Swiss data

Laurent Donzé, Rédina Berkachy
Applied Statistics and Modelling, Department of Informatics, University of Fribourg, Switzerland

Abstract
The measure of poverty is an excellent field to apply fuzzy statistics. Indeed, nowadays, this measure is fast always considered as multidimensional. Furthermore, the evaluation of a poverty level, which is on a large scale subjective and varying between individuals and upon situations, has undoubtfully a fuzzy content. We propose, first, to show in a fuzzy approach how individual and global assessments can be implemented. Second, we develop a fuzzy ANOVA method. We use then these two theoretical tools in order to evaluate the level of poverty, concerning financial conditions, in Switzerland. We test the difference in poverty between two groups of population, Swiss and foreigners.

Keywords
Fuzzy Statistics; Signed Distance; FANOVA; Linguistic Questionnaire; Poverty Measure

1. Introduction

The measure of poverty has challenged the economists and statisticians during decades. It appears that nowadays the measure has to be multidimensional, in the sense that several factors impacting the poverty have to be taking into account. Whatever the measure considered, the problem of evaluating the level of poverty remains. This task is mainly subjective and generally depends on personal considerations. Regarding this latter point, which shows that imprecision and vagueness could be essential, we propose to model the measure of poverty by a fuzzy approach. Indeed, we are not the first to advocate such a modelisation, and for instance we can cite Belhadj (2011), Belhadj and Limam (2012), Chatterjee, Mukherjee, and Kar (2014), Miceli (1998), and Mussard and Pi Alperin (2005).

On another side, the surveys intending to capture the poverty level in a given population, for example, those produced by national statistical offices, are mainly conceived as linguistic questionnaires. The proper evaluation of such questionnaires becomes thus a priority in producing meaningful poverty indices. In some former works, we showed how to implement fuzzy individual and global evaluations of linguistic questionnaires. We also demonstrated

how useful is the signed distance measure in a, e.g. defuzzification process, and in computing these evaluations (see Berkachy and Donzé (2015, 2016a,b)).

In the following, we propose a fuzzy individual and global evaluations of a subset of the Swiss survey SILC (Swiss Federal Statistical Office (2014)). This survey is conducted every year and aims the Swiss income and living conditions. A significant part of the questionnaire is written in linguistic terms, and as such, it is well suited to this kind of approach. Furthermore, we test by a fuzzy anova study for two attributes of poverty the difference between Swiss citizens and foreigners. Bourquin (2016)'s Master thesis was at our knowledge the first one to apply a fuzzy approach to analyse the SILC survey. She measured by different categories of interest, and for several attributes, individual and global fuzzy poverty. Based on the same data, we intend to complete her study. However, our analysis will differ concerning two points. First, as we will show below, our fuzzy measures will be defuzzified by the signed distance measure. Second, we will adopt a different weighting scheme.

Let us shortly in sections 2 and 3 define and present the individual and global assessments, as well as the fuzzy ANOVA (FANOVA). These theoretical results will be applied in section 4, the empirical part of the study. For more explanations, one can fruitfully read, e.g. Berkachy and Donzé (2015, 2016a,b, 2018).

2. Individual and global assessments

Let us assume a linguistic questionnaire divided in main and sub-items, denoted respectively by B_j and $B_{jk}, j = 1, \dots, r, k = 1, \dots, m_j$. We denote respectively by B_j and B_{jk} the associated weights with the constraints: $0 \leq b_j \leq 1, \sum_{j=1}^{r} b_j = 1$, $0 \leq b_{jk} \leq 1$ and $\sum_{k=1}^{m_j} b_{jk} = 1$. A sub-item list of m linguistic terms, $L_q, q = 1, \dots, m$, which we suppose fuzzy. The sampling weight is denoted by $a_i, i = 1, \dots, N$, where N is the size of the sample. Finally, we define the following indicator function:

(1) $\delta_{jkqi} = \begin{cases} 1 & \text{if the observation i has an answer for the linguistic } L_q \\ 0 & \text{otherwise} \end{cases}$

Assuming that there are no missing values – thus latter assumption could be easily relaxed -, the global evaluation P of the linguistic questionnaire is given by:

(2) $$P = \sum_{j=1}^{r} b_j \sum_{k=1}^{m_j} b_{jk} \sum_{q=1}^{m} \frac{\sum_{i=1}^{N} \alpha_i \delta_{jkqi}}{\sum_{i=1}^{N} \alpha_i} d(\tilde{L}_q, \tilde{0}),$$

where $d(L_q, 0)$ is the signed distance of L_q measured from the fuzzy origin 0̃. If a fuzzy linguistic term is characterised by a triangular isosceles membership function, i.e. $L_q = (t_{q-1}, t_q, t_{q+1}), q = 1, \dots, m$, the signed distance $d(L_q, 0)$ is

$$d(\tilde{L}_q, \tilde{0}) = \frac{1}{4}(t_{q-1} + 2t_q + t_{q+1}),$$

and it follows that

$$P_i^{(j)} = \frac{1}{4}\sum_{k=1}^{m_j} b_{jk} \sum_{q=1}^{m} \delta_{jkqi}(t_{q-1} + 2t_q + t_{q+1}). \quad (3)$$

The individual P_I,for an observation I, and the global evaluation P can then easily be computed:

$$P_i = \sum_{j=1}^{r} b_j P_i^{(j)} \quad \text{and} \quad P = \frac{\sum_{i=1}^{N} \alpha_i P_i}{\sum_{i=1}^{N} \alpha_i}$$

3. Fuzzy one-way ANOVA with signed distance

Suppose that we want to perform an Analysis of variance (ANOVA) of a variable X by a factor with K levels. Let $k = 1, \ldots, K$ indicate a specific level with n_k observations. The total number of observations is $n = \sum_{k=1}^{K} n_k$. One unit in a given level k is indexed by j. We denote by X_{kj} the $j-th$ observation of the $k-th$ level. In a fuzzy approach, the output variable X is taken as fuzzy. We denote also by $\tilde{X}_{kj}$ the fuzzy equivalent of X_{kj}. The fuzzy mean $\tilde{\mu}_k$, for a given level k, and the fuzzy sample mean (overall mean) can be estimated respectively by:

$$\bar{\tilde{X}}_{k\bullet} = \frac{1}{n_k}(\tilde{X}_{k1} \oplus \ldots \oplus \tilde{X}_{kn_k}) \quad \text{and} \quad \bar{\tilde{X}}_{\bullet\bullet} = \frac{n_1}{n}\bar{\tilde{X}}_{1\bullet} \oplus \ldots \oplus \frac{n_K}{n}\bar{\tilde{X}}_{K\bullet}. \quad (4)$$

We are now able to express the fuzzy sums of squares related respectively to the treatment $(\widetilde{SSTR}_{\bar{X}})$, the error $(\widetilde{SSE}_{\bar{X}})$ and the total $(\widetilde{SST}_{\bar{X}})$ as follows:

$$\widetilde{SST}_{\tilde{X}} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\tilde{X}_{kj} \ominus \bar{\tilde{X}}_{\bullet\bullet}) \otimes (\tilde{X}_{kj} \ominus \bar{\tilde{X}}_{\bullet\bullet}), \quad (5)$$

$$\widetilde{SSTR}_{\tilde{X}} = \sum_{k=1}^{K} n_k(\bar{\tilde{X}}_{k\bullet} \ominus \bar{\tilde{X}}_{\bullet\bullet}) \otimes (\bar{\tilde{X}}_{k\bullet} \ominus \bar{\tilde{X}}_{\bullet\bullet}), \quad (6)$$

$$\widetilde{SSE}_{\tilde{X}} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\tilde{X}_{kj} \ominus \bar{\tilde{X}}_{k\bullet}) \otimes (\tilde{X}_{kj} \ominus \bar{\tilde{X}}_{k\bullet}). \quad (7)$$

The latter sums of squares cannot be so easily computed. Nevertheless, approximating the fuzzy differences by means of the signed distance can obviously simplify the estimations. Thus, we propose to compute the following crisp analogues of these sums of squares

$$SST_{\tilde{X}} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}\left(d(\tilde{X}_{kj}, \bar{\tilde{X}}_{\bullet\bullet})\right)^2, \quad (8)$$

$$SSTR_{\tilde{X}} = \sum_{k=1}^{K} n_k\left(d(\bar{\tilde{X}}_{k\bullet}, \bar{\tilde{X}}_{\bullet\bullet})\right)^2, \quad (9)$$

$$(10) \qquad SSE_{\tilde{X}} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(d\left(\tilde{X}_{kj}, \bar{\tilde{X}}_{i\bullet}\right) \right)^2 .$$

As the expressions (8), (9) and (10) are now crisp, the well-known decomposition of the sum of squares can be easily verified:

$$SST_{\tilde{X}} = SSTR_{\tilde{X}} + SSE_{\tilde{X}}.$$

and it follows that we can, analogously to the classical ANOVA case, derive a test statistic. Let $F_{\tilde{X}}$ be such a crisp test statistic, where $F_{\tilde{X}} = \frac{MSTR_{\tilde{X}}/(K-1)}{MSE_{\tilde{X}}/(n-K)}$. Under the classical assumption of normality, we have $F_{\tilde{X}} \sim F_{K-1,n-K}$.

4. Empirical analysis

The 2014 SILC data represents more or less 17,000 persons from a sample of Swiss households. We take a subset of the database and keep only the active population, i.e. employed persons which age greater or equal to 18. We focus our analysis on the **financial situation** with a poverty attribute consisting of the two sub-items "good deprivation" and "satisfaction". These are of course only a part of the components of a multidimensional poverty measure. The table 1 describe our variables.

4.1 Individual and global assessments

After modelling each of these sub-items by triangular fuzzy numbers, and combining them by aggregation rules, we defuzzified the obtained aggregated result by the signed distance measure, and proceeded to the computation of the individual and global assessments. The distribution of the (crisp) financial situation (finance) is sketched in figure 1. The global assessment, in this case the mean, is equal to 8.587. It is a relatively high value, which means a weak level of poverty according to finance. This measure is 8.668 for the Swiss citizens and 8.099 for the foreigners, i.e. a sightly lesser (bad) value for the foreigners. As the distribution is crisp, we can without problem perform traditional analysis with it. For instance, a T-test of the difference between the latter two global assesments (Swiss vs. Foreigners) can be done without other restrictions. In the particular case, we observe a significant difference in global assessment of the financial situation between the two sub-populations.

4.2 Fuzzy ANOVA

We propose to execute a fuzzy ANOVA in order to test in a fuzzy context the factor foreign on the two sub-items deprivation and satisfaction. Membership functions for these two latter variables have to be defined. We opted for triangular isosceles fuzzy numbers (Table 2). The results are listed in table 3 and show significant differences with respect to the factor foreign.

5. Conclusion

We have shown how individual and global assessments can be produced. Such results has to be completed by inferences. Thanks to the crisp distribution of the output, we are able to perform simple T-tests. Furthermore, we describe a simple procedure to effectuate fuzzy ANOVA. These two theoretical results have been successfully applied on data concerning the measure of poverty in Switzerland.

APPENDIX

Figures

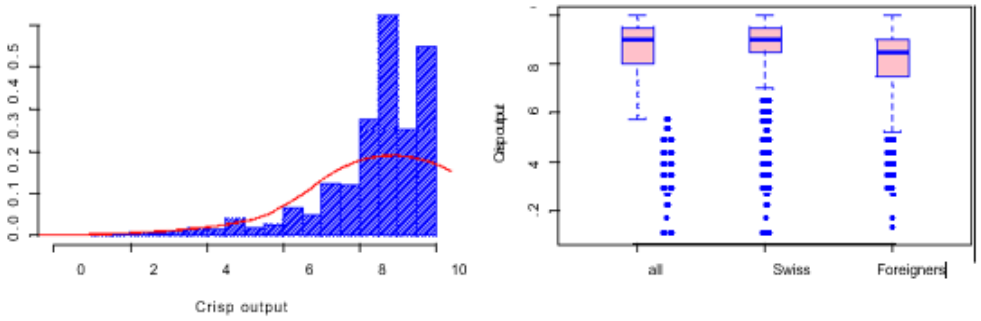


Figure 1: Individual Assessments (Attribute: Financial situation (FINANCE)

Tables

Variables	Description
FINANCE	Crisp financial situation (On a scale of 0 to 10; 0: very bad; 10: very good)
DEPRIVATION	Material deprivation depending on a financial situation (0,1,...,5; 5: all five types of deprivation occured; 0: no deprivation)
SATISFACTION	Satisfaction concerning the financial situation of the houshold (0 to 10 on a Likert scale; 0: unsatisfied; 10: entirely satisfied)
FOREIGN	Nationality (Swiss, Foreigner)

Table 1: Variables

DEPRIVATION		SATISFACTION			
Values	$\mu \bar{X}$	Values	$\mu \bar{X}$	Values	$\mu \bar{X}$
0	(-1,0,1)	0	(-1,0,1)	6	(5,6,7)
1	(0,1,2)	1	(0,1,2)	7	(6,7,8)
2	(1,2,3)	2	(1,2,3)	8	(7,8,9)
3	(2,3,4)	3	(2,3,4)	9	(8,9,10)
4	(3,4,5)	4	(3,4,5)	10	(9,10,11)
		5	(4,5,6)		

Table 2: Triangular isosceles membership functions

DEPRIVATION

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FOREIGN	1.00	72.20388	72.20388	202.45974	0.00
Residuals	13735	4898.35825	0.35663		

SATISFACTION

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FOREIGN	1.00	72.20388	72.20388	202.45974	0.00
Residuals	13735	4898.35825	0.35663		

Table 3: Fuzzy ANOVA for the variables DEPRIVATION and SATISFACTION with respect to the variable FOREIGN

References

1. Belhadj, Besma (Sept. 2011). New Fuzzy Indices of Poverty by Distinguishing Three Levels of Poverty. Research in Economics 65.3, pp. 221–231. url: https://ideas.repec.org/a/eee/ reecon/v65y2011i3p221-231.html.
2. Belhadj, Besma and Limam, Mohamed (2012). Unidimensional and Multidimensional Fuzzy Poverty Measures: New Approach. Economic Modelling 29.4, pp. 995–1002. issn: 0264-9993. doi: http : / / dx . doi . org / 10 . 1016 / j . econmod . 2012 . 03 . 009. url: http : / / www . sciencedirect.com/science/article/pii/S026499931200065X.
3. Berkachy, Rédina and Donzé, Laurent (Nov. 7, 2015). "Linguistic questionnaire evaluation: global and individual assessment with the signed distance defuzzification method". In: Advances in Computational Intelligence, Proceedings of the 16th International Conference on Fuzzy Systems FS'15, Rome, Italy. The 16th International Conference on Fuzzy Systems FS'15 (Nov. 7, 2015). Vol. 34. Rome, Italy, pp. 13–20.
    - (2016a). "Individual and Global Assessments with Signed Distance Defuzzification, and Characteristics of the Output Distributions Based on an Empirical Analysis". In: Proceedings of the 8th International Joint Conference on Computational Intelligence (IJCCI 2016) – Volume 2: FCTA. IJCCI 2016, pp. 75–82. isbn: 978-989-758-201-1. doi: 10.5220/0006036500750082.
    - (2016b). Linguistic Questionnaire Evaluation: an Application of the Signed Distance Defuzzification Method on Different Fuzzy Numbers. The Impact on the Skewness of the Output Distributions. International Journal of Fuzzy Systems and Advanced Applications 3, pp. 12– issn: 2313-0512.
    - (Nov. 8, 2018). "Fuzzy one-way ANOVA using the signed distance method to approximate the fuzzy product". In: Rencontres Francophones sur la Logique Floue et ses Applications 2018. Ed. by

Collectif LFA. LFA 2018. LFA. Arras, France: CÉPADUÈS-ÉDITIONS, pp. 253–264. isbn: 978-2-36493-677-5.

4. Bourquin, Joëlle (Oct. 21, 2016). "Mesures floues de la pauvreté. Une application au phenomena des working poor en Suisse. Travail de Master sous la direction du Prof. Dr. Laurent Donzé". MA thesis. Fribourg: University of Fribourg. 86 pp.
5. Chatterjee, Amitava, Mukherjee, Supratim, and Kar, Samarjit (2014). Poverty Level of Households:
6. A Multidimensional Approach Based on Fuzzy Mathematics. Fuzzy Information and Engineering 6.4, pp. 463–487. issn: 1616-8658. doi: http://dx.doi.org/10.1016/j. fiae . 2015 . 01 . 005. url: http : / / www . sciencedirect . com / science / article / pii / S1616865815000060.
7. Miceli, David (Dec. 1998). Measuring poverty using fuzzy sets. English. Research rep. NATSEM National Center for Social and Economic Modelling, University of Canberra., p. 41.
8. Mussard, Stéphane and Pi Alperin, María (2005). Multidimensional Decomposition of Poverty: A Fuzzy Set Approach. Cahiers de recherche. Departement d'Economique de la Faculte d'administration à l'Universite de Sherbrooke. url: http://EconPapers.repec.org/RePEc: shr:wpaper:05-08.
9. Swiss Federal Statistical Office (2014). Income and Living Conditions (SILC). Survey 2014. Neuchâtel, Switzerland.

Generalized active learning and design of statistical experiments for manifold-valued data

Langovoy, Mikhail
KIT, Kriegsstr. 77, 76133 Karlsruhe, Germany

Abstract

Characterizing the appearance of real-world surfaces is a fundamental problem in multidimensional reflectometry, computer vision and computer graphics. For many applications, appearance is sufficiently well characterized by the bidirectional reflectance distribution function (BRDF). We treat BRDF measurements as samples of points from high-dimensional non-linear non-convex manifolds. BRDF manifolds form an infinite-dimensional space, but typically the available measurements are very scarce for complicated problems such as BRDF estimation. Therefore, an efficient learning strategy is crucial when performing the measurements.

In this paper, we build the foundation of a mathematical framework that allows to develop and apply new techniques within statistical design of experiments and generalized proactive learning, in order to establish more efficient sampling and measurement strategies for BRDF data manifolds.

Keywords

Manifold-valued data; BRDF; proactive learning; sampling strategy.

1. Introduction

In computer graphics and computer vision, usually either physically inspired analytic reactance models, like Cook and Torrance (1981) or He et al. (1991), or parametric reflectance models chosen via qualitative criteria, like Phong (1975), or Lafortune et al. (1997), are used to model BRDFs. These BRDF models are only crude approximations of the reflectance of real materials. In multidimensional reflectometry, an alternative approach is usually taken. One directly measures values of the BRDF for different combinations of the incoming and outgoing angles and then fits the measured data to a selected analytic model using optimization techniques.

There were numerous efforts to use modern machine learning techniques to construct data-driven BRDF models. Brady et al. (2014) proposed a method to generate new analytical BRDFs using a heuristic distance-based search procedure called Genetic Programming. In Brochu et al. (2008), an active learning algorithm using discrete perceptional data was developed and applied to learning parameters of BRDF models such as the Ashikhmin - Shirley model Ashikhmin and Shirley (2000), while Langovoy et al. (2016)

treated active learning for the Cook - Torrance model Cook and Torrance (1981). Analysis of BRDF data with statistical and machine learning methods was discussed in Langovoy (2015b), Langovoy (2015a), Sole et al. (2018), Doctor and Byers (2018).

2. Active learning and design of experiments

In general, BRDF is a 5-dimensional manifold, having 4 angular and 1 wavelength dimension. Note that even a set of 1-dimensional manifolds is infinite-dimensional (and k-dimensional manifolds are not to be confused with parametric k −dimensional families of functions). At the same time, a typical measuring device only takes between 50 and 1000 points for all the BRDF layers together. In view of this, the available measurement points are indeed very scarce for a complicated problem such as BRDF estimation. Therefore, an efficient sampling strategy is required when performing the measurements. Since sets of BRDF measurements are, in fact, observed random manifolds, we are dealing here with manifold-valued data.

Statistical design of experiments (see Fisher et al. (1960), Cox and Reid (2000)) is a well developed area of quantitative data analysis. However, previous research in this field was often more concerned with (important) topics such as manipulation checks, interactions between factors, delayed effects, repeatability, among many others. This shifted the focus away from considering design of statistical experiments on structured, constrained, or infinite-dimensional data. In contrast, BRDF measurements are carried out in strictly defined settings and by qualified experts. Therefore, there is less room for human or random errors and influences. On the other hand, BRDF measurements are collections of points representing manifolds, so defining even the simplest statistical quantities in this case turns out to be a nontrivial and conceptual task.

Overall, our methodology represents a far-reaching generalization of the active machine learning framework, also generalizing the proactive learning setup of Donmez and Carbonell (2008). Active learning, as a special case of semi-supervised machine learning, oftentimes deals with finite sets of labels and aims at solving classification or clustering problems with a finite number of classes. While there have been a number of promising practical applications, most of the existing theory deals with analysis of performance of specific algorithms (query by committee, A^2 algorithm, or importance weighted approach, among a few others) under rather restrictive conditions on the loss functions, incoming distributions, and other components of the learning model. For recent developments, we refer to Agarwal et al. (2013), Beygelzimer et al. (2009), Dasgupta and Hsu (2008).

3. Main definition

In the most basic case, the bidirectional reflectance distribution function (BRDF), $f_r(\omega_i, \omega_r))$ is a four-dimensional function that defines how light is reflected at an opaque surface. The function takes a negative incoming light direction, ω_i, and outgoing direction, ω_r, both defined with respect to the surface normal $\mathbf{n}$, and returns the ratio of reflected radiance exiting along ω_r to the irradiance incident on the surface from direction ω_i. The BRDF was first defined by Nicodemus in Nicodemus (1965). The defining equation is:

$$(1)\ f_r(\omega_i, \omega_r) = \frac{dL_r(\omega_r)}{dE_i(\omega_i)} = \frac{dL_r(\omega_r)}{L_i(\omega_i)\cos\theta_i d\omega_i}$$

where L is radiance, or power per unit solid-angle-in-the-direction-of-a-ray per unit projected-areaperpendicular-to-the-ray, E is irradiance, or power per unit surface area, and θ_i is the angle between ω_i and the surface normal, $\mathbf{n}$. The index i indicates incident light, whereas the index r indicates reflected light.

Suppose we have measurements of a BRDF available for the *set of incoming angles*

$$(2)\Omega_{inc} = \left\{\omega_i^{(p)}\right\}_{p=1}^{P_{inc}} = \left\{\left(\theta_i^{(p)}, \varphi_i^{(p)}\right)\right\}_{p=1}^{P_{inc}}.$$

Here $P_{inc} \geq 1$ is the total number of incoming angles where the measurements were taken. Say that for an incoming angle $\left\{\omega_i^{(p)}\right\}$ we have measurements available for angles from the *set of reflection angles*

$$(3)\ \Omega_{refl} = \bigcup_{p=1}^{P_{inc}} \Omega_{refl}(p),$$

where

$$\Omega_{refl}(p) = \left\{\omega_r^{(q)}\right\}_{q=1}^{P_{refl}(p)} = \left\{\left(\theta_r^{(q)}, \varphi_r^{(q)}\right)\right\}_{q=1}^{P_{refl}(p)},$$

where are $\left\{P_{refl}(p)\right\}_{p=1}^{P_{inc}}$ (possibly different) numbers of measurements taken for corresponding incoming angles. Our aim is to infer the BRDF manifold (1) from the above observations.

In general, the connection between the true BRDF and its measurements is described via a stochastic transformation T, i.e., $f_r(\omega_i, \omega_r) = T\big(f_r(\omega_i, \omega_r)\big)$, where $T: \mathcal{M} \times \mathcal{P} \times \mathcal{F}_4 \to \mathcal{F}_4$, with $\mathcal{M} = (M, \mathfrak{A}, \mu)$ is an (unknown) measurable space, $\mathcal{P} = (\Pi, \mathfrak{B}, \mathbb{P})$ is an unknown probability space, $\mathcal{F}_4$ is the space of all Helmholtz-invariant energy preserving 4-dimensional BRDFs, and

$\mathcal{F}_4$ is the set of all functions of 4 arguments on the 3-dimensional unit sphere S^3 in $\mathbb{R}^4$.

In order to evaluate the influence of measurement errors and to be able to measure the quality of fit of BRDF models, one needs a "measure of distance" between BRDFs. There are many choices of distances and quasi-distances available: $L_p, 1 \leq p < +\infty, L_\infty$, Sobolev distances, Kullback-Leibler information divergence Kullback and Leibler (1951), Mahalanobis (1936), chi-squared distance used in correspondence analysis Langovaya et al. (2013). In computer science literature on BRDFs, there are few papers that study the quality of fit of BRDF models to real data. Most of these studies use the (most standard) L_2 −norm. An alternative approach was taken in Langovoy et al. (2014), where a perception-inspired quasi-metric for the space of BRDFs was proposed.

4. Active manifold learning strategies

In BRDF sampling, the equispaced-angular grid pictured in Figure 1(a) is the standard. However, as was shown in Langovoy et al. (2016), this choice of measurement points leads to very inefficient sampling. Another strategy is in using uniformly distributed points on a sphere, see Figure 1(c). Since it was already understood in the community (see Höpe and Hauer (2010)) that the standard grid is suboptimal, there were multiple heuristic attempts to propose trickier grids that better reflect the typical structure of BRDF models. A good example is shown in Figure 1(b). Ideally, the main goal of this research is to find the best sampling strategy; this strategy has to retain its optimality at least for a class of reasonable criteria, and for a sufficiently general classes of both BRDFs as well as of estimating procedures.

On the other hand, any result showing that new strategy is better than the default strategy, at least for one specific loss function, for one specific BRDF, and one specific estimating procedure, is already instrumental in understanding the general picture of learning BRDF manifolds from scarce expensive data. This basic case is straightforwardly formulated in the language of mathematical optimization, so we are able to obtain theoretical guarantees on learning accuracy, at least for some special cases. Let us outline a possible mathematical framework for BRDF sampling, in a basic case to begin with.

Consider BRDF $f \in \mathcal{F}_4$. Suppose that f is measured on the finite set $\Omega_{means}(n)$

$$(4)\ \Omega_{means}(n) = \left\{\left(\theta_i^{(p)}, \varphi_i^{(p)}, \theta_r^{(q)}, \varphi_r^{(q)}\right) \middle| \left(\theta_i^{(p)}, \varphi_i^{(p)}\right) \in \Omega_{inc}, \left(\theta_r^{(q)}, \varphi_r^{(q)}\right) \in \Omega_{refl}(p)\right\},$$

where

$$(5)\ n = |\Omega_{meas}(n)| = \sum_{p=1}^{P_{inc}} \left|\Omega_{refl}(p)\right|.$$

Definition 1. Cost function $Cost$ of a measurement configuration Ω_{meas} is a Lebesgue measurable function $Cost\colon \mathbb{R}^{|\Omega_{meas}|} \to \mathbb{R}_+$.

Let $Dist$ be a function (measurable for a suitably chosen σ −algebra) such that $Dist : \mathcal{F}_4 \times \mathcal{F}_4 \to \mathbb{R}_+$. For our purposes, we typically like $Dist$ to be inducing either a quasi-distance or a pseudo-distance on $\mathcal{F}_4^0$, where $\mathcal{F}_4^0 \subseteq \mathcal{F}_4$ is a sufficiently reach subset. As an example, a perception-based μ_{BRDF} from Langovoy et al. (2014), was often used in our practical experiments. Standard L_p −distances are easier for theoretical comparisons.

Definition 2. Sampling strategy Ω is a sequence $\Omega = \{\Omega_{n_0}\}_{n_0=1}^{\infty}$ where for each n_0 there exists an integer $n \geq n_0$ such that $\Omega_{n_0} = \Omega_{meas}(n)$ for some measurement configuration $\Omega_{meas}(n)$ defined according to (4), and for any integers $n_1 \geq n_2$ it holds that $|\Omega_{n_1}| \leq |\Omega_{n_2}|$.

Consider arbitrary fixed statistical estimator of BRDFs, $\varepsilon_n : \mathbb{R}^n \to \mathcal{F}_4$.

Definition 3. Let $\Omega = \{\Omega_{n_0}\}_{n_0=1}^{\infty}$ be a sampling strategy, and suppose that $C_{max} : \mathbb{N} \to \mathbb{R}_+$ be a known function. We say that the strategy Ω has uniformly admissible costs with the majorant C_{max}, if for all $n \geq 1$ it holds that $Cost(\Omega_n) < C_{max}(n)$. We say that Ω has asymptotically uniformly admissible costs with the majorant C_{max}, if there exist $n_{min} \in \mathbb{N}$ such that for all $n \geq n_{min}$ it holds that $Cost(\Omega_n) < C_{max}(n)$.

Consider two sampling strategies: Ω^1, Ω^2. Suppose that both strategies have uniformly admissible costs. The problem of generalized active learning for BRDF sampling can be stated in the following way: find a sampling strategy $\Omega_{meas} = \{\Omega_{meas}(n)\}_{n=n_{min}}^{\infty}$ such that for all $n \geq n_{min}$

$$(6)\ \Omega_{meas}(n) = \arg\min_{\Omega: Cost(\Omega) < C_{max}} Dist(\varepsilon_n(f;\Omega), f)$$

Definition 4. Suppose $f \in \mathcal{F}_4$ is a particular (possibly unknown) BRDF. Let Ω^1, Ω^2 sampling strategies be sampling strategies with C_{max} − uniformly admissible costs. We say that strategy Ω^1 is asymptotically more efficient for learning f than the strategy Ω^2, and write $\Omega^1 \succcurlyeq_f \Omega^2$, iff

$$(7)\ \limsup_{n\to\infty} \frac{Dist(\varepsilon_n(f;\Omega^1(n)), f)}{Dist(\varepsilon_n(f;\Omega^2(n)), f)} < 1.$$

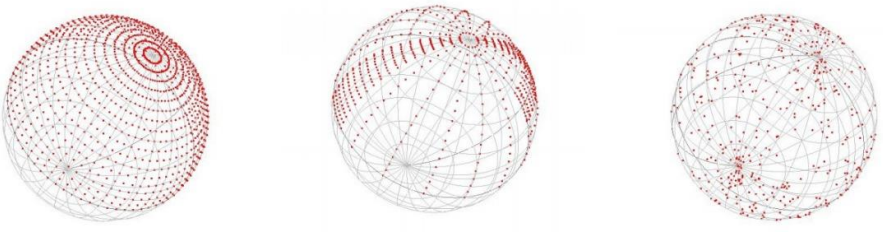
Notice that, for the task of evaluating sampling quality, expected errors (over classes of BRDFs) are more interesting than maximal errors (over the same classes). Indeed, maximal errors are often dominated by degenerate counterexamples, while we are interested in a typical case behavior of our learning procedures. Therefore, we are typically interested in expected errors of the form $\mathbb{E}_{\mathcal{F}'_4}\left(Dist(\varepsilon_n(f;\Omega(n)), f)\right) \longrightarrow \min$, where $\mathcal{F}'_4 \subseteq \mathcal{F}_4$ is a sufficiently reach subset. Clearly, the choice of quasi-metric $Dist$ plays a crucial role.

Definition 5. Suppose $\mathcal{F'}_4 \subseteq \mathcal{F}_4$ is a subset of the set of BRDFs. Let Ω^1, Ω^2 sampling strategies be sampling strategies with C_{max} −uniformly admissible costs. We say that strategy Ω^1 is asymptotically more efficient for learning BRDFs of the class $\mathcal{F'}_4$ than the strategy Ω^2, and write $\Omega^1 < \Omega^2$, if

$$(8)\ \lim_{n\to\infty} \sup \frac{\mathbb{E}_{\mathcal{F'}_4}(Dist(\varepsilon_n(f;\Omega^1(n)),f))}{\mathbb{E}_{\mathcal{F'}_4}(Dist(\varepsilon_n(f;\Omega^2(n)),f))} < 1.$$

Notice that this problem is neither a classification nor a regression task, as we are picking points to estimate manifolds from noisy data.

A special case of this Definition was used in Langovoy et al. (2016) in order to propose more efficient BRDF sampling strategies for industrial applications.



(a) (b) (c)

Figure 1: Sampling strategies for BRDF manifold learning. (a) Standard grid, inefficient sampling. (b): Tricky grid, heuristic choice. (c): Uniformly distributed point on a sphere

5. Conclusions

BRDF manifolds form an infinite-dimensional space, but typically the available measurements are very scarce and expensive. Therefore, an efficient sampling strategy is crucial when performing the measurements. We built a mathematical framework that allows to develop and apply new techniques within statistical design of experiments and generalized proactive learning, in order to establish more efficient sampling and measurement strategies for manifold-valued BRDF data.

Acknowledgements

This work was partially supported by the "DRIMPAC - Unified DR interoperability framework enabling market participation of active energy consumers" project funded by the EU H2020 Programme, grant agreement no. 786559.

References

1. Alekh Agarwal, Leon Bottou, Miroslav Dudik, and John Langford. Para-active learning. arXiv preprint arXiv:1310.8243, 2013.
2. Michael Ashikhmin and Peter Shirley. An anisotropic phong brdf model. Journal of graphics tools, 5 (2):25{32, 2000.
3. Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 49{56. ACM, 2009.
4. Adam Brady, Jason Lawrence, Pieter Peers, and Westley Weimer. genbrdf: Discovering new analytic brdfs with genetic programming. ACM Trans. Graph., 33(4):114:1{114:11, July 2014. ISSN 0730- 0301. doi: 10.1145/2601097.2601193. URL http://doi.acm.org/10.1145/2601097.2601193.
5. Eric Brochu, Nando D Freitas, and Abhijeet Ghosh. Active preference learning with discrete choice data. In Advances in neural information processing systems, pages 409{416, 2008.
6. Robert L Cook and Kenneth E Torrance. A reectance model for computer graphics. In ACM Siggraph Computer Graphics, volume 15, pages 307{316. ACM, 1981.
7. David Roxbee Cox and Nancy Reid. The theory of the design of experiments. CRC Press, 2000.
8. Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In Proceedings of the 25th international conference on Machine learning, pages 208{215. ACM, 2008.
9. Katarina Z Doctor and Je_ M Byers. Optimal sampling of brdf's of varying complexity. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pages 4123{4126. IEEE, 2018.
10. Pinar Donmez and Jaime G Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In Proceedings of the 17th ACM conference on Information and knowledge management, pages 619{628. ACM, 2008.
11. Sir Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Ronald Aylmer Fisher, and Statisticien G_en_eticien. The design of experiments, volume 12. Oliver and Boyd Edinburgh, 1960.
12. Xiao D He, Kenneth E Torrance, Fran_cois X Sillion, and Donald P Greenberg. A comprehensive physical model for light reection. In ACM SIGGRAPH Computer Graphics, volume 25, pages 175{186. ACM, 1991.
13. Andreas Hope and Kai-Olaf Hauer. Three-dimensional appearance characterization of diffuse standard reection materials. Metrologia, 47(3):295, 2010. URL http://stacks.iop.org/0026-1394/47/i=3/a=021.
14. Solomon Kullback and Richard A Leibler. On information and su_ciency. The Annals of Mathematical Statistics, pages 79{86, 1951.

15. Eric PF Lafortune, Sing-Choong Foo, Kenneth E Torrance, and Donald P Greenberg. Non-linear approximation of reectance functions. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pages 117{126. ACM Press/Addison-Wesley Publishing Co., 1997.
16. Anna Langovaya, Sonja Kuhnt, and Hamdi Chouikha. Correspondence analysis in the case of outliers. In Classi_cation and Data Mining, pages 63{70. Springer, 2013.
17. M. Langovoy. Statistical analysis of brdf data for computer graphics and metrology. In IAENG Transactions on Engineering Technologies. Springer, 2015a.
18. M. Langovoy. Machine Learning and Statistical Analysis for BRDF Data from Computer Graphics and Multidimensional Reectometry. IAENG International Journal of Computer Science, 42(1): 23{30, 2015b.
19. M. Langovoy, G. W• ubbeler, and C. Elster. Novel metric for analysis, interpretation and visualization of BRDF data. Submitted, 2014.
20. M. Langovoy, F. Schmaehling, and G. Wuebbeler. Numerical comparison of sampling strategies for BRDF data manifolds. Measurement, 94:578{584, 2016.
21. Prasanta Chandra Mahalanobis. On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta), 2:49{55, 1936.
22. F. E. Nicodemus. Directional reectance and emissivity of an opaque surface. Applied Optics, 4:767 { 775, jul 1965. doi: 10.1364/AO.4.000767.
23. Bui Tuong Phong. Illumination for computer generated pictures. Commun. ACM, 18(6):311{317, June 1975. ISSN 0001-0782. doi: 10.1145/360825.360839. URL http://doi.acm.org/10.1145/360825.360839.
24. Sergei Sobolev. On a theorem of functional analysis.
25. Aditya Sole, Ivar Farup, Peter Nussbaum, and Shoji Tominaga. Bidirectional reectance measurement and reaction model fitting of complex materials using an image-based measurement setup. Journal of Imaging, 4(11):136, 2018.

Model Selection for Covariates Clustering

Thierry Dumont, Ana Karina Fermin
Université Paris Nanterre

Abstract

In this talk we study the problem of inference in high dimensional setting where each variable is assumed to be Boolean. We will present a method to select an appropriate partition of the variables such that variables that are not grouped together are assumed to be independent.

Keywords

Clustering; high-dimensionality; covariate partition; thresholding; model selection; slope heuristic; binary data; MovieLens dataset.

1. Introduction

In many fields of application the data scientist has to handle a lot of covariates that are possibly depending on each other. Examples include social networks, the Netix problem, metagenomics surveys.

We consider here the estimation problem of the law of p binary covariates. Even in the simple case of binary covariates, $2^p - 1$ parameters are needed to describe their distribution, therefore inference becomes quickly intractable when p increases.

In this paper we focus on this binary case and propose a method to overcome this curse of dimensionality phenomena. We propose to project the data into low dimensional models. The covariates are clustered in K blocs such as covariates belonging to different blocs are assumed to be independent. Let $p_1, \ldots, p_K$ be the number of covariates in each bloc, such a clustering involves a number of parameters equal to $\sum_{i=1}^{K}(2^{Pi} - 1)$.

As an illustration, if each bloc contains one variable only, then K = p, the $p_i'^s$ are all equal to 1 and the number of parameters associated with this clustering equals p. This is the simplest model. On the opposite, if all the covariates belong to the same bloc, $K = 1$ and $p_1 = p$. This corresponds to the most expensive model with a number of parameters equal to $2^P - 1$.

To any such partition in blocks, we can associate a maximum likelihood estimate of the density s* of the observations. We aim at finding a good covariate partition.

We propose a statistical view for this partition selection problem. Since the number of possible partitions exponentially grows with the number of covariates, we use a procedure based on the empirical correlation matrix to

restrict the considered set of partitions. A similar approach has been developped in the Gaussian setting in Devijver and Gallopin (2018).

Once the subset of considered partitions has been built we select the best partition using a classical model selection approach based on a penalized likelihood criteria.

The estimator of s* is then defined as the maximun likelihood estimator dened by the chosen data-driven partition.

We implemented our method and applied it to real data. We used the MovieLens dataset made of ratings of 137000 users and consider the top 1000 most rated movies.

This paper is organized as follows. After introducing the notations used throughout the paper, we present our three step method : data-driven pre-selection of the set of partitions of interest, partition selection using a penalized likelihood approach and calibration of the penalty using the classical slope heuristic. The performance of the method is studied using synthetic data in Section 4.1 and using the MovieLens dataset in Section 4.2.

2. Covariates partitions

2.1 Basic notations

Let $p \in \mathbb{N}$ be a the number of covariates. Consider the index set $\{1,\dots,p\}$. In the following, we denote by $y^{(j)}$ the jth component of a vector y and by $y^{(B)} = y^{(j)}; j \in B)$ the group of variables from a cluster $B \subseteq \{1,\dots,p\}$.

Throughout the article, $m = \{B_1, B_2, \dots, B_K\}$ will denote a partition of the covariates into K disjoint clusters $B_1, B_2, \dots, B_K$ with $\cup_{k=1}^{K} B_k = \{1, \dots, p\}$.

Denote by $p_k = |B_k|$ the number of variables in the cluster k.

Denote by M the set of all possible partitions of variables. The set M is large: its size corresponds to the Bell's number which exponentially growth with p.

2.2 Model collection associated with a partition

Let $m \in M$ be a covariates partition. We associate with m a set of probability densities with respect with the uniform measure on $\{0,1\}^p$ defined by

$$S_m = \left\{ s(y) = \prod_{k=1}^{K} s_k\left(y^{(B_k)}\right) \right\}$$

where, for any $k \in \{1, \dots, K\}$, s_k is a probability density on $\{0,1\}^{pk}$.

3. The Method

Suppose that we observe some data $y_1, y_2, \dots, y_n \in \{0,1\}^p$ considered as an n i.i.d. realizations of an unknown probability distribution $s^\star$ on $\{0,1\}^p$. On the

basis of this sample we are interested in estimating the distribution $s^\star$ from the data when the dimension p is large.

For this purpose, on the basis of the observations we consider, for each $m \in M$, an estimator $\hat{s}_m$ of the density $s^\star$ that belongs to S_m. One wish to select among the family of estimators $(\hat{s}_m)\ m \in M$ the one that realizes the "best" compromise between data adjustment and the corresponding model's complexity. However, such a selection requires an exhaustive exploration of all possible partitions which is intractable.

We propose in Section 3.1 a method to considerably restrict the set of all partitions to a set $\widehat{M}$ that naturally arises from thresholding the empirical correlation matrix. Section 3.2 provides a criteria that allows to choose among the candidate densities $(\hat{s}_m)_{m\in\widehat{M}}$.

3.1 Data-driven partition set

Let $C = \big(c(i,j)\big)_{p\times p}$ be the empirical $p \times p$ correlation matrix associated with the data $y = (y_1, y_2, \ldots, y_n)$ with $y_i \in \{0,1\}^p$. For any i and j in $\{1,\ldots,p\}$, $c_{i,j} = cor\big(y^{(i)}, y^{(j)}\big)$.

Let $\lambda \in [0,1)$ be a threshold, de ne the partition m_λ associated with λ: define the adjacency matrix $A_\lambda = (a_\lambda(i,j) = 1_{|c(i,j)|>\lambda})_{(i,j)\in p\times p}$.

let $G = (V, E)$ be the graph associated with A_λ, where the vertexes (nodes) V represent the covariates and E the collection of non oriented edges. For two nodes i and j, $\{i,j\} \in E$ if $a_\lambda(i,j) = 1$ that is if $\big|c(i,j)\big) > \lambda$. The partition m_λ corresponds to the connected components of this graph. Remark that if λ is between two consecutive values of the $|c(i,j)|'s$, $c_1 \leq \lambda < c_2$, then the partition m_λ is the same as $m_{|c1|}$.

Therefore if $\Lambda = \{|c_{i,j}| < 1\} \cup \{0\}$, we consider $\widehat{M} = \{m_\lambda\}_{\lambda\in\Lambda}$ the resulting collection of partitions. It is straightforward to show that $|\widehat{M}| \leq p$.

Model Selection for binary Variable Clustering

3.2 Data-driven partition selection

Let $\widehat{M}$, subset of M, be a random collection of partitions $(y_1,\ldots,y_n)$ −measurable. We wish to select among $\widehat{M}$ a model m that allows theorical bounds on the distance between the true density $s^\star$ and $\hat{s}_m$. We propose a penalized version of the maximum likelihood estimator defined as following.

Next, let $\hat{m}$ be defined as

$$\hat{m} = \underset{m\in\widehat{M}}{argmin}\{-\frac{1}{n}\sum_{t=1}^{n}\log(\hat{s}_m(y_t)) + pen(m)\} \qquad (3.1)$$

where, if $m = \{B_1, B_2, \ldots, B_K\}$,

$pen(m) = \kappa D_m/n$, with $D_m = \sum_{k=1}^{K}(2^{|B_k|})$ being the dimension of the model m and κ is a constant parameter that needs to be calibrated. The penalized density estimator is then defined as $\hat{s}_{\widehat{m}}$. We show that the following oracle type inequality holds under mild assumptions on the model.

Theorem 3.1. *Suppose that there exists* $0 < \epsilon < 2^{-p}$ *such that, for any* $m = (B_1,\ldots,B_K)$ *and any* $s=\prod_{k=1}^{K} s_k \in S_m$, *for all* $k \in \{1, \ldots., K\}$, $s_k \geq \epsilon^{\frac{Pk}{P}}$.

Then there exists $k > 0$ *such that, if* $\widehat{m}$ *is defined as* (3.1),

$$\mathbb{E}[d_H^2(s, \hat{s}_{\widehat{m}})] \leq C\left(\mathbb{E}\left[\inf_{m\in\widehat{M}}\left\{\inf_{s\in S_m} KL(s, s^*) + pen(m)\right\}\right] + \frac{p\log(2p)}{n}\right), \qquad (3.2)$$

for some absolute positive constant C

The proof of Theorem 3.1 relies on the oracle inequality for random collections of models developed in Meynet and Maugis-Rabusseau (2012) and bracketing entropy controls inspired by Bontemps and Toussile (2013).

3.3 Penalty calibration and slope heuristic

Note that our theorem ensures that there exists a κ large enough for which the estimate has good properties, but does not give an explicit value for κ. In practice, κ has to be chosen. We have used the slope heuristic, introduced by Birgé and Massart (2007) and described for instance in Baudry, Maugis and Michel (2012). It provides a practical method to find a good κ.

It is based on the idea that there exists a minimal value κ_{min} such that

- if $\kappa \leq \kappa_{\text{min}}$, the penalized estimator chooses some too complex models,
- if $\kappa > \kappa_{\text{min}}$, the penalized estimator chooses a model for which the estimation error is controlled. In practice, a good choice is to use $\kappa = 2\kappa_{\text{min}}$ so that the penalty used is

$$pen(m) = 2_{k_{min}} D_m/n.$$

It remains to nd this κ_{min}. Two criterion exist. In the first one, called the jump criterion, κ_{min} is estimated as the smallest κ such that the dimension of the model selected is much smaller than the dimension of the most complex m. In the second one, called the slope criterion, we use the fact that if the penalty grows faster with the dimension than the log likelihood then the model chosen will not have a large dimension. As our penalty is proportional to the dimension, it suffices to estimate the slope of the log likelihood in the saturated models with respect to the dimension to obtain a good estimate of κ_{min}.

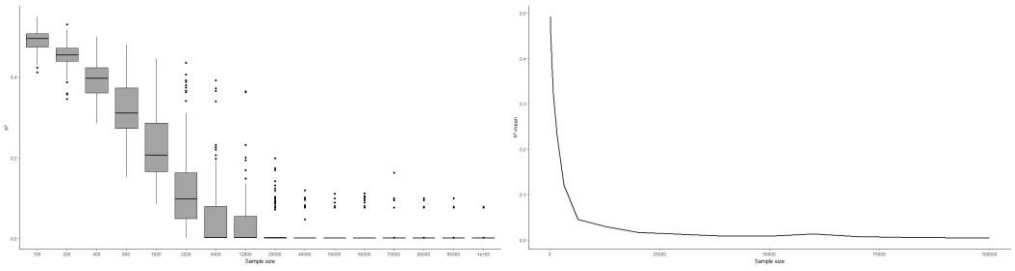


Figure 1: Boxplots (first plot) and empirical means (second plot) of the distances h^2 for each sample size

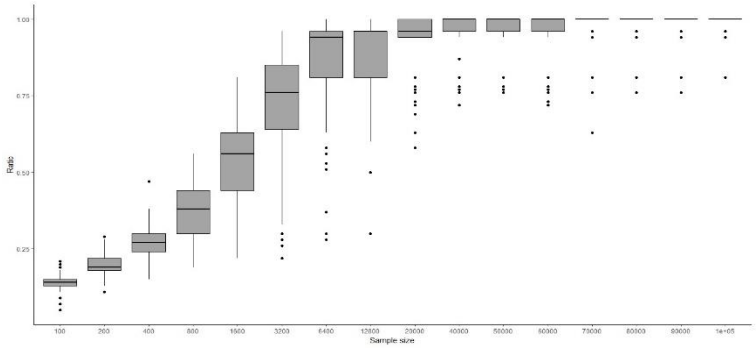


Figure 2: Boxplots of the ratio of variables belonging to the right group for each sample size

4. Numerical results

In Section 3 we presented our approach in nding the optimal variable clusters unsing a n-sample of a multivariate Bernoulli distribution. Our four steps estimation procedure consists in

1. Build the set of partitions of interest $\widehat{M}$ using the thresholding method described in 3.1,
2. For each considered partition $m \in \widehat{M}$, compute the maximum log-likelihood of the associated model ℓ_m and the dimension D_m of its parameter space,
3. Use the slope heuristic method to approach the optimal penalty constant $\hat{\kappa}_{opt}$,
4. select the model $\widehat{m}$ minimizing among $\widehat{M}$ the criterion $-\ell_m + \hat{\kappa}_{opt} D_m / n$

In Section 4.1 we apply our procedure on simulated data. It will allow us to appreciate the performance of the procedure by comparing our estimator with the true model used to generate the sample. In Section 4.2 we illustrate the performance of the procedure on the MovieLens dataset. We will provide an

interpretation of our results when the data is made of movies ratings in order to motivate the practical use of the statistical tool presented in this paper.

4.1 Simulated data set

To study the performance of our procedure on synthetic data we choose a large number of covariates (p = 100). We randomly a ect these covariates into blocs $\{B_1^*, \dots, B_{K^*}^*\}$ whose sizes p_p^*, $k = 1, \dots, K^*$ vary from 1 to 7. Then, for each $k \in 1, \dots, K^\star$, the distribution s_k^* of $Y^{(B_k)}$ is drawn uniformly in the set of probability measures on $\{0,1\}p_k^\star$. The synthetic data sets manipulated in this section are samples of the product measure $s^* = \prod_{k=1}^{K} s_k^*$. Figures 1 and 2 present the performance depending on the sample size n. For each considered size n, 200 independent n-samples are drawn from $s^\star$. Our procedure is applied on each sample providing a partition of the covariates $\{B_1,\dots,B_K\}$ and an estimator $\hat{s}_{\hat{m}} = \prod_{k=1}^{K} \hat{s}_k$ of s^*. In Figure 1 are displayed, for each sample size n the boxplots and the empirical mean of $h^2(\hat{s}_{\hat{m}}, s^*$ over the 200 experiences. Figure 1 illustrates the convergence $\hat{s}_{\hat{m}}$ towards $s^\star$ in terms of the Hellinger distance.

Model Selection for binary Variable Clustering

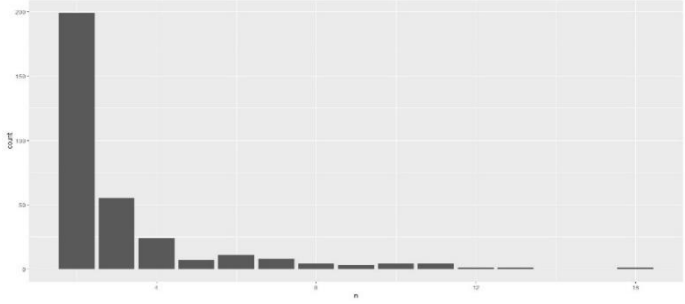


Figure 3: MovieLens groups sizes

title	genres
Unbreakable (2000)	Drama|Sci-Fi
300 (2007)	Action|Fantasy|War|IMAX
Signs (2002)	Horror|Sci-Fi|Thriller
Others, The (2001)	Drama|Horror|Mystery|Thriller
Men in Black II (a.k.a. MIIB) (a.k.a. MIB 2) (200>	Action|Comedy|Sci-Fi
I, Robot (2004)	Action|Adventure|Sci-Fi|Thriller
Ring, The (2002)	Horror|Mystery|Thriller
Terminator 3: Rise of the Machines (2003)	Action|Adventure|Sci-Fi
Mr. & Mrs. Smith (2005)	Action|Adventure|Comedy|Romance
I Am Legend (2007)	Action|Horror|Sci-Fi|Thriller|IMAX
War of the Worlds (2005)	Action|Adventure|Sci-Fi|Thriller
X-Men: The Last Stand (2006)	Action|Sci-Fi|Thriller
Day After Tomorrow, The (2004)	Action|Adventure|Drama|Sci-Fi|Thriller
Ocean's Twelve (2004)	Action|Comedy|Crime|Thriller
King Kong (2005)	Action|Adventure|Drama|Fantasy|Thriller
Transformers (2007)	Action|Sci-Fi|Thriller|IMAX

Figure 4: Titles and genre of the biggest group of movies selected

In Figure 2 we compute the boxplots of the proportion of covariates whose estimated bloc corresponds to one of the true bloc $B_k^\star$. This graphic allows us to appreciate the quality of the model selection side of the procedure.

4.2 Movielens data set

MoviLens dataset Harper and Konstan (2015) contains ratings from 137753 users on 27278 movies(excluding movies with no rating values). Ratings on a 1-5 scale and each user has rated at most 367 movies. We restrict our study to the rst 1000 most often rated movies. For each movie and each user we study the variable equal 1 if the user rated the movie and 0 otherwise. Our selection method applied to the dataset selected a partition made of 322 groups of movies whose size vary from 2 to 16. Figure 3 represents the distribution of the number of variables by group in the partition. Figure 4 represents the movies that belong to the biggest group. We notice that most of them are action/Sci- /Adventure movies released in the early 2000. Similarly most of the groups are made of movies with similar genres and years. Other examples of groups forming the selected partition are provided by Figure 5.

5. Discussion and conclusion

Figures 4, 5 illustrate the quality of the variables clustering provided by the method. We also provide a consistent estimator of the target distribution s^*. This estimator may be used to understand the joint behavior of the variables belonging to the same bloc. Conditioning $\hat{s}_{\hat{m}}$ can also allow prediction on new partially observed dataset.

title	genres
Forrest Gump (1994)	Comedy|Drama|Romance|War
Jurassic Park (1993)	Action|Adventure|Sci-Fi|Thriller
Terminator 2: Judgment Day (1991)	Action|Sci-Fi
Fugitive, The (1993)	Thriller
Speed (1994)	Action|Romance|Thriller
Clear and Present Danger (1994)	Action|Crime|Drama|Thriller
Waterworld (1995)	Action|Adventure|Sci-Fi
Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
Cliffhanger (1993)	Action|Adventure|Thriller
Net, The (1995)	Action|Crime|Thriller
Crimson Tide (1995)	Drama|Thriller|War

title	genres
Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Sci-Fi
Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Sci-Fi
Star Wars: Episode V - The Empire Strikes Back (1>	Action|Adventure|Sci-Fi
Raiders of the Lost Ark (Indiana Jones and the Ra>	Action|Adventure
Indiana Jones and the Last Crusade (1989)	Action|Adventure
Indiana Jones and the Temple of Doom (1984)	Action|Adventure|Fantasy

title	genres
Harry Potter and the Sorcerer's Stone (a.k.a. Har>	Adventure|Children|Fantasy
Harry Potter and the Chamber of Secrets (2002)	Adventure|Fantasy
Harry Potter and the Prisoner of Azkaban (2004)	Adventure|Fantasy|IMAX
Harry Potter and the Goblet of Fire (2005)	Adventure|Fantasy|Thriller|IMAX
Harry Potter and the Order of the Phoenix (2007)	Adventure|Drama|Fantasy|IMAX

title	genres
Little Mermaid, The (1989)	Animation|Children|Comedy|Musical|Romance
Lady and the Tramp (1955)	Animation|Children|Comedy|Romance
Jungle Book, The (1967)	Animation|Children|Comedy|Musical
101 Dalmatians (One Hundred and One Dalmatians) (>	Adventure|Animation|Children
Bambi (1942)	Animation|Children|Drama

title	genres
Aladdin (1992)	Adventure|Animation|Children|Comedy|Musical
Lion King, The (1994)	Adventure|Animation|Children|Drama|Musical|IMAX
Beauty and the Beast (1991)	Animation|Children|Fantasy|Musical|Romance|IMAX

Figure 5: Titles and genres of movies in the selected partition

References

1. Baudry, J.-P., Maugis, C. and Michel, B. (2012). Slope heuristics: overview and implementation. Statistics and Computing 22 455 470.
2. Birgé, L. and Massart, P. (2007). Minimal Penalties for Gaussian Model Selection. Probability Theory and Related Fields 138 33 73.
3. Bontemps, D. and Toussile, W. (2013). Clustering and variable selection for categorical multivariate data. Electron. J. Statist. 7 2344 2371.
4. Devijver, E. and Gallopin, M. (2018). Block-Diagonal Covariance Selection for High-Dimensional Gaussian Graphical Models. Journal of the American Statistical Association 113 306-314.
5. Harper, F. M. and Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Trans. Interact. Intell. Syst. 5 19:1 19:19.
6. Meynet, C. and Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering Research Report.

Female labour force participation rate in Malaysia: Where are we?

Riyanti Saari, Nur Layali Mohd Ali Khan
Department of Statistics Malaysia

Abstract

Recognising the prominence impact of women in socioeconomic growth and nation building, gender equality is included as a core element of the global 2030 agenda for sustainable development. Although the overall well-being of women saw improvement in areas such as health and education, women is still at a disadvantage in the world of work. Globally, the labour force participation rate (LFPR) of female is 50 per cent. At the national level, female's LFPR has increased from 45.0 per cent in 1982 to 54.7 per cent in 2017. A World Bank 2012 study estimated that women's greater participation could provide a growth dividend of up to 0.4 per cent a year. Accordingly, enhancing the role of women in development is one of the priority areas in the Eleventh Malaysia Plan. In line with the global and national aspirations, policies and programmes are reviewed and formulated to further elevate women's involvement in Malaysia's workforce. These initiatives targeted to lift female's LFPR to 56.4 per cent by 2020. Therefore, this study will observe the trend and pattern of women's participation in the labour market from 1982 to 2017 based on the Labour Force Survey data. It will elaborate on the labour force demographic and socioeconomic characteristics of women. In addition, this paper will include the comparison of women's involvement in the labour market in selected countries.

Keywords

Labour Force Survey; Labour supply; Labour market information; Female labour force

1. Introduction

According to the United Nations Department of Economic and Social Affairs [UNDESA] (2019), there were 2.3 million females out of 4.6 billion populations in 1982. The share of female sustained at 49.6 per cent as the world population reached 7.7 billion in 2019. UNDESA (2019) projected that as the world population grows to 8.5 billion in 2030, female will continue to make up almost half of the population. This almost perfect balance entails equity in terms of opportunity and access to education, health and economic resources.

In 2019, Malaysia's population has reached 32.6 million, where for every 100 females, there are 107 males (Department of Statistics, Malaysia [DOSM], 2019a). In 2030, it is projected that 38.1 million population will occupy this country, with the sex ratio of 108 males for every 100 females (DOSM, 2019a). Although males seemed to outnumbered females, females are projected to outlived males. DOSM (2019b) predicted that females born in 2018 would have the life expectancy of 77.6 years while males would live until 72.7 years.

As increase in life expectancy and decrease in fertility rate are coupled with improved educational attainment, female roles have expanded beyond caregiver and nurturer in the family. Ability to generate income, specifically through employment, is one of the most effective ways to achieve economic independence. A study by Luci (2009) found that economic growth promotes women's labour market participation only with active labour market policies to facilitate the entry of women. Thus, she proposed that policies to promote economic sustainability should be combined with policies to increase decent and productive work opportunities for women. From the viewpoint of Kabeer (2012), cultural and traditional roles assignments of breadwinners and household nurtures attributed in most regions of the world to males' higher labour force participation. In relation to this, Duflo (2012) believed that paving ways for female into the labour market would provide a strong catalyst for countries to strengthen the economy. Ghanghas (2018) emphasised that reducing women's time of doing household chores can lead to their economic development and hence empowerment of women. Hence, it is fair to deduce that females' roles in advancing the national social and economic landscape is equally as important as males'. In Malaysia, empowering females to participate in the economic and social development is used as a mean to improve the quality of life. The Mid-Term Review of Eleventh Malaysia Plan has included improving female labour force participation rate (LFPR) as one of the initiatives to empower national human capital as the engine of economic growth. In this sense, it is targeted that female LPFR will attain 56.4 percent by 2020.

LFPR is extensively used to assess the labour market and serves a useful assessment of the labour market along with employment and unemployment rate. ILO (2016) recommended using LFPR as one of the indicators to determine the size and composition of a country's human resources, and to understand the labour market behaviour of different categories of the population. The level and pattern of LFPR depend on employment opportunities and the demand for income, which may differ from one category of persons to another (ILO, 2016).

LFPR is generally higher for male than for female in any given country (ILO, 2019). Australia posted female LFPR of 44.6 per cent in 1982 compared to male LFPR of 77.4 per cent. In 2018, the gap is narrowed down as female recorded LFPR of 60.5 per cent while male LFPR dropped to 71 per cent. Similar trend is

observed for Canada where female LFPR increased from 52.1 per cent (1982) to 61.3 per cent (2018) while male LFPR decreased from 77.1 per cent (1982) to 69.6 per cent (2018). The female LFPR in the Unites States of America went up from 52.6 per cent (1982) to 57.1 per cent (2018) as opposed to male LFPR which went down from 76.6 per cent (1982) to 69.1 per cent (2018). In France, female LFPR rose from 41.8 per cent in 1982 to 51.6 per cent in 2018 while male LFPR within similar period reduced from 68.7 per cent to 60.3 per cent. Within the Asia Pacific region, Japan registered female LFPR of 52.5 per cent in 2018 as against 48.0 per cent in 1982. On the other hand, male LFPR declined to 71.2 per cent compared to 79.5 per cent in 1982. LFPR of female in Singapore improved from 45.2 per cent in 1982 to 59.8 per cent in 2017, while male LFPR within the same period decreased from 81.5 per cent to 76 per cent.

Therefore, this paper aims to assess the male and female LFPR in Malaysia and describe the LFPR of both groups by sociodemographic characteristics. It is hoped that the findings will be able to shed some light on how far Malaysia has grown in terms of encouraging female participation within the labour market. Subsequently, this could spark further concern to facilitate and sustain female in the labour market within the context of decent work and providing work life balance to ensure viable economic and social growth for Malaysia.

2. Methodology

The study utilised data of the Labour Force Survey (LFS) conducted by DOSM. Demographic and socioeconomic characteristics of population in and outside the labour force was profiled using LFS data for selected years between 1982 and 2018. LFS was conducted through household approach to produce national and states estimates of labour force, employment and unemployment. The survey adopted a stratified two stage sampling design. The first stage unit of sample selection is the enumeration blocks (EBs) consisting of 80 to 120 living quarters (LQs), while the second stage unit was the LQs within the EBs. All persons in the selected LQs were canvassed. The detailed methods of LFS are available in the LFS Report (DOSM, 2019c). The sample units were systematically drawn with equal probability of being selected at every stage of selection. The response rates of the annual LFS were more than 85 per cent for all the years involved. Analyses were based on the household members aged 15 to 64 in the LFS data sets. Additionally, the statistics on demographic transition of the population was obtained from the population projections based on the 2010 Population and Housing Census.

Labour force refers to population in the working age group of 15 to 64 who are either employed or unemployed. Malaysia presently maintained the maximum age limit of 64 years in line with the population structure (DOSM, 2019c). LFPR is the ratio of labour force to the working age population

expressed in terms of percentage, used as one of the indicators to determine share of working age population in the labour market. Those not classified as employed or unemployed were identified as outside labour force. This category consists of homemakers, students, retirees, disabled persons and those who are not interested in working.

An employed person was further classified into categories of occupation, sector and employment status consisting of employer, employee, own account worker and unpaid family worker. Employer operates a business, a plantation or other trade and employs one or more workers to help him/her. Meanwhile, own account worker operates his/her own farm, business or trade without employing any paid workers in a continuous basis. Another status of employment was unpaid family worker, a person who works without pay or wages on a farm, business or trade operated by another member of the family.

Education attainment refers to the highest level in which a person has completed schooling or is currently attending, in a public or private educational institution that provide formal education. There are four levels of education attainment which are no formal education, primary, secondary and tertiary.

Survey data were analysed using Statistical Package for Social Science (SPSS version 22) and Microsoft Office Excel. Weighting the survey data was required to infer the sample to represent the survey population. Several descriptive data analysis techniques were utilised to identify the trend of LFPR as well as to profile the demographic characteristics of populations in and outside the labour force.

3. Results

Out of 8.4 million working age population in 1982, there was 64.8 per cent or 5.4 million labour force while 35.2 per cent or 2.9 million outside labour force (Table 3.1). After 36 years, the working age population grew 2.7 per cent per annum to reach 22.4 million, while the labour force grew at a faster rate of 2.9 per cent, registering 14.8 million labour force or LFPR of 68.3 per cent. Of this, male working age population expanded 2.8 per cent annually while labour force registered slower growth of 2.7 per cent.

Table 3.1: Principal labour force statistics, Malaysia, 1982, 2010 & 2018

Sex	Year	Labour force	Employed	Outside labour force	LFPR	Unemployment rate
		('000)	('000)	('000)	(%)	(%)
Male + Female	1982	5,431.4	5,249.0	2,944.6	64.8	3.4
	2010	12,303.9	11,899.5	7,023.0	63.7	3.3
	2018	15,280.3	14,776.0	7,094.4	68.3	3.3
Male	1982	3,562.3	3,465.3	611.5	85.3	2.7
	2010	7,955.5	7,707.8	2,071.7	79.3	3.1
	2018	9,330.2	9,041.8	2,271.3	80.4	3.1

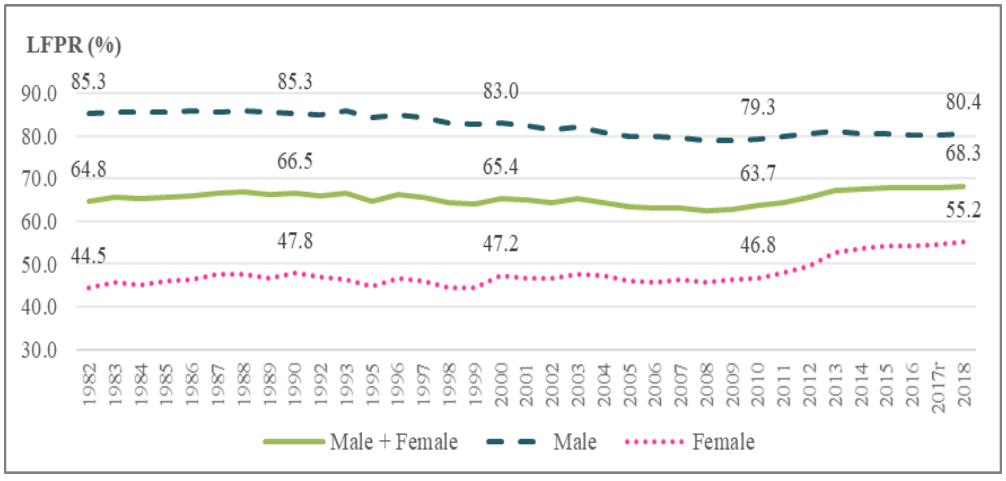
Female	1982	1,869.1	1,783.7	2,333.1	44.5	4.6
	2010	4,348.4	4,191.7	4,951.2	46.8	3.6
	2018	5,950.1	5,734.2	4,823.2	55.2	3.6

Source: LFS, DOSM

Although female registered lower LFPR, they are evidently catching up. As female population within the working age increased 2.6 per cent annually to 10.8 million, female labour force grew at a much faster rate of 3.2 per cent to register 5.7 million. Among the female who join the labour market in 1982, 4.6 per cent were unemployed. Female unemployment rate improved by 1.0 percentage points to register unemployment rate of 3.6 per cent in 2018. Even if male always registered lower unemployment rate, it was observed that male unemployment rate went up from 2.7 per cent in 1982 to 3.1 per cent in 2018.

Chart 3.1 indicates the LFPR of male and female population aged 15 to 64 years in Malaysia from 1982 until 2018. In terms of level, a notable gap was witnessed between male and female LFPR. While male LFPRs were often more than 80 per cent, female LFPR never exceeded 50 per cent prior to 2010. Throughout the period of 1982 to 2010, the LFPR hovered within 63 per cent to 66 per cent. Male LFPR depicted a rather obvious downward trend from 85.3 per cent (1982) to 79.3 per cent (2010) while female LFPR rose marginally from 44.5 per cent in 1982 to 46.8 per cent in 2010. Post 2010, the national LFPR registered an increase of 4.6 percentage points against 2010 to mark an all-time high of 68.3 per cent in 2018. This was mainly attributed by the hike of 8.4 percentage points in female LFPR as opposed to a rise of 1.1 percentage points of male LFPR.

Chart 3.1: LFPR by sex, Malaysia, 1982-2018



Note: LFS was not conducted in 1991 and 1994. The absence of LFS for the two years was due to resources constraint as the organisation prioritised the implementation of Population and Housing Census in 1991 and the Agriculture Census in 1994

Source: LFS, DOSM

The overall and male LFPRs in 1982 and 2018 across age groups (Chart 3.2) portrayed a somewhat reversed "U" shape. High LFPRs were recorded among the 'prime age' of 25 to 54 years old while lower LFPRs were registered by youth aged 15 to 24 and those aged 55 and older. Although LFPR for female seemed to peak at the prime age, the pattern was not the same as national and male LFPRs. In 1982, the LFRP peaked at the age group of 20-24, and later again was high for the age group of 40-44. However, this was not evident when pattern of female LFPR in 2018 is observed. Instead, female LFPR in 2018 was the highest for age group 25-29 years old with a gradual downward pattern thereafter.

LFPR by marital status at two points of time is shown in Table 3.2. As a whole, LFPR for all marital status went up within the period of 1982 and 2018 except for those never married. However, an interesting observation worth noting was the within the period, LFRP of married and widowed male declined 3.9 percentage points and 3.6 percentage points respectively. In the case of female, the LFPR of those who were married surged 18.1 percentage points to 57.3 per cent. In addition, the LFPR of divorced/permanently separated female went up 12.7 percentage points (79.1%) while widowed female posted increase of 5.9 percentage points (48.1%).

In 1982, the highest LFPR was registered for those with tertiary education at 79.1 per cent while those with no formal education posted the lowest LFPR at 56.8 per cent. Although no formal education continued to post the lowest LFPR, the highest LPFR was posted by those with primary education at 70.7 per cent. The highest male LFPR was recorded for those with primary education at 94.7 per cent in 1982 and 89.3 per cent in 2018. The highest female LFPR was recorded for those with tertiary education in 1982 (70.2%) and 2018 (64.3%). Nevertheless, female LFPR with tertiary education dropped 5.9 percentage points in 2018 while LFPR for those with primary education and secondary education registered increases of 5.5 percentage points and 8.6 percentage points respectively.

Chart 3.2: LFPR by sex and age group, Malaysia, 1982 & 2018

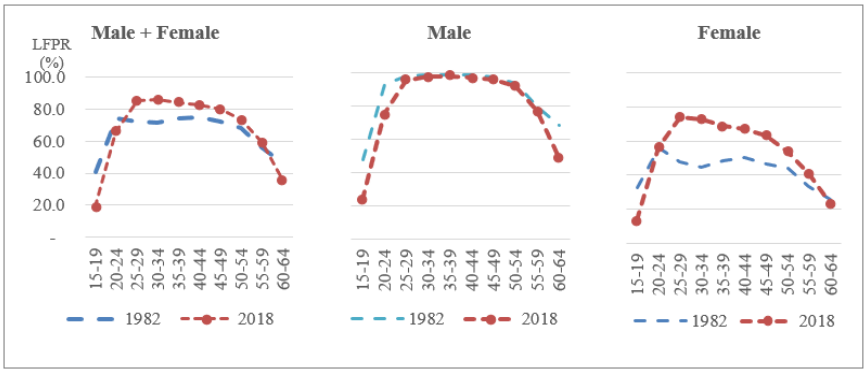


Table 3.2: LFPR by marital status, Malaysia, 1982 & 2018

(%)

Marital status	Male + Female		Male		Female	
	1982	2018	1982	2018	1982	2018
Never married	63.5	60.5	71.0	67.2	53.8	51.2
Married	66.7	74.4	95.9	92.0	39.2	57.3
Widowed	46.7	53.2	75.3	71.7	42.2	48.1
Divorced/permanently separated	72.0	82.3	87.2	88.9	66.4	79.1

Source: LFS, DOSM

Table 3.3: LFPR by educational attainment, Malaysia, 1982 & 2018

(%)

Marital status	Male + Female		Male		Female	
	1982	2018	1982	2018	1982	2018
No formal education	56.8	55.8	89.6	72.1	44.8	41.8
Primary	69.6	70.7	94.7	89.3	42.3	47.8
Secondary	61.8	68.3	75.7	81.8	43.8	52.4
Tertiary	79.1	68.8	84.3	73.8	70.2	64.3

Source: LFS, DOSM

4. Discussion and Conclusion

Based on the assessment of the LFPR, it was observed that prior to 2013, female LFPR never manage to surpass 50 per cent. It started to escalate since 2013 to attain 55.2 per cent in 2018. This could largely be attributed to digital and social media revolution that encouraged new form of work activities especially entrepreneurship which can be conducted within the comfort of home and allow female to balance work and family responsibilities. In the meantime, male LFPR which exceeded 80 per cent most of the time seemed to experience slight decrease.

From the perspective of age group, the LFRP in 1982 showed a double peaked pattern at the age group of 20-24, suggesting first entrance into the labour market and later again was high for the age group of 40-44, signifying reentrance. However, this was not evident when pattern of female LFPR in 2018 is observed. Instead, female LFPR in 2018 was the highest for age group 25-29 years old. This could be attributed by longer time taken to complete education as female embarked on pursuant of higher educational attainment and delayed marriage. A gradual decline of LFPR detected across the older age groups could be one of the indications that career took a back step as motherhood took up the centre stage especially for child bearing and child rearing.

As for marital status, it is noticed that divorced/permanently separated female recorded the highest LFPR since this group required financial

independence more than the other groups. Meanwhile, it is fascinating to see that married female LFPR experienced obvious increase. This could again be attributed by better opportunity to education which later lead to better access into the labour market; as well as various form of work activities made possible through digital disruptions.

Education has always been the determinant for female participation in the labour market, as the findings pointed out that the highest female LFPR was for those with tertiary education. However, improved female LFPR of those with primary and secondary education also suggested that work has changed form and income generating activities do not always require tertiary education. Such instances are prevalence in services sector as dependent contractors work in e-hailing businesses and food delivery services.

To further improve future studies, assessment should be made beyond LFPR to look at the characteristics of female employment, and to also investigate the characteristics of female outside labour force.

Since female make up almost half of the country's population, and with marginal downward trends of male LFPR, it is high time that policies and programmes are designed to facilitate decent and fair opportunity for female to participate in the labour market. This include breaking down unnecessary barriers to include female equal access in decision making positions, allowing flexible and family friendly working hours and environment and encouraging reentrance of female talent into the labour market.

References

1. Dayioğlu, M., & Kirdar, M. G. (2010). Determinants of and trends in labor force participation of women in Turkey. Working Paper No. 5, State Planning Organization of the Republic of Turkey and World Bank Welfare and Social Policy Analytical Work Program, Ankara.
2. Department of Statistics, Malaysia (2019a). Current Population Estimates, Malaysia, 2019. Putrajaya: Department of Statistics, Malaysia.
3. Department of Statistics, Malaysia (2019b). Abridged Life Tables, Malaysia, 2016-2018. Putrajaya: Department of Statistics, Malaysia.
4. Department of Statistics, Malaysia. (2019c). Labour Force Survey, Malaysia, 2018. Putrajaya: Department of Statistics, Malaysia.
5. Department of Statistics, Malaysia (2016). Population Projections (Revised), Malaysia, 2010-2040. Putrajaya: Department of Statistics, Malaysia.
6. Duflo, E. (2012). Women empowerment and economic development. Journal of Economic literature, 50(4), 1051-79.
7. Ejaz, M. (2007). Determinants of female labor force participation in Pakistan an empirical analysis of PSLM (2004-05) micro data. The Lahore Journal of Economics, 12(S), 203-235.

8. Ghanghas, A. (2018). Empowerment of women: Concept, policy approach and implications. International Journal of Law, Volume 4, Issue 1, 36-40.
9. International Labour Organization. (2019, July). ILOStat. Retrieved from International Labour Organization: https://www.ilo.org/ilostat/faces/oracle/webcenter/portalapp/pagehierarchy/Page27.jspx?subject=EAP&indicator=EAP_DWAP_SEX_AGE_RT&datasetCode=A&collectionCode=YI&_afrLoop=2290402432999516&_afrWindowMode=0&_afrWindowId=11e5ea3er2_1#!%40%40%3Findicator%3DEAP
10. International Labour Organization (2016). Key Indicators of the Labour Market, Ninth edition. Geneva: International Labour Offfice.
11. Kabeer, N. (2012). Women's economic empowerment and inclusive growth: labour markets and enterprise development. International Development Research Centre, 44(10), 1-70.
12. Luci, A. (2009). Female labour market participation and economic growth. International Journal of Innovation and Sustainable Development, 4(2/3), 97-108.
13. Ministry of Economic Affairs. (2018). Mid-Term Review of the Eleventh Malaysia Plan 2016-2020: New Priorities and Emphases. Putrajaya: Ministry of Economic Affairs.
14. Tan Sin Yin, Loh Yue Fang, Aminag Ahmand and Nithyaroobini Munia (2010). Categorical data analysis on labour force data in Malaysia. Proceedings of the 6th IMT-GT International Conference on Mathematics, Statistics and its Application, November 3-2, 2010, Grand Season Hotel, Kuala Lumpur.
15. United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, custom data acquired via website.

Measuring the economic impact of tourism industry in Malaysia: An Input-Output Analysis

Rusnani Hussin, Siti Rahmah Seh Omar, Mohd Azam Aidil Abd Aziz
Department of Statistics, Malaysia

Abstract

Tourism industry effects positively to the Malaysia's economy in the forms of output, income and value added. Tourism industry is anticipated to promote economic activity such as Accommodation, Food and beverage, Passenger transport, Travel agencies, Cultural, sports and recreation and Shopping. The paper validates those expectations by examining the economic impacts of tourism industry in Malaysia. Input-output model is adopted as the main methodology in this study. The aim of this paper is to estimate the direct and indirect economic impacts on tourism. Results are consistent with our expectations that tourism has benefited other related tourism industries such as Travel agencies and Accommodation.

Keywords

Tourism Industry; Economic Impact; input-output analysis; multipliers

1. Introduction

Tourism sector contributes towards various economic sectors and thus generate multiplier in the economy. Job opportunities related to tourism could be enhanced as this industry is labour-intensive. The development of this industry is supported by the rapid advancement in the online service facilities. Hence, cost effectiveness in tourism marketing aspects could be optimised.

Tourism industry has become one of the significant industries towards boosting national economy and impact positively to other related industries. Registered travel agencies, accommodation providers, transport operators, food items, souvenirs, handicrafts and independent tour guides have grown rapidly to meet the needs of tourists. Amongst tourism industries, Accommodation, Food and beverage, Passenger transport, Travel agencies, Cultural, sports and recreation and Shopping has spurred the development of the services sector, as well as small and medium enterprises. The impact of this industry development can create jobs, generate domestic income and drive exports as well as generate the country's economy as a whole. According to statistics released in the Tourism Satellite Account report, tourism industry in Malaysia continues to grow in 2017. The share of Gross Value Added Tourism Industry (GVATI) increased to 14.9 per cent in 2017 compared to 14.8 per cent in 2016, amounted to RM201.4 billion. The growth was driven by

retail, food & beverage and lodging industries. Tourism Direct Gross Domestic Product (TDGDP) recorded RM82.6 billion, grew 7.8 per cent from 2016 to 2017. The direct impact of the Tourism industry on GDP reflects the domestic spending and government spending in the tourism industry. Hence, tourism industry has direct impact and indirect impact to the economy

Objective of study is to measuring the economic impact of tourism industry in Malaysia using InputOutput Analysis and to identify the direct and indirect effect tourism industry to economic. This paper is divided into four sections, including this one. Section 2 discusses the methodology that is followed, while section 3 discusses the main findings for the economic impacts of tourism industry. Finally, section 4 provides concluding remarks.

a. Literature review

Hanafiah, et. al (2013) defines that tourism development is a double-edged sword for local communities and attitude directly affects the current and future industry development. The involvement and participation of the community is important towards the success of the tourism development plan. Hanafiah et. al (201) found that Tioman Island community supported future tourism development. It is clear that the role of local residence is vital to support tourism development and maintain continuous growth.

Mazumder, (2013), defines that normal and ratio multipliers were measured to demonstrate the contribution made by tourism industry and its linkages with the other sectors of the economy. International tourist expenditure generate output the most since output multiplier is seven times higher than import multiplier. The value of import multiplier signals the amount of leakage as a result from insignificant tourist expenditure. The findings proves that tourism expenditure results in output generated and higher value added. The multiplier analysis is found to be effective as an appropriate policy making in regulating tourism industry.

However, multiplier analysis of tourism industry in Malaysia showed that this industry is contributing significantly to the economy and proves as a potential sector to enhance economic growth towards a developed nation by 2020.

Fletcher, 1994; Frechtling, 1999; Crompton & Shuster, 2001 and Tyrrell & Johnston, (2001) agreed that until now the leading approach for economic impact estimation of tourism is input-output analysis. They indicated that it has long been recognised that tourism can have an impact on economic activity. Most tourism related industries such as transportation, tax regulation and special events would impact on the patterns of local economic activity and overall economy.

Chou and Huang (2016) defines that, in particular, with the third phase of the “opening up to Chinese tourism” policy implementation, the output value,

GDP and employment increase dramatically in 2009. The GDP created by the inbound tourism is reached to 0.352% of total GDP. The number of inbound Chinese visitors increases in line with the value of macroeconomic variables. It is acknowledged that the input-output model has limitations. In the future, the impact analysis can be conducted using the tourism CGE model, which is currently under updating. Furthermore, if the expenditure function in the evaluation system can be utilized to estimate the tourism expenditure for Chinese visitors, the evaluation of the tourism policy is expected to increase its accuracy.

Hunziker and Krapf (1942) proved that tourism influences towards national economy. They showed that tourism could affect positively and negatively to the national income depending on the inward and outward direction of tourist flows. Consequently, tourism firstly brings about a redistribution of national income, dividing the world into tourist-generating and tourist-receiving countries, regions and destinations; and, secondly, it also leads to a redistribution of income between sectors and companies within the economy, with the latter reflecting the fact that tourism consumption differs from personal consumption.

Camelia (2009) indicated the role and importance of different economic value added, incomes and employment and it analyses the existing connection in an economy. Input-Output Analysis is an economic tool used to measure the impact of an existing, proposed or anticipated business operation, decision or event on the economy. I-O analysis was used to measure the impact of hotels and restaurants for two different years, 2000 and 2005 respectively, the latest available I-O table year for the Romanian economy. The multipliers, estimated on the basis of the IO analysis, are defined as the system of economic transactions that follow a disturbance in an economy. The multipliers can be used to identify the degree of structural interdependence between each sector and the rest of the economy.

2. Methodology:

To determine tourism multipliers through input-output techniques, the 2015 Malaysian national inputoutput transaction tables were used. Sectoral gross output and household income for 2015 were obtained from the 2015 Input-Output transaction table.

Value added data have been taken from the 2015 Malaysian Input-Output transaction table. The 2015 Input-Output tables are 124 x 124 sector. In this study, the 124 sectors in a transaction table are aggregated into 12 sectors and focus on six tourism related sectors sector which are Accommodation, Food and beverage, Passenger transport, Travel agencies, Cultural, sports and recreation and Shopping. Then, this transaction table is constructed into a technical coefficient matrix in order to conduct further analysis. Then, coefficient table is developed to derive multipliers. The aggregating process follows the guideline of Handbook of Input- Output Table Compilation and Analysis, UN, New York, 2012 (UN, 2012).

There are various techniques created to determine the multipliers. The most frequently used approach is the input-output technique. The major benefit of input-output analysis is that it provides detailed information on direct, indirect and induced effects towards the local economy (Loomis and Walsh, 1997). The methodology adopted in this study is based on Leontief input-output techniques where structure of an economy is analysed in terms of inter-relationships between economic sectors (e.g. Miller and Blair, 1985).

The input-output technique of a specific economy signifies the flow of goods and services among its different industries for a particular time period. In the framework of input-output technique, the relationships between economic sectors can be defined in a system of linear equations where total output formed by each sector is either consumed as an intermediate input by other sector, or, internally by the producing sector itself, or, by the final demand sector, or both. To explain, let there be an economy with n-producing sectors and a final demand sector. Assume that the economy can be categorized into n sectors. If we denote by Xi the total output (production) of sector i and by fi the total final demand for sector i's product, in which sector i distributes its product through sales to other sectors and to final demand:

$$Xi = \sum_{j=1}^{n} Xij + fi \quad (1)$$

Let aij be the technical (input) coefficient which represents the amount (value) of sector i's output needed to produce one unit (one Ringgit) of sector j's output; thus, using the assumption of constant production coefficient, we get:

$$aij = x_{ij}/X_j \text{ or } x_{ij} = a_{ij}X_j.$$

Which means that the total value of purchases of goods and services by sector j from sector i is a_{ij} X_j.

Therefore, for a given target of final demand on goods and services, f, this relation defines how much each producing sector must produce in order to satisfy a particular bundle of final demand on goods and services, i.e., Equation (1) in reduced matrix form can be written as:

$$X = AX+F \quad (2)$$

The equation (2) can be found as:

$$X = [I - A]^{-1} F \quad (3)$$

It is mandatory that [I - A] should be a non-singular matrix meaning that the determinant of [I - A] does not equal to zero to have a unique solution in the form of $[I - A]^{-1}$. When the Leontief inverse matrix is assumed to be $[I - A]^{-1} = Z$, then z_{ij}'s stand for the elements of the Leontief inverse matrix. Each element of the $[I - A]^{-1}$ shows the direct and indirect requirements of output of sector i per unit of final demand. Multipliers are those that estimate the effects of exogenous changes on (a) outputs of the sectors in the economy, (b) the value added that is created by each sector in the economy because of the new outputs, and c) income earned by households in each sector because of the new outputs. The concept of multipliers rests upon the difference between the initial effect of an exogenous change and the total effects of that change. The total effects can be defined either as the direct and indirect effects (found from an input–output model that is open with respect to households) or as direct, indirect and induced effects (found from a model that is closed with respect to households).

3. Result

The multipliers that incorporate direct and indirect effects are also known as simple multipliers. When direct, indirect and induced effects are captured, they are often called total multipliers.

$(I - A)^{-1} = I + A + A2 + A3 + \cdots$ will be used, it seems to us preferable to associate "initial" with the I term, "direct" with A, and "indirect" with the remaining terms, $A2 + A3 + \cdots$

Multiplier Analysis
Central to any analysis related to measure the contribution of an activity are economic multipliers, which are derived from the inverse coefficients or total requirements table. In developing multipliers related to tourism sectors in the Malaysian economy, the following procedures are followed. First, Malaysian input-output transaction table is aggregated to 12 sectors. Then followed by the construction of direct requirements matrix and develop the direct, indirect, and induced requirement matrix.

Output Multipliers
An output multiplier for sector j is defined as the total value of production in all sectors of the economy that is necessary in order to satisfy a Ringgit's worth of final demand for sector j's output

Income Multipliers
In this section we explore impacts on households; the approach is exactly the same whether we measure this impact in terms of earnings (monetary). In what follows, we illustrate using income :

$$m(h)j = \sum_{i=j}^{n} an + 1, ilij \qquad (4)$$

Value-Added Multipliers
Another kind of multiplier relates the new value added created in each sector in response to the initial exogenous shock to that initial shock. The principles are identical, and the results in (4) again remain valid. The only new information required is a set of sectoral value-added coefficients:

$$v'_c = v' \hat{x}^{-1}$$

This section discusses the impact from tourism sector towards Malaysian economy. The multipliers computed are output, income and value added multipliers. This study adopted the Input-Output technique to derive the multipliers for Malaysian tourism industry.

This study highlighted the output multiplier resulted from six tourism sector which are Accommodation, Food and beverage, Passenger transport

services, Travel agencies and other reservation services, Cultural, sports and recreation services, Retail trade.

Average tourism multipliers

An average tourism of Type I output multiplier showed that for every Ringgit increase in the tourist expenditure would eventually increase output by 1.81 Ringgit. Consequently, tourism average income multiplier leads to the creation of incomes namely salaries, wages, profits, rents and interest as a result of tourist expenditure. The average Type I income multiplier of tourism indicates that for every Ringgit tourist spending generates 0.31 Ringgit of Malaysian household incomes. Type I value added multiplier was 0.81 Ringgit on average.

Output Multiplier

On average, Type II output multiplier showed that one Ringgit increase in demand for tourism leads to a total of 2.44 Ringgit of output generated in the economy (Table 1). Of 2.44 Ringgit, 0.45 Ringgit created by direct effect 1.36 Ringgit by indirect effect while 0.63 Ringgit by induced effect. Among the related Tourism sectors, Travel agencies yield the highest output multiplier of Type I (1.95) and Type II (2.72). Of 2.72 Ringgit, the contributions of direct, indirect and induced effect are 0.54, 1.42 and 0.77 Ringgit respectively. The second most important sector is Passenger transport with multiplier Type I (1.91) and Type II (2.49). Of 2.49 Ringgit, 0.50 Ringgit created by direct effect, 1.42 Ringgit by indirect and 0.57 Ringgit by induced effect. Food and beverage and Accommodation generated output multiplier above tourism industry average with 2.48 and 2.44 Ringgit respectively.

Income Multiplier

Income multiplier estimates the amount of income generated to Malaysian household residents as a result of unit increase in Ringgit of tourism expenditure. On average, Type II income multiplier showed that one Ringgit increase in demand for tourism leads to a total of 0.41 Ringgit (Table 2). Of 0.41 Ringgit, 0.18 Ringgit created by direct effect 0.13 Ringgit by indirect effect while 0.10 Ringgit by induced effect. Travel Agencies posted the highest Income multiplier Type I (0.38) and Type II (0.50). Of 0.50 Ringgit, direct, indirect and induced were 0.20 Ringgit, 0.17 Ringgit and induced 0.13 Ringgit respectively. Accommodation secured above average income multiplier with Type I (0.37) and Type II (0.49).

Value Added Multiplier

On average, Type II value added multiplier showed that one Ringgit increase in demand for tourism leads to a total of 1.09 Ringgit of value added

generated in the economy (Table 3). Of 1.09 Ringgit, 0.46 Ringgit created by direct effect 0.34 Ringgit by indirect effect while 0.28 Ringgit by induced effect. Among the related Tourism sectors, Accommodation contributed the highest value added multiplier Type I (0.85) and Type II (1.18). Of 1.18 Ringgit, the contributions of direct, indirect and induced effect are 0.55, 0.30 and 0.33 Ringgit respectively. The second most important sector is Travel agencies with multiplier Type I (0.82) and Type II (1.16). Of 1.16 Ringgit, 0.40 Ringgit created by direct effect, 0.43 Ringgit by indirect and 0.34 Ringgit by induced effect.

4. Conclusion:

This paper measured the multipliers of the tourism industry for the Malaysian economy using input output technique. Tourism industry in Malaysia has been identified as a key driver in the services sector. 11th MP focus on capturing high yield tourists to stimulate the industry's contribution to the economy. Domestic tourism will be harnessed to further increase the vibrancy of the industry. Based on Type I multipliers, on average for every Ringgit increase in the tourist expenditure would eventually generate output by 1.81 Ringgit. In addition, for every Ringgit tourist spending generated an average income multiplier of 0.31 Ringgit. This industry has resilient inter-sectoral linkages with other sectors of the economy.

On average, Type II output multiplier showed that one Ringgit increase in demand for tourism leads to a total of 2.44 Ringgit of output generated in the economy. Of 2.44 Ringgit, 0.45 Ringgit created by direct effect 1.36 Ringgit by indirect effect while 0.63 Ringgit by induced effect. Results showed Travel agencies was the highest Type I output and income multipliers with 1.95 and 0.38 Ringgit respectively. For Type I value added multiplier, Accommodation registered the highest with 0.85 Ringgit. Meanwhile Type II output, income and value added multiplier showed that Travel agencies with 2.72, 0.50 and 1.16 Ringgit respectively. Type I value added multiplier was 0.81 Ringgit on average.

Nonetheless, on this study, evidently the tourism industry is contributing significantly to the Malaysian economy in terms of generating output, income and value added. Thus, it also can be used as a tool for policy analysis and economic planning in tourism Industry.

Table 1. Output Multiplier

	Direct	Indirect	Induced	Type I	Rank	Type II	Rank
Tourism Sector	1	2	3	(1+2)		(1+2+3)	
Accommodation	0.390	1.305	0.750	1.694	5	2.444	4
Food and beverage	0.500	1.385	0.598	1.885	3	2.483	3
Passenger transport	0.497	1.418	0.571	1.915	2	2.485	2
Travel agencies	0.536	1.418	0.766	1.954	1	2.720	1
Cultural, sports and recreation	0.471	1.352	0.551	1.823	4	2.373	5
Shopping	0.336	1.268	0.539	1.603	6	2.142	6
Average	0.455	1.357	0.629	1.812		2.441	

Table 2. Income Multiplier

	Direct	Indirect	Induced	Type I	Rank	Type II	Rank
Tourism Sector	1	2	3	(1+2)		(1+2+3)	
Accommodation	0.253	0.116	0.123	0.369	2	0.492	2
Food and beverage	0.173	0.121	0.098	0.294	3	0.392	3
Passenger transport	0.148	0.132	0.094	0.281	4	0.374	4
Travel agencies	0.205	0.172	0.126	0.377	1	0.503	1
Cultural, sports and recreation	0.125	0.146	0.090	0.271	5	0.361	5
Shopping	0.174	0.091	0.088	0.265	6	0.353	6
Average	0.180	0.130	0.103	0.309		0.413	

Table 3. Value Added Multiplier

	Direct	Indirect	Induced	Type I	Rank	Type II	Rank
Tourism Sector	1	2	3	(1+2)		(1+2+3)	
Accommodation	0.551	0.298	0.333	0.849	1	1.182	1
Food and beverage	0.407	0.364	0.265	0.771	5	1.036	5
Passenger transport	0.396	0.358	0.253	0.754	6	1.007	6
Travel agencies	0.397	0.425	0.340	0.822	3	1.162	2
Cultural, sports and recreation	0.440	0.372	0.244	0.812	4	1.056	4
Shopping	0.585	0.247	0.239	0.832	2	1.071	3
Average	0.463	0.344	0.279	0.807		1.086	

References

1. Mohd Hafiz Hanafiah*, Mohd Raziff Jamaluddin, Muhammad Izzat Zulkifly, (2013), Local Community Attitude and Support towards Tourism Development in Tioman Island, Malaysia Faculty of Hotel and Tourism Management, Universiti Teknologi MARA, Procedia - Social and Behavioural Sciences 105 (2013) 792 – 800.
2. Mohammad Nurul Huda Mazumder, (2013), Does Tourism Contribute Significantly to the Malaysian Economy? Multiplier Analysis Using I-O Technique, Faculty of Management, Multimedia University, International Journal of Business and Management Vol. 4 No.7.
3. Larry Dwyera,*, Peter Forsythb, Ray Spurrca Qantas Professor, (2004), Evaluating tourism's economic effects: new and old approaches, Malaysia Faculty of Hotel and Tourism Management, Travel and Tourism Economics, University of New South Wales, NSW 2052, Australia, Tourism Management 25 (2004) 307–317.
4. Chou, Chang-Erh and Huang, Yi-Chen, (2016). "Accurately Estimate Tourism Impacts: Tourism Satellite Account and Input-Output Analysis" Travel and Tourism Research Association: Advancing Tourism Research Globally. 36.
5. Hunziker and Krapf, (1942). The science of systems for tourism development, Annals of Tourism Research, 1988.
6. Camelia SURUGIU, The Economic Impact of Tourism. An Input-Output Analysis, (2009), PhD, Junior Researcher, National Institute for Research and Development in Tourism
7. Ronald E. Millerv and Peter D. Blair(2009), Input-Output analysis Fundamental and extension, Second edition, Cambridge University Press The Edinburgh Building, Cambridge CB2 8RU, UK
8. Department of Statistics Malaysia (2018). Malaysia Input-Output Table 2015: Findings. Department of Statistics Malaysia: Putrajaya.
9. Eurostat (2008). European manuals of supply, use and input-output tables: methodologies and working papers. Office for Official Publications of the European Communities, Luxembourg.
10. Department of Statistics Malaysia (2018). Tourism Satellite Accounts 2017: Findings. Department of Statistics Malaysia: Putrajaya.

Stress transfer modeling Dueto earthquake activity (Case study on Java Island and Bali Nusa Tenggara Islands)

Hasih Pratiwi[1], UpikHandayani[2], Andreas Rony Wijaya[2]
1 Statistics Department, Universitas Sebelas Maret, Surakarta, Indonesia
2 Mathematics Department, Universitas Sebelas Maret, Surakarta, Indonesia

Abstract

Indonesia has a high seismic activity, so we need a disaster management to reduce the risk. Point process is one of the stochastic aspect in disaster management. In point process, a sequence of earthquakes is seen as a collection of points that are random in a space and time, and a conditional intensity function is used to determine the probability of earthquake occurrence. An elastic rebound theory explains a stress release model considering increased stress in an area and stress released during an earthquake over a period of time. On analysis of earthquake data, a region has interactions with neighboring regions of stress transfer that has not been considered in the stress release model, it developed into a linked stress release model. The aimof this research is to estimate parameters so can be of conditional intensity function of linked stress release model on earthquake data in Java Island and Bali-Nusa Tenggara Islands. The conditional intensity function of linked stress release model is the development of the conditional intensity function of stress release model by considering interregional interactions. Maximum likelihood method is used to obtain parameters estimate of the conditional intensity function of linked stress release model. Based on the conditional intensity function of the linked stress release model on earthquake data in Java and Bali-Nusa Tenggara, there is influence of stress transfer caused by earthquake in the regions which is dominated by the proportion of stress from the region itself. Intensity of earthquake in Java is relatively higher than in the Bali-Nusa Tenggara.

Keywords

Point process; earthquake; stress release model; intensity function

1. Introduction

Indonesia is in areas that have high seismic activity and is prone to earthquake disaster. This condition is because Indonesia is located between the meeting of three large plates, that is the Eurasian Plate, the Indo-Australian Plate, and the Pacific Plate. The earthquake disaster can certainly have a big impact. Therefore, a natural disaster management is needed to minimize the impact of the earthquake disaster.

The management of earthquake disasters is still being studied in terms of both seismology and stochastic aspects. One of the stochastic model that describes earthquake's phenomenon is a point process. In this process, earthquakes are seen as a collection of random points in a space, with each point stating the time and / or location of the earthquake (Sunusi et al. [4]). In the point process to find out the intensity of events per unit of time can be known through the conditional intensity function (Ogata [6]). The conditional intensity function is the probability of the occurrence of an event in a very small time interval with the conditions of the events in the previous time.

In the seismology, elastic rebound theory still has a major role. The elastic rebound theory that proposed by Reid is a classic theory for earthquakes. This theory showed that the elastic pressure in the seismic region accumulates due to tectonic plate movement and is released when the pressure exceeds the limit of plate strength. This theory also showed that large earthquakes are usually followed by passive periods, whereas in fact large earthquakes are followed by active periods and sometimes followed by earthquakes that have nearly the same magnitude (Lu et al. [1]).

Vere-Jones [2] proposed a stress release model, which is a stochastic version of the elastic rebound theory. The stress release model considers the pressure that increases in an area and the pressure had been released during an earthquake in a certain period. In 1994, Zheng and Vere-Jones [5] applied a stress release model to earthquake data in China, Iran, and Japan. One of the interesting phenomenon to be observed in his research is that the biggest event is often followed by other major events that are located far enough from the first event. In the earthquake analysis, an area has interaction with other regions. Zheng and Vere-Jones [5] noted several clues to the grouping of treatment transfers of pressure and interaction interregions. Transfer of stress and interaction interregions cannot be considered in the stress release model. The model linked stress release is the development of the stress release model proposed for spatial analysis of the earthquake's occurrence through the transfer of pressure over large areas of the earth's crust.

Jawa, Bali and Nusa Tenggara are part of Indonesia's seismotectonics. This area is traversed by Mediterranean mountains and there is a subduction zone due to a meeting between the Eurasian Plate and the Indo-Australian Plate. The limit of this meeting is in the form of an Oceanic Trench in the south of the cluster of Java, Bali and Nusa Tenggara Islands. This condition causes Java, Bali, and Nusa Tenggara to have high seismic activity.

Through a linked stress release model, it can be used to analyze stress transfers and interactions between regions on the three islands. In the June-August 2018 a series of earthquakes occurred on the islands of Lombok and Sumbawa which were felt to reach Bali and some parts of Java. The biggest earthquake was a magnitude of 7.2 SR which occurred on 5th August 2018

which was centered in East Lombok. Then on 8th August 2018, an earthquake with a magnitude of 5.9 SR was based in Malang, East Java. This research discusses the conditional intensity function for modeling the transfer of pressure due to earthquakes through a linked stress release that is applied to earthquake data in Java Island and Bali-Nusa Tenggara Islands.

2. Methodology

The first step is to determine the form of stress release model and determine the hazard function of the stress release model. Based on the hazard function that has been obtained, then the conditional intensity function of the stress release model is reconstruct. Based on the conditional intensity function of the stress release model then reconstruct the conditional intensity function of the linked stress release model. The conditional intensity function of the linked stress release model obtained applied to earthquake data in Java and Bali-Nusa Tenggara, determine the initial parameter value based on existing earthquake data and then estimate the parameters. From the results of parameter estimation, it is obtained the plot of the conditional intensity function of the linked stress release model. Furthermore, we interpret the results of parameter estimation and conditional intensity function of the linked stress release model.

The elastic rebound theory is a classical theory that can explain the occurrence of earthquakes. Based on this theory, the point process model is constructed into a stress release model. According to Lu et al. [1], in the stress release model, this stress region which controls the probability of the occurrence of earthquakes. The level of $X(t)$ increases deterministically between earthquake and is reduced stochastically as a result of an earthquake. The current value $X(t)$ can be represented in the form

$$X(t) = X(0) = \rho t - S(t) \tag{1}$$

where $X(0)$ the initial value, ρ is constant loading rate from external tectonic forces, and $S(t)$ is the accumulated stress release from earthquake within the region over the period $[0, t)$ that is $S(t) = \sum_{0 \le t_a < t} S(t_a)$, where t_a and $S(t_a)$ denote respectively time of occurrence and stress release associated with the a -earthquake. The stress release value during an earthquake ($S(t_a)$) is estimated from the magnitude.

Kanamori and Anderson [3] showed that magnitude M is proportional to the logarithm of seismic energy released during an earthquake according to the relation $M = \frac{2}{3}\log E + const$. For simplicity, the stress drop during an earthquake is supposed proportional to the square root of the energy released, i.e., $S \propto \frac{1}{E^2}$. Then we have the formula

$$S(t_a) = 10^{0.75(M - M_0)}$$

where M_0 is the normalized magnitude.

3. Result and Discussion

Linked Stress Release Model

Large earthquakes are usually followed by other large earthquakes that are far enough away from the first event and large events can inhibit subsequent events. Inter-regional interactions can affect the time and magnitude of an earthquake event that is explained by the stress transfer. Suppose for region, the stress function of equation (1) is written as

$$X_i(t) = X_i(0) + \rho_i t - S_i(t) \quad (2)$$

Suppose that the accumulated stress release during an earthquake event in region j over a period of time $(0.t)$ is written as $S_j(t)$. The displaced stress from region j to region i has a fixed proportion of positive or negative values symbolized by θ_{ij} with $i = 1,2,3,\dots,m$ and $j = 1,2,3,\dots,m$. Accumulated stress for several regions i after stress transfer from region j with$i = 1,2,3,\dots,m$ is

$$\begin{aligned} S_i(t) &= \theta_{i1}S_1(t) + \theta_{i2}S_2(t) + \cdots + \theta_{im}S_m(t) \\ &= \sum_j \theta_{ij}S_j(t) \end{aligned} \quad (3)$$

Based on equations (2) and (3), the pressure function for some subregions i is written as

$$X_i t = X_i(0) + \rho_i t - \textstyle\sum_j \theta_{ij}S_j(t) \quad (4)$$

Equation (4) is a linked stress release model.

Conditional Intensity Function

The hazard function $\Psi(x)$ states the probability of an earthquake occurring in a time interval $(t, t + \Delta t)$ approaching $\Psi\big(X(t)\big)\Delta t + o(t)$ for Δt which is quite small. It is assumed that the hazard function $\Psi(x)$ is an exponential function written as

$$\Psi(x) = \exp(\alpha + \beta x) \quad (5)$$

with $\alpha \epsilon R$ and $\beta \geq 0$. The α parameter describes the initial pressure value and the β parameter describes the combined strength and heterogeneity of the earth's crust in the area. The probability of an earthquake occurring can be determined using the conditional intensity function of the linked stress release model. The conditional intensity function $\lambda_i(t)$ with the history condition $H_t = \{(t_a, M_a); a = 1,2,\dots,n\}$ is a hazard function of $X_i(t)$ pressure, which is written as

$$\lambda_i(t|H_t) = \Psi(X_i(t)) \quad (6)$$

Substituting equations (4) and (5) into equation (6) is obtained

$$\lambda_i(t|H_t) = \exp\left(\alpha + \beta X_i(0) + \beta\rho_i\left(t - \textstyle\sum_j \frac{\theta_{ij}}{\rho_i} S_j(t)\right)\right). \quad (7)$$

Suppose that $\alpha + \beta X_i(0)) = \alpha_i, \beta\rho_i = b_i$, and $\frac{\theta_{ij}}{\rho_i} = c_{ij}$, thus equation (7) is written as

$$\lambda_i(t|H_t) = \exp\left(a_i + b_i(t - \sum_j c_{ij} S_j(t))\right) \quad (8)$$

The parameter of the conditional intensity function in equation (8) is estimated using the maximum likelihood method with its likelihood function

$$L = [\prod_{a=1} \lambda_i(t|H_t)]\exp\left(-\int_{T_2}^{T_2} \lambda_i(t|H_t)dt\right), \quad (9)$$

and the Log-likelihood function

$$\log L = \sum_{\alpha=1} \log \lambda_i(t|H_t) - \int_{T_2}^{T_2} \lambda_i(t|H_t)dt, \quad (10)$$

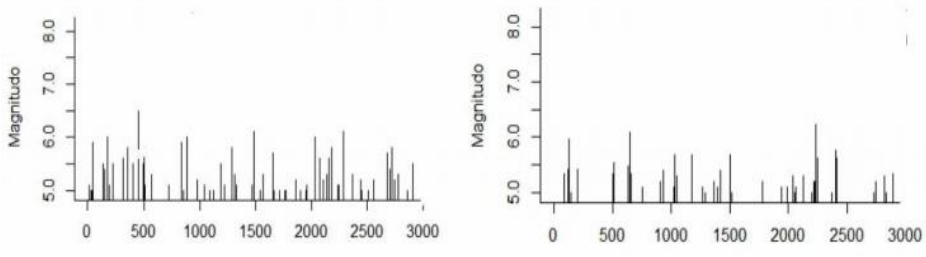
with time internal (T_1, T_2).

This section provides the application of conditional intensity function from the linked stress release model on earthquake data on Java and Bali-Nusa Tenggara. The earthquake data is secondary data sourced from the United States Geological Survey. Earthquake data includes t_a, m_a, component area $i; j$, with t_s states that the time of the a-earthquake and m_a states the magnitude of the a-earthquake. The period of the earthquake data occurred from January 2010 to December 2017 with a magnitude of ≥5 mb and a depth of <70 km. Plot of time and magnitude of earthquake in Java and Bali-Nusa Tenggara is shown in Figure 1.

(a) (b)

Figure 1: Plot magnitude and time for earthquake data in Java (a) and Bali-Nusa Tenggara (b)

Based on Figure 1 (a), there were three earthquakes in Java with the largest magnitude, namely in the south of Java Island on 3 April 2011 with a magnitude of 6.5 mb, in the southeast of Adipala on 25 January 2014 with a magnitude of 6.1 mb, and in the northwest of Bunisari at 6 April 2016 with a magnitude of 6.1 mb. Based on Figure 1 (b), there were three earthquakes in



Bali-Nusa Tenggara with the largest magnitude, namely in the Sumbawa on 8 May 2010 with a magnitude of 6.0 mb, in the south of Bali Island on 13 October 2011 with a magnitude of 6.1 mb, and in the west of Komerda on 12 February 2016 with magnitude 6.3 mb.

Table 1: Estimated parameters of the conditional intensity function of the linked stress release model on earthquake data in Java and Bali-Nusa Tenggara.

Region	Parameter	Parameter estimation
Java	a1	-1.56865330
	b1	0.01200591
	c11	2.00653527
	c12	-0.50834458
Bali-Nusa Tenggara	a2	-1.47601543
	b2	0.01246046
	c21	0.32708324
	c22	0.89687253

Based on equation (8), the parameter vector of the conditional intensity function of the linked stress release model is $\theta = (a_i; b_{i,}; c_{ij})$ '. Then an parameters estimation of the conditional intensity function from the linked stress release model uses the maximum likelihood method and Newton Raphson. The parameter estimation results of the conditional intensity function of the linked stress release model on earthquake data in Java and Bali-Nusa Tenggara are shown in Table 1.

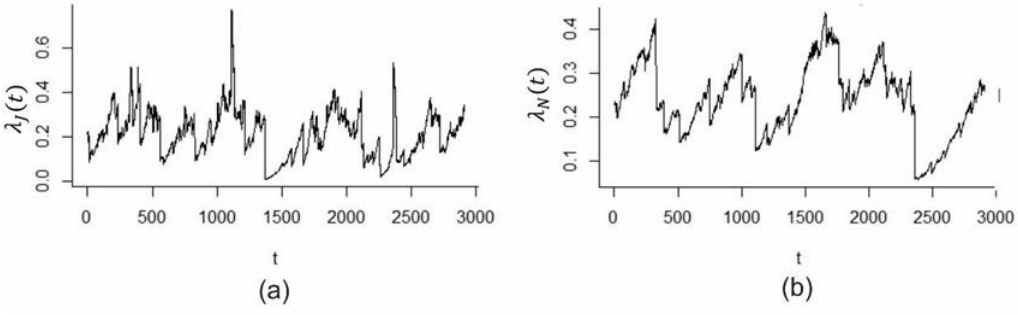
Based on the conditional intensity function of the linked stress release model in equation (4.9) and parameter estimation in Table 1, the conditional intensity function for Java and Bali-Nusa Tenggara respectively is

$\hat{\lambda}_J(t|H_t)$= exp (-1.56865330+0.01200591(t-(2.00653527$S_t(t) - 0.50834458S_2(t)$

and

$\hat{\lambda}_N(t|H_t)$= exp (-1.47601543+0.01246046(t-(0.32708324$S_t(t)$ $0.89687253(t)$

Based on the estimation of linked stress release model parameters on earthquake data in Java, it is seen that c_{11} is greater than c_{12} , it means that the proportion of pressure in the Java region is greater due to pressure in the Java region itself and c_{22} greater than c_{21} means the proportion of pressure in the Bali-Nusa Tenggara region is greater due to pressure in the Bali-Nusa Tenggara region itself. From the results that has been obtained it can be seen



that the intensity of earthquakes for both regions is more dominated by the influence of pressure in the region itself than other regions. Plot of the conditional intensity function of a linked stress release model on earthquake data in Java and BaliNusa Tenggara is shown in Figure 2. It can be seen that the intensity of earthquakes in Java is relatively higher than in Bali-Nusa Tenggara.

Figure 2: Intensity function conditional of linked stress release model on earthquake data in Java (a) and Bali-Nusa Tenggara (b).

4. Conclusion

Conditional intensity function of the linked stress release model is the development of the intensity function of the stress release model by considering inter-regional interactions. Conditional intensity function of the model can be presented by hazard function of stress movement and interaction among areas. Based on earthquake data in Java and Bali-Nusa Tenggara, there is influence of stress transfer to earthquake intensity in the region which is dominated by the proportion of pressure from the region itself. Intensity of the earthquake in Java is relatively higher than the one in Bali-Nusa Tenggara.

References

1. C. Lu, D. Harte, and M. Bebbington, "A Linked Stress Release Model for Historical Japanese Earthquake: Coupling among Major Seismic Regions", Earth Planets Space, no. 51, pp. 907-916, 1999.
2. D. Vere-Jones, "Earthquake Prediction – A Statistician's View", J. Phys Earth. Vol. 26, pp. 129- 146, 1978.
3. H. Kanamori and D. L. Anderson, "Amplitude of the Earth's Free Oscillations and Long-Period Characteristic of the Earthquake Source", Geophys, no. 80, pp. 1075-1078, 1975.
4. N. Sunusi, A. K. Jaya, A. Islamiyati, and Raupone, "Studi Temporal Point Process pada Analisa Prakiraan Peluang Waktu Kemunculan Gempa, Mitigasi dan Manajemen Sumber Daya Alam", Research report, Mathematics Department, FMIPA, Universitas Hasanuddin, Makassar, 2013.
5. X. Zheng and. Vere-Jones, "Applications of Stress Release Models to Historical Earthquake from North China", Pure Appl. Geophys, vol. 4, no. 135, pp. 559-576, 1991.
6. Y. Ogata, "Seismicity Analysis Trough Point Process Modelling: A Review", Pure an Applied Geophysics, no. 155, pp. 471-507, 1999.

Regime-switching state-space models with applications to brain imaging

David Degras[1], Chee Ming Ting[2], Hernando Ombao[3]
[1] University of Massachusetts Boston
[2] Universiti Teknologi Malaysia
[3] King Abdullah University of Science and Technology

Abstract

State-space models (SSMs) with regime switching can efficiently identify recurring patterns of variation and recurring dynamics in nonstationary multivariate time series. These models have been successfully applied in various fields such as econometrics, signal processing, control engineering, and object tracking. In this work we focus on the implementation of switching SSMs in high dimension via the Expectation-Maximization (EM) algorithm. The EM algorithm provides a relatively simple way to compute the maximum likelihood estimator (MLE) of the model parameters. However, in switching SSMs, exact calculations are intractable as they grow exponentially with the time series length. Even approximate calculations are burdensome with high dimensional data. In addition, the EM algorithm has a tendency to get stuck in non-optimal stationary points of the likelihood function, a tendency further compounded in high-dimension. Considering two common switching SSMs, one with switching dynamics and the other with switching observation process, we make several practical contributions: 1) we propose novel robust initialization methods for the EM algorithm, 2) we develop a parametric bootstrap procedure for statistical inference, 3) we provide an efficient implementation of the EM algorithm for all discussed models in a comprehensive MATLAB package publicly available at https://github.com/ddegras/switch-ssm. These contributions make it possible to reliably calculate the MLE in a reasonable time, even with very long and/or high-dimensional time series. We evaluate the statistical performance of the MLE in a simulation study and compare it to a popular alternative approach (sliding windows correlation followed by k-means clustering). We also present applications to the study of dynamic functional connectivity in large electroencephalography (EEG) datasets.

Keywords

Nonstationary time series, Computational statistics, High-dimensional data, EM algorithm

1. Introduction

Regime-switching state-space models (in short, switching SSMs) form a powerful class of time series models used in fields as varied as econometrics [7], speech recognition [12], computer vision [1], and tecently neuroimaging [11]. Switching SSMs can flexibly track nonstationary behavior and identify (possibly low-dimensional) latent factors in time series. These models are particularly suitable in situations where dependencies between study variables are modulated by an underlying regime of activity. In econometrics, for example, such regimes could be "growth cycle" and "recession cycle".

Several computational methods are available for switching SSMs. Bayesian approaches include Gibbs sampling [8], variational Bayes [4], and sequential Monte Carlo [3]. Frequentist approaches are typically based on the maximum likelihood estimator (MLE) and the Expectation-Maximization (EM) algorithm (2, 5, 7, 13]. In practice, switching SSMs have been mostly applied to low-dimensional and telatively short time series. The case of high-dimensional and/or long time series, which is our focus here, poses considerable numerical challenges both for model fitting and statistical inference. Also, to our knowledge, there are currently no publicly available software packages for switching SSMs.

This work aims to facilitate the implementation of switching SSMs with large datasets. We study two broadly applicable switching SSMs and their implementation via the EM algorithm. Our contributions are as follows. First, we provide two new initialization methods for the EM based on least square regression, K-means clustering. and dichotomic search. Indeed, the choice of starting points is often key to the successful convergence of optimization algorithms, especially for large datasets and models with many parameters. Second, we provide numerical optimization tools to handle constraints on the model parameters such as equality constraints, fixed coefficients constraints, or scaling constraints.

Such constraints can prove important both for model interpretability and for numerical stability and convergence of the EM. Third, we develop a parametric bootstrap method for the statistical inference of model parameters. In our experience, likelihood-based inference is not tractable in high-dimensional switching SSMs: the proposed bootstrap offers a viable alternative that can easily be computed in parallel. Fourth, we implement our approach in a suite of MATLAB functions available at https://github.com/ddegras/svitch-ssa. Applications of our switching SSMs to large electro-encephalography (EEG) data from an epilepsy study and a brain computer interface study will be presented orally (but not here for reasons of space).

The paper is organized as follows. Section 2 gives a general account of switching SSMs and introduces our study models. Section 3 briefly describes

the EM algorithm for switching SSMs and outlines our numerical contributions. Section 4 presents a new parametric bootstrap of the MLE for statistical inference.

2. Linear state-space models with regime switching

Linear state-space models with regime switching can be viewed as a combination of linear dynamical systems and hidden Markov models:

$$y_t = C_{S_t} x_t + W_t$$
$$X_t = A_{S_t} x_{t-1} + V_t \quad (1)$$

where at time $1 \le t \le T$, y_t is the measurement vector of size N, x_t is the hidden state vector of size $r (r \le N)$, W_t represents measurement errors, and V_t denotes random innovations to the state process. The first and second equations are called observation equation and state equation, respectively. The switching variable S_t indicates the regime under which the system (1) operates at time t. The sequence $(S_t)_{1 \le t \le T}$ is a homogeneous Markov chain taking values in a finite set of regimes, say $\{1, \dots, M\}$, with initial state probabilities $\pi_j = P(S_1 = j)$ and transition probabilities $Z_{ij} = P(S_t = j | S_{t-1} = i)$ for $1 \le i, j \le M$. Under regime $S_t = j$, A_j is the transition matrix that governs the dynamics of the state vector x_t and C_j is the observation matrix that maps the hidden state vector x_t to the observed measurements y_t. Conditionally on the regimes $(S_t)_{1 \le t \le T}$, the measurement errors w_t are independent over time and have normal distribution $N(0. R_{S_t})$. Similarly, the innovations v_t are independent over time, mutually independent with the w_t and have normal distribution $N(0. Q_{S_t})$ conditionally on $(S_t)_{1 \le t \le T}$. Hence at time t, if $S_t = j$, the parameters at play in model (1) are (A_j, C_j, Q_j, R_j). Note that regular (non-switching) linear SSMs correspond to the case $M = 1$.

Model (1) is very general and must be specialized for practical purposes. A first specification is

$$y_t = C_{x_i} + W_t$$
$$x_t = \sum_{\ell=1}^{P} A_{\ell, S_t} x_{t-\ell} + v_t \quad (2)$$

This model, which we call the *switching dynamics* model as in [10], posits a common observation matrix C and error covariance R for all regimes $1, \dots, M$. In other words the observation equation does not depend on the regime S_t. On the other hand, the dynamics of the state equation switch with S_t. At time t, conditional on S_t, x_t is a vector autoregressive (VAR) process of order p with transition matrices A_{ℓ, S_t}, and innovation covariance Q_{S_t} ($1 \le \ell \le p$ denotes the lag). The dependencies between observations, state vectors, and regimes under this model are depicted in Figure 1 (left panel).

Another possible specification of model (1) is: (3)

$$y_t = C_{S_t} x_{tS_t} + w_t$$

$$x_t = \sum_{\ell=1}^{P} A_{\ell_j} x_{(t-\ell)j} + v_{tj}, 1 \le j \le M$$

Following [10], we refer to model (3) as the *switching observations* model. Here, both the observation matrix and observed state vector depend on the regime S_t. There are in fact M different state vectors $x_{tj}, 1 \le j \le M$, evolving independently according to a VAR(p) model determined by the A_{ℓ_j} and Q_j. At time t, only one of these state vectors is observed through C_{S_t}. Dependencies between observations, state vectors, and regimes under this model are depicted in Figure 1 (right panel). Model (3) can be viewed as a mixture-of-experts neural network wherein the SSMs specified by $y_t = C_j x_{tj} + w_t$ and $x_t = \sum_{\ell=1}^{P} A_{j\ell} x_{(t-\ell)j} + v_{tj}$ $(1 \le j \le M)$ are experts and (S_t) is a gating network [4].

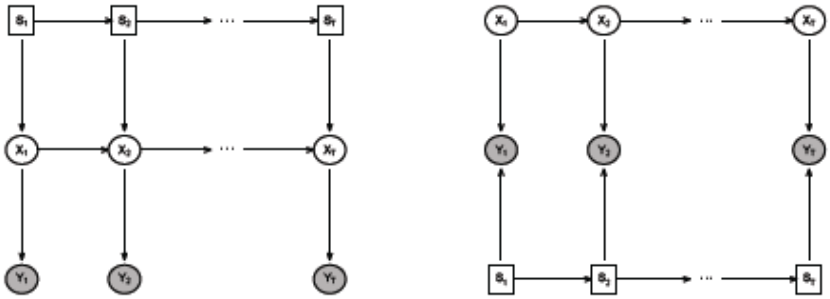


Figure 1: Directed acyclic graph representation of the studied switching space-models. Left: switching dynamics (2). Right: switching observations (3). Square nodes represent discrete variables and oval ones are Gaussian. Shaded nodes are observed while white one are hidden.

3. Model fitting by the EM algorithm

A general presentation of the EM algorithm can be found in [9] and its specific implementation in model (1) is described in [7, 10]. For reasons of space, we omit a full presentation here and focus on our new developments. Let $\theta = \{(\boldsymbol{A}_j, \boldsymbol{C}_j, \boldsymbol{Q}_j, \boldsymbol{R}_j, \mu_j, \sum_j) : 1 \le j \le M; \pi; \boldsymbol{Z}\}$ be the collection of all parameters in model (1), with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)'$ and $\boldsymbol{Z} = (Z_{ij})_{1 \le i,j \le M.}$ For brevity we denote the measurements $(\boldsymbol{y}_t)_{1 \le t \le T}$ by $\boldsymbol{y}_{1:T}$, the state vectors $(\mathbf{x})_{1 \le t \le T}$ by $\mathbf{x}_{1:T}$, etc. We recall that only the measurements $\boldsymbol{y}_{1:T}$ are observed whereas both the state vectors $\mathbf{x}_{1:T}$ and regimes $S_{1:T}$ are unobserved. We denote the complete likelihood function (i.e., if $\boldsymbol{y}_{1:T}$, $\mathbf{x}_{1:T}$, $S_{1:T}$ were all observed) by $L_c(\theta)$. We also denote the probability measure associated to model (1) by P_θ and expectation under P_θ by E_θ.

3.1 E-step

Given a current estimate $\hat{\theta}$ of θ, the E-step consists in taking the conditional expectation of the complete data log-likelihood given the observed data $\boldsymbol{y}_{1:T}$ while assuming that $\hat{\theta}$ is the true model parameter. This yields the Q −function

$$Q(\theta;\hat{\theta}) = E_{\hat{\theta}}(\log L_c(\theta)|\boldsymbol{y}_{1:T})$$

which serves as an approximation to the (incomplete data) log-likelihood function. The calculation of $Q(\theta;\hat{\theta})$ requires the following quantities:

$$\begin{aligned} W^j_{t|\tau} &= P_{\hat{\theta}}(S_t = j|\boldsymbol{y}_{1:\tau}), & \mathbf{P}^j_{t|\tau} &= E_{\hat{\theta}}(x_t x'_t|S_t = j, \boldsymbol{y}_{1:\tau}),\\ W^{ij}_{t-1,t|\tau} &= P_{\hat{\theta}}(S_{t-1} = i, S_t = j|\boldsymbol{y}_{1:\tau}), & \mathbf{P}^{\bigcirc j}_{t-1|\tau} &= E_{\hat{\theta}}(x_{t-1} x'_{t-1}|S_t = j, \boldsymbol{y}_{1:\tau}),\\ x^j_{t|\tau} &= E_{\hat{\theta}}(x_t|S_t = j|\boldsymbol{y}_{1:\tau}), & \mathbf{P}^j_{t,t-1|\tau} &= E_{\hat{\theta}}(x_t x'_{t-1}|S_t = j, \boldsymbol{y}_{1:\tau}), \end{aligned} \tag{4}$$

where $\tau = t-1$ for prediction, $\tau = t$ for filtering, and $\tau = T$ for smoothing. The quantities in (4) can be computed approximately with the Kim filtering algorithm, also known as Hamilton filtering [7, 10]. In essence, the calculation of (4) relies on a forward-backward algorithm similar to the Kalman filter and Rauch-Tung-Striebel smoother for SSMs. Here however, exact calculations in the filtering step require conditioning at each time t on all M^t possible histories of the switching variables $S_1, \dots, S_t$, meaning that the computational cost grows exponentially with time, which is not feasible. Instead, at each time t the Kim filter only considers the recent history of the switching variables (say, the M^2 possible values of (S_{t-1}, S_t)) to calculate relevant probabilities and expectations, which it then collapses to (approximately) recover (4) (for $\tau = t$). Further approximations are required in the smoothing step to make calculations tractable. In practice the Kim filter/smoother saves considerable computational time and provides accurate approximations to the Q −function and the log-likelihood, at least in low dimensional settings. Up to additive constants, the Q −function expresses as

$$\begin{aligned} Q(\theta;\hat{\theta}) = &\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{M} W^j_{t|T}\left(\log|\mathbf{R}_j^{-1}| - tr\mathbf{R}_j^{-1}\left(\mathbf{y}_t\mathbf{y}'_t - 2\mathbf{C}_j\mathbf{x}^j_{t|T}\mathbf{y}'_t + \mathbf{C}_j\mathbf{P}^j_{t|T}\mathbf{C}'_j\right)\right)\\ &+\frac{1}{2}\sum_{t=2}^{T}\sum_{j=1}^{M} W^j_{t|T}\left(\log|\mathbf{Q}_j^{-1}| - tr\mathbf{Q}_j^{-1}\left(\mathbf{P}^j_{t|T} - 2\mathbf{P}^j_{t,t-1|T}\boldsymbol{A}'_j + \boldsymbol{A}_j\mathbf{P}^{\bigcirc j}_{t-1|T}\boldsymbol{A}'_j\right)\right)\\ &+\frac{1}{2}\sum_{j=1}^{M} W^j_{t|T}\left(log|\textstyle\sum_j^{-1}| - tr\textstyle\sum_j^{-1}\left(\mathbf{P}^j_{1|T} - 2\mathrm{x}^j_{1|T}\boldsymbol{\mu}'_t + \boldsymbol{\mu}_j\boldsymbol{\mu}'_j\right)\right)\\ &+\sum_{j=1}^{M} W^j_{1|T}\log\pi_j + \sum_{t=2}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} W^j_{t-1,t|T}\log\pi_{ij}. \end{aligned} \tag{5}$$

Whereas the literature typically describes the EM algorithm for the general model (1), its extension to models (2) and (3) is not entirely straightforward, a least not when the maximum lag p in these models is greater than 1. In this case, additional modeling assumptions must be made on the joint distribution of $\mathbf{x}_{1:p}$ or alternatively of $\mathbf{x}_{(2-p):1}$. Also, conditioning must be carefully done in the smoothing part of the E-step to avoid issues of degeneracy and numerical inaccuracy. (For example, one should not naively condition $\mathbf{x}_{(t-p+1):t}$ on $\mathbf{x}_{(t-p):(t+1)}$ and vice-versa).

3.2 M-step

The M-step consists in maximizing the Q −function (5) with respect to $\boldsymbol{\theta}$. In the absence of constraints on $\boldsymbol{\theta}$, this amounts to a simple least squares problem in linear regression and the solution $\widehat{\boldsymbol{\theta}}$ can be found analytically. But even then, some regularization may be required to ensure that the transition matrices $\mathbf{A}_{\ell l}$ define invertible, stationary processes. Typical parameter constraints are fixed coefficients constraints (e.g. make covariance matrices $\boldsymbol{\Sigma}_j, \mathbf{Q}_j$ and/or $\mathbf{R}$ diagonal), equality constraints across regimes, and scaling constraints on the norms of $\mathbf{C}$ or $\mathbf{C}_j$. We rigorously enforce any number of these constraints in our software package via a projected gradient approach.

3.3 Initialization

Given that the EM algorithm is only guaranteed to converge to a stationary point of the likelihood function, choosing good starting points is essential to increase the chances that the EM converges to a global maximum. We propose two initializations methods for the switching dynamics model (2) and then adapt them to the switching observation model (3). We assume here that the higher-order model parameters M, p, r, are fixed.

Initialization for the switching dynamics model (method 1)

1. Perform the singular value decomposition (SVD) of the data: $\mathbf{Y} = \mathbf{UDV}'$, where $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$(the rows of $\mathbf{Y}$ should be centered on zero), $\mathbf{U}$ is of dimension $N \times m$ with $m = min(N, T)$ and $\mathbf{U}'\mathbf{U} = \mathbf{I}_m, \mathbf{V}$ Nis of dimension $T \times m$ with $\mathbf{U}'\mathbf{U} = \mathbf{I}_m$, and $D = \mathrm{diag}(d_1, \ldots, d_m)$ with $d_1 \geq \ldots \geq d_m \geq 0$. Let $\mathbf{U}_r$ and $\mathbf{V}_r$ be the submatrices obtained by taking the first r columns (i.e. singular vectors) of $\mathbf{U}$ and $\mathbf{V}$, and let $\mathbf{D}_r = diag(d_1, \ldots, d_r)$. Initialize the estimated observation matrix and estimated state vectors as $\widehat{\mathbf{C}} = \mathbf{U}_r$ and $\widehat{\mathbf{X}} = (\hat{\mathrm{x}}_1, \ldots, \hat{\mathrm{x}}_T) = \mathbf{D}_r\mathbf{V}'_r$.
2. Initialize $\widehat{\mathbf{R}}$ as the diagonal matrix containing the sample variances of the (rows of the) residual matrix $\mathbf{Y} - \widehat{\mathbf{C}}\widehat{\mathbf{X}}$.
3. Force the $\widehat{\boldsymbol{\mu}}_j$ and $\widehat{\boldsymbol{\Sigma}}_j$ to be equal across regimes ($1 \leq j \leq M$) and set them to the sample mean and sample variance of xb1,...,xbp.

(Alternatively, the $\widehat{\boldsymbol{\mu}}_j$ can be set to zero and the $\widehat{\boldsymbol{\Sigma}}_j$ can be made diagonal for numerical stability.)

4. Divide the time range $\{1,\dots,T\}$ into κ subintervals $\tau_k = \{t_k + 1, \dots, t_{k+1}\}$ where $t_k = \lfloor kT/k \rfloor$ for $0 \le k \le \kappa$ and $\lfloor\cdot\rfloor$ denotes the integer part. Fit a VAR(p) model to each subseries $(x_t)_{t\in\tau_k}$ by ordinary least squares (OLS) and call $\widehat{\mathbf{A}}^{(k)}$ and $\widehat{\mathbf{Q}}^{(k)}$ the corresponding estimates of the transition matrices and innovation covariances.
5. Partition the $(\widehat{\mathbf{A}}^{(k)}, \widehat{\mathbf{Q}}^{(k)}), 1 \le k \le \kappa$, into M clusters with the K-means algorithm. Take the initial estimates $(\widehat{\mathbf{A}}_j, \widehat{\mathbf{Q}}_j)$, $1 \le j \le M$, as the cluster centers. Alternatively, refit the VAR(p) model by OLS to each of the M subseries associated with the clustering and take the resulting $(\widehat{\mathbf{A}}_j, \widehat{\mathbf{Q}}_j)$ as the initial estimates.
6. Let $\hat{S}_t$ be the estimated regime at time t: $\hat{S}_t = j$ if $t\epsilon\tau_k$ and $(\widehat{\mathbf{A}}^{(k)}, \widehat{\mathbf{Q}}^{(k)})$ belongs to cluster j. Set $\hat{\pi}_j = 1$ if $\hat{S}_j = j, \hat{\pi}_j = 0.01$ otherwise, and rescale so that $\sum_{j=1}^{M} \hat{\pi}_j = 1$. Set $\hat{\pi}_{ij} = \#\{t: \hat{S}_{t-1} = i,\ \hat{S}_t = j\}/\#\{t: \hat{S}_{t-1} = i\}$ for $1 \le i, j \le M$. For each i, replace any $\hat{\pi}_{ij}$ less than $0.01max(\hat{\pi}_{i1}, \dots, \hat{\pi}_{iM})$ by this value and rescale so that $\sum_{j=1}^{M} \hat{\pi}_j = 1$.

Initialization for the switching dynamics model (method 2)

Method 2 is identical to Method 1 in all aspects except for step 4 which uses a more sophisticated segmentation algorithm for the time series $(\hat{\mathrm{x}}_t)$, namely binary segmentation. Initially, a VAR(p) model is fitted to $(\hat{\mathrm{x}}_t)$ over its entire range $\{1,\dots,T\}$, yielding a sum of squared errors $\mathrm{SSE}(1,T)$. For each time point $1 \le t \le T$, one fits a VAR(p) over each of the subintervals $\{1,\dots,t\}$ and $\{t+1,\dots,T\}$, yielding a total sum of squares $\mathrm{SSE}(1,t) + \mathrm{SSE}(t+1,T)$. One then selects the time $\tau = \mathrm{argmin}_{1\le t\le T}\{\mathrm{SSE}(1,t) + \mathrm{SSE}(t+1,T)\}$ as a candidate change point. If the reduction in SSE is sufficient, say, $(\mathrm{SSE}(1,\tau) + \mathrm{SSE}(\tau+1,T)) \le (1-\epsilon)\mathrm{SSE}(1,T)$ for some small $\epsilon > 0, \tau$ is accepted as a change point and the initial time range $\{1,\dots,T\}$ is split in two subintervals $\{1,\dots,\tau\}$ and $\{\tau+1,\dots,T\}$. The process is then iterated for each new subinterval and so on so forth until no new change points are found. The resulting change points are denoted by $t_1 < \cdots < t_{k-1}$ with $t_0 = 0$ and $t_k = T$ as before. In practice, the tolerance can be selected by trial and error until a reasonable number κ of segments has been obtained. One may also impose a minimal distance between successive change points for faster computations and better interpretability of results.

The initialization method for the switching observations model builds on the above initializations. For reasons of space, we do not present it in this paper.

4. Statistical inference of model parameters

MLEs in linear Gaussian SSMs are consistent and asymptotically normal [6]. This result can likely be extended to the switching SSM (1) under mild assumptions on the Markov chain (S_t). One can thus in principle perform statistical inference on $\boldsymbol{\theta}$ using the limiting distribution of the MLE. We have found however that due to the high dimension of $\boldsymbol{\theta}$, common techniques to estimate the limiting covariance matrix (i.e. the inverse of the Fisher information matrix) are numerically unfeasible. As an alternative approach, we propose a parametric bootstrap method that enjoys a simple and easily parallelizable implementation.

1. Apply the EM algorithm of section 3 to the data $\boldsymbol{y}_{1:T}$ and denote by $\widehat{\boldsymbol{\theta}}$ the MLE of $\boldsymbol{\theta}$.
2. Draw a bootstrap replicate S_t^* of S_t according to the probabilities $\hat{\pi}$.
3. For $2 \leq t \leq T$, draw a bootstrap replicate S_t^* of S_t according to the probabilities $(\hat{Z}_{s_{t-1}^*,1},\ldots,\hat{Z}_{s_{t-1}^*,M})$.
4. Draw a bootstrap replicate $\mathbf{x}_1^*$ of $\mathbf{x}_1$ from $N\big(\widehat{\boldsymbol{\mu}}_{S_1^*}, \widehat{\boldsymbol{\Sigma}}_{S_1^*}\big)$.
5. For $2 \leq t \leq T$, draw a bootstrap replicate $\mathbf{x}_t^*$ of $\mathbf{x}_t$ from $N\big(\widehat{\mathbf{A}}_{S_{t-1}^*}\mathrm{x}_t^*, \widehat{\mathbf{Q}}_{S_1^*}\big)$.
6. For $1 \leq t \leq T$, draw a bootstrap replicate $\mathbf{y}_t^*$ of $\mathbf{y}_t$ from $N\big(\mathbf{C}_{S_t^*}\mathrm{x}_t\ , \widehat{\mathbf{R}}_{S_1^*}\big)$.
7. Apply the EM to the bootstrap sample $\boldsymbol{y}_{1:T}^*$ and denote by $\widehat{\boldsymbol{\theta}}^*$ the bootstrap replicate of $\widehat{\boldsymbol{\theta}}$.
8. Repeat steps 2–7 a large number of times, say $50 \leq B \leq 200$, to obtain the probability distribution of the bootstrap estimator $\widehat{\boldsymbol{\theta}}^*$ conditional on $\boldsymbol{y}_{1:T}$.

References

1. C. Bregler. Learning and recognizing human dynamics in video sequences. In Proc. IEEE Conf. Comp. Vision and Pattern Recognition, pages 568–574, 1997.
2. C. B. Chang and M. Athans. State estimation for discrete systems with switching parameters. IEEE Transactions on Aerospace and Electronic Systems, AES-14(3):418–425, 1978.
3. A. Doucet, N. J. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump markov linear systems. IEEE Trans. Signal Processing, 49:613–624, 2001.
4. Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. Neural Computation, 12(4):831–864, 2000.
5. J. D. Hamilton. Analysis of time series subject to changes in regime. J. Econometrics, 45(12):39–70, 1990.
6. J. D. Hamilton. Time series analysis. Princeton University Press, Princeton, NJ, 1994.
7. C.-J. Kim. Dynamic linear models with Markov-switching. J. Econometrics, 60(1-2):1–22, 1994.
8. C.-J. Kim and C. R. Nelson. State-Space Models with Regime Switching: Classical and GibbsSampling Approaches with Applications. The MIT Press, 1999.
9. G. J. McLachlan and T. Krishnan. The EM algorithm and extensions. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.
10. K. P. Murphy. Switching Kalman filters. Technical report, University of California Berkeley, 1998.
11. H. Ombao, M. Fiecas, C.-M. Ting, and Y. F. Low. Statistical models for brain signals with properties that evolve across trials. NeuroImage, 2017.
12. A.-V. I. Rosti and M. Gales. Rao-Blackwellised gibbs sampling for switching linear dynamical systems. In In Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP), pages 809–812, 2004.
13. R. H. Shumway and D. S. Stoffer. Dynamic linear models with switching. J. Amer. Statist. Assoc., 86(415):763–769, 1991.

Nu-Support Vector Regression for the identification of outliers in High Dimensional Data

Abdullah Mohammed Rashid[1], Habshah Midi[1], Waleed Dhhan[2&3], Jayanthi Arasan[1]

[1]Institute for Mathematical Research, University Putra Malaysia

[2]Babylon Municipalities, Ministry of Construction, Housing, Municipalities and Public Works, Babylon, Iraq.

[3]Scientific Research Centre, Nawroz University (NZU), Duhok, Iraq.

Abstract

High-dimensional data (HDD) refer to the situation where the number of unknown parameters which are to be estimated is one or several orders of magnitude larger than the number of samples in the data. As High-dimensional data occur as a rule rather an exception in areas like information technology, bioinformatics or astronomy, it is imperative to use efficient technique of modelling and analyzing such data to avoid misleading conclusion. Analyzing such data encounter many challenges and outliers turn out to be the major challenge when dealing with this data. It is important to detect outliers because they have on adverse effect on the values of the various estimates, which lead to a misleading conclusion. Several parametric and non-parametric methods have been developed to detect outliers, but these methods cannot deal with high dimensional data (HDD). The fixed parameters support vector regression (FP-SVR) is put forward to remedy this problem. Nonetheless, the FP-SVR which employs Eps-SVR is not very successful in the identification of mild outliers and other contamination scenarios. To remedy this problem, we propose to use Nu-SVR to detect extreme and mild outliers. The merit of our proposed method is confirmed by well-known examples and simulation study.

Keywords

Outliers; Robustness; Statistical Learning Theory; Support Vector Regression.

1. Introduction

Outliers come back terribly ofttimes in the real information set, and that they usually go unmarked. There are many forms of outliers in regression issues. Any observation that has a large residual is referred to as residuals outlier. Observations that are extreme in the y-direction are known as y-outliers or vertical outliers and that they are answerable for model failure. High leverage points are those observations that are far in X-direction, and that they also are moving the regression model. Habshah et al. (2009) highlighted that it's crucial to detect multiple high leverage points as they're answerable for

misleading conclusion concerning the fitting of regression model. Pena and Yohai (1995) noted that HLPs are in the main answerable for masking and swamping of outliers in regression models. It is currently evident that HLPs is also the prime source of collinearity influential observations (Imon and Khan, 2003). Habshah et al. (2011) recognized that high leverage point collinearity influential observations are those HLPs that can choosed the pattern of multicollinearity. There are many good papers of identification of high leverage points in linear model (Hadi 1993, Habshah et al 2009, Limet at, 2016 and Alguraibawi et al 2015). However, there are not much work has been focused on identification high leverage points in high dimensional data. High-dimensional statistics refers to statistical inference when the number of unknown parameters is of much larger order than sample size (Bühlmann, P., & Van De Geer, S. 2011). In real-life applications, samples are always subject to noise, or outliers.

The support vector machine (SVM) is one of the most important techniques used to deal with problems in high-dimensional data. Recently, many techniques have been developed which depends on SVM. Dhhan et al (2015) and Rana et al (2018) has developed method which depends on fixed parameters support vector regression (FP-SVR) to detect outliers and high leverage points (HLP's) for high dimensional data. Unfortunately, (FP-SVR) which employs Eps-SVR is not very successful in the identification of mild outliers and other contamination scenarios. To remedy this problem, we propose to use Nu-SVR to detect extreme and mild outliers. Section 2 briefly describe FP-SVR and propose method of Nu-SVR will discuss in Section 3, numerical example and simulation study are presented in Section 4, finally, concluding remark are given in Section 5.

2. Fixed Parameters Support Vector Regressions

A practical procedure involving fixed parameters ε −tube SV Regression (SVR) has been suggested in order to elevate the performance of the standard SVR in detecting outliers. This method is suitable to be applied as it has many advantages in terms of time taken as it consumes less time than the conventional methods and also able to detect outliers without even getting rid of them.

The advantage of non-sparseness of the ε −insensitive loss function has been utilized in the fixed parameters ε −tube SVR. As appeared in Ceperic et al. (2014) and Guo et al. (2010), the SVR model will rely on most training data if the value of threshold ε is very small and hence giving the non-sparse solution. At the point when the ε parameter is more than zero, almost certainly, a portion of the outliers are not considered as support vectors (fall inside the $\varepsilon -$ zone), inferring the requirement for further iterations for

identifying outliers accurately. Essentially, the recognition of outliers should be possible by utilizing the non-sparse $\varepsilon -$ tube loss function is given by

$$L_\varepsilon(y_i) = \begin{cases} 0 & if|y_i - f(x)| \le 0 \\ |y_i - f(x)| & otherwise \end{cases}$$

Thus, the convex optimization problem mentioned can be rephrase as shown below:

$$minimize\ \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{n}(\xi_i - \xi_i^*)$$

$$subject\ to \begin{cases} y_i - w.\Phi(x_i) - b \le \xi_i \\ w.\Phi(x_i) + b - y_i \le \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1,2 \ldots, n \end{cases}$$

Therefore, the last regression function of the non-sparse $\varepsilon\ -$ tube SVR and the load vector could be appointed to the following condition:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x_i.x) + b$$

$$w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\Phi(x_i)$$

As exhibited by Rojo-Álvarez et al. (2003), controlling the free parameters of SVM (C, ε and the part parameter h) outliers are less taken care, or it permits the decrease in the effect of outliers in the solution. As indicated by Üstün et al. (2005), the strength and robustness of the SVR depends primarily on the choice of C, in light of the fact that the most astounding α_i^* and α_i values, by meaning of the Lagrange system, are equivalent to C. More absolutely, an extremely high value of C creates in SVs with a high variance among α_i^* and α_i values, bringing signifivant loads. The most elevated Lagrange multipliers belong to the rare data point in the training data, is considered as an outlier as mentioned by (Jordaan and Smts 2004). The load vector increases at any increment in estimated point C, including the presence of outliers. In this circumstance, it is not difficult to control the effect of outliers dependent on C and the kernel characteristics.

The characteristic of kernel functions can also become a`nother reason to be taken into consideration. Based on Williams (2011), the SVM calculation is sensitive to the tuning decision (the class of kernel), thus it is essential to understand how kernel function works. The information basically follows two kinds of kernel such as exponential radial basis function and the linear function. Thus, two demonstrative techniques are going to be introduced for the suitability in a wide range of data

2.1 Radial Basis Function (RBF)

The Gaussian Radial Basis, is the commonly used type of kernel which is given by

$$\mathrm{K}(\mathrm{x}, \mathrm{x_j}) = \exp\left[-\frac{\|\mathrm{x} - \mathrm{x_j}\|^2}{2\mathrm{h}^2}\right]$$

where x is the explanatory variable, x_j is the fractions of x and h is the bandwidth kernel function. According to Rana et al. (2018), outliers can be detected for only variables Z_i by using cut-off-points as follows:

$$\mathrm{CP_{RBF}} = 2\mathrm{Med}|\mathrm{Z}_i| + 2\sqrt{\mathrm{var(Med)}}$$

where

$$Zi = f(x)$$

As this approach involves detecting all the outlier points by applying only with one iteration, the computational cost would be less than those of the conventional techniques. Additionally, it is suitable for non-expert users because it introduces fixed set of parameters. In the experimental result sections, the RBF kernel function is utilized with ($h = 1, \varepsilon = 0, C = 10000$), using the predicted values to detect outliers.

3. Nu- Support Vector Regression

Another class of learning calculation, spurred by consequences by the results of statistical learning theory (Vapnik, 1995) has been involved by Support Vector (SV) machines. They represent the decision boundary in terms of a typically small subset (Schölkopf et aI., 1995) of all training examples, called the Support Vectors which is initially created for example acknowledgment. Vapnik devised the so-called ε −insensitive loss function, according to (Deng et al, 2012) can be handle ε-SVR during a similar approach. $\varepsilon -$SVR is changed because the equivalent $v -$support vector regression ($v -$SVR), wherever the parameter ε is replaced by a meaningful parameter v.

$$L_v(y_i) = \begin{cases} 0 & if|y_i - f(x)| \le v, \\ |y_i - f(x)| - v & otherwise \end{cases}$$

which does not penalize errors below some $v > 0$, chosen a priori in order for this property to carry over to the case of SV Regression (Schölkopf et al, 1999). The primary issue of $v -$SVR based on (Chang et al, 2002) to Introducing the corresponding kernel $K(x, x*) = (\Phi(x).\Phi(x*))$ and can be rewritten as follows

$$minimize\ \frac{\|w\|^2}{2} + C\left(v\varepsilon + \frac{1}{\ell}\sum_{i=1}^{\ell}(\xi_i - \xi_i^*)\right)$$

$$subject\ \big((w, x_i) + b\big) - y_i \le \epsilon + \xi_i$$

$$y_i - \big((w, x_i) + b\big) - y_i \le \epsilon + \xi_i^*$$
$$\xi_i^* \ge 0, \quad \varepsilon \ge 0$$

Here and beneath, it is comprehended that $i = 1, \ldots, \ell,$ and that bold Greek letters signify ℓ −dimensional vectors of the relating factors. Presenting a Lagrangian with multipliers $\alpha_i^*, ⍰_i^*, \beta \ge 0$, we get the Wolfe double issue. In addition, based on Boser et al. (1992), we substitute a bit k for the dot item, comparing to a dot item in some element space identified with info space by means of a nonlinear guide Φ,

$$K(x, y) = (\Phi(x).\Phi(y)).$$

This prompts the $v - SVR$ Enhancement Issue: for $v \ge 0, C > 0$.

$$maximize\ W(\alpha^*) = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*)^* y_i - \frac{1}{2} \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*)\big(\alpha_j - \alpha_j^*\big)\, K\big(x_i, x_j\big)$$

Subject to

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0, \qquad 0 \le \alpha_i^* \le 0, \qquad \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \le C.v$$

Therefore, the last regression function of the Nu-SVR and the load vector can be defined as follows

$$f(x) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) k(x_i.x) + b$$

Another point for consideration is the characteristic of kernel functions. According to Williams (2011), the SVM algorithm is sensitive to the tuning choice (the type of kernel), so it is important to understand how kernel function works.

3.1 Bessel Function

Bessel function are one type of kernel, which is given by

$$k(x_i, x) = \frac{Bessel_{(v+1)}^{n}(\sigma \|x_i - x\|)}{(\|x_i - x\|)^{-n(v+1)}}$$

Outliers can be detected by constricting a cut-off-points for any Z_i the cut-off points for Nu-SVR based on Bessel function is given by

$$CP_{Bessel} = 1.05 * median|Z_i| + 3MAD(Z_i)$$

where

$$MAD(Z_i) = b\ med\{|Z_i - med(Z_i)|\}$$

As this approach involves detecting all the outlier points by applying it one iteration, the computational cost would be less than those of the conventional techniques. In the experimental result sections, the Bessel kernel function is utilized using the predicted values to detect outliers.

4. Numerical Example

The performance of the proposed method has been illustrated in rank-deficient data. An artificial data with $n = 20$ and $p = 50$ have been considered. Each variable is generated from normal distribution N (0,1), the good data is contaminated by replacing three observations (2,4&6) with arbitrary large numbers equal to 20. The results for Nu-SVR and FP-SVR will be compared to show the efficiency for the proposed method.

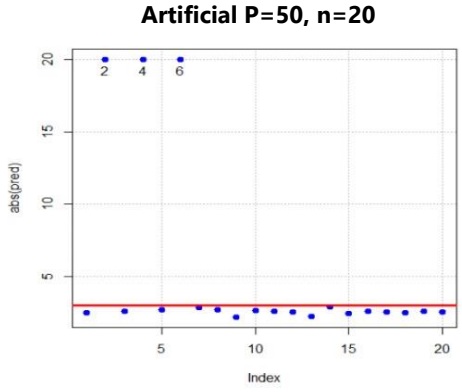


Figure 1: The number of detected outliers, Nu-SVR

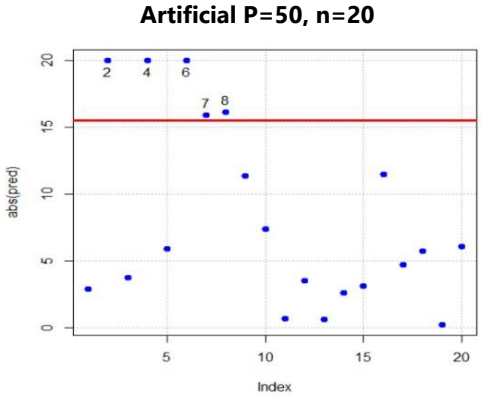


Figure 2: The number of detected outliers, FP-SVR

It can be observed from Figer1 & Figer2, the Bad performance of FP-SVR in the detection of outliers two observations are swamped as outliers on the other hand, the Nu-SVR successfully detects three observations as outliers.

A simulation study is convicted to further access the performance of our proposed method the same process is performed by contaminating the data with certain percentage of outliers, The replication is done for 1000 times and the result is displayed in Table1 it is interesting to now that percentage of correct detection of Nu-SVR is closer to 100% with low percentage of masking and swamping. Nonetheless, the FP-SVR is very poor. Whereby it's percentage of detection is very low with high masking effect.

Table1: Percentage of correct identification of BLP, masking and swamping for simulation data with 200 predictors (p=200)

θ	n	% Correct detection		% Masking		% Swamping	
		FP-SVR	Nu-SVR	FP-SVR	Nu-SVR	FP-SVR	Nu-SVR
5%	20	0.2	100	99.8	0	0	0.19
	40	8.05	100	91.95	0	0	0.01
	100	70.88	100	29.15	0	0	0.144
	150	75.16	93.333	24.84	6.6666	0	0.44333
10%	20	1.2	100	98.8	0	0	0
	40	44.05	100	55.95	0	0	0
	100	79.58	100	20.42	0	0	0.037

	150	86.58	100	13.42	0	0	0.128
15%	20	4.1	100	95.9	0	0	0
	40	53.5	100	46.5	0	0	0
	100	81.51333	100	18.48667	0	0	0.003
	150	85.45778	97.7778	14.54222	2.222	0	0.04866667
20%	20	10.575	100	89.425	0	0	0
	40	57.5375	100	42.4625	0	0	0
	100	83.73	100	16.27	0	0	0
	150	88.60333	100	11.39667	0	0	0.00266667

5. Conclusion

It is crucial to detect outliers in high dimensional data as it may give misleading conclusion about fitting of regression model. The FP-SVR has been developed to identify outliers in high dimensional data. Nevertheless, it is not very successful in detection outliers in high dimensional data. Hence, we developed Nu-SVR to remedy this problem. The numerical example signify that Nu-SVR is very successful in detecting high dimensional data in small and large samples.

References

1. Alguraibawi, M., Midi, H., and Imon, A. H. M. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, 1:1-12
2. Bagheri, A. (2011). *Robust Estimation Methods And Robust Multicollinearity Diagnostics For Multiple Regression Model in the Presence of High Leverage Collinearity-Influential Observations* (Doctoral dissertation, Universiti Putra Malaysia).
3. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM
4. Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
5. Ceperic, V., Gielen, G., & Baric, A. (2014). Sparse $$\varepsilon $$-tube support vector regression by active learning. *Soft Computing*, *18*(6), 1113-1126.
6. Chang, C. C., & Lin, C. J. (2002). Training v-support vector regression: theory and algorithms. *Neural computation*, *14*(8), 1959-1977.
7. Deng, N., Tian, Y., & Zhang, C. (2012). *Support vector machines: optimization-based theory, algorithms, and extensions*. Chapman and Hall/CRC.

8. Dhhan, W., Rana, S., & Midi, H. (2015). Non-sparse ϵ-insensitive support vector regression for outlier detection. *Journal of Applied Statistics*, *42*(8), 1723-1739.
9. Guo, G., Zhang, J. S., & Zhang, G. Y. (2010). A method to sparsify the solution of support vector regression. *Neural Computing and Applications*, *19*(1), 115-122.
10. Habshah, M., Norazan, M. R., & Rahmatullah Imon, A. H. M. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, *36*(5), 507-520.
11. Imon, A. H. M. R., & Khan, M. A. I. (2003). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. *Int. J. Stat. Sci*, *2*, 37-50.
12. Jordaan, E. M., & Smits, G. F. (2004, July). Robust outlier detection using SVM regression. In *IEEE International Joint Conference on Neural Networks* (Vol. 3, pp. 2017-2022).
13. Lim, H. A. and H. Midi (2016). Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics* 3(31): 859-877.
14. Rojo-Álvarez, J. L., Martínez-Ramón, M., Figueiras-Vidal, A. R., García-Armada, A., & Artés-Rodríguez, A. (2003). A robust support vector algorithm for nonparametric spectral analysis. *IEEE Signal Processing Letters*, *10*(11), 320-323
15. Rana, S., Dhhan, W., & Midi, H. (2018). FIXED PARAMETERS SUPPORT VECTOR REGRESSION FOR OUTLIER DETECTION. *Economic Computation & Economic Cybernetics Studies & Research*, *52*(2).
16. Schölkopf, B., Bartlett, P. L., Smola, A. J., & Williamson, R. C. (1999). Shrinking the tube: a new support vector regression algorithm. In *Advances in neural information processing systems* (pp. 330-336).
17. Schölkopf, B., Bartlett, P., Smola, A., & Williamson, R. (1998). Support vector regression with automatic accuracy control. In *ICANN 98* (pp. 111-116). Springer, London.
18. Vapnik, V. (1995). *The nature of statistical learning theory*, 1st ed. Springer, New York
19. Williams, G. (2011). Decision trees. In *Data Mining with Rattle and R* (pp. 205-244). Springer, New York, NY.
20. Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642-656
21. Üstün, B., Melssen, W. J., Oudenhuijzen, M., & Buydens, L. M. C. (2005). Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta*, *544*(1-2), 292-305

The robustness of two step estimation against heteroskedasticity and outliers in panel data

Habshah Midi[1], Nor Mazlina Abu Bakar[1,2]
[1]Institute of Mathematical Research, Universiti Putra Malaysia
[2]Centre of Management Sciences, Faculty of Economics and Management Sciences, Universiti Sultan Zainal Abidin, Terengganu, Malaysia.

Abstract
Robust methods in the literature are mainly proposed to withstand heteroscedasticity or outlying values separately. However, when abnormal data or outliers are present together with heteroskedasticity, two important least square assumptions are simultaneously violated and requires immediate solutions. In this study, Two Step Heteroscedasticity- and Outlier-robust or TSHO is proposed to withstand the influence of outliers and at the same time able to counter the heteroskedastic condition. TSHO assigned lower weights to outlying observations and able to produce more reliable fixed effect estimates than existing methods. This is confirmed by empirical evidence provided in the study via application on numerical data.

Keywords
panel data; heteroscedasticity, outliers; robust

1. Introduction

Robustness with respect to outliers is widely discussed for non-panel data regression where a large body of literature can be found in estimating robust parameters. On the contrary, robustness in the econometric literature is mainly discussed with respect to heteroskedasticity. Very little attention is given in developing robust estimators against outliers for panel data regression in the presence of heteroskedasticity. Feasible Generalized Least Square or FGLS is the conventional method used to protect against the effects of heteroskedasticity for fixed effect panel data model (Stock and Watson, 2008). In the presence of heteroskedastic errors, regression using Feasible Generalized Least Squares (FGLS) offers potential efficiency gains over Ordinary Least Squares (OLS) (Miller and Startz, 2018). However, the method is a modified version of OLS and very much affected towards outliers. Bias values will be produced and hence, the breakdown of FGLS. To the best of our knowledge, a study regarding both problems of heteroskedasticity and outlying values in panel data is non-existent. Thus, this study is considered to be among the first (if not the first) in solving simultaneous problems of heteroskedastic and non-normal errors. Therefore, this study is carried forward to achieve two main objectives. The first objective is to propose Two

Step Heteroskedasticity-Outlier (TSHO) robust estimator which consists of two important steps to guard against the vulnerability of outliers and also heteroskedastic errors. Crucial steps are taken to dampen the heteroskedastic condition and lower weights are assigned to outlying observations to produce more reliable fixed effect estimates. The second objective is to investigate the performance the newly proposed method and the performance behaviour of existing methods such as Robust Within Group Generalized M or RWGM (Bramati and Croux, 2007) and FGLS under the violations of the least square assumptions of non-normal and heteroskedastic errors. Empirical evidence on the performance of the newly proposed method, Two Step HO (TSHO) will be provided by comparing its performance with the existing methods under different data centering procedures.

The paper proceeds as follows. The next section presents the proposed TSHO based on weighted least squares. The method's first step consists of a procedure to correct heteroskedasticity. On the other hand, different weights are introduced in the second step to dampen the effects of outlying values. Section 3 provides the results of TSHO when applied to real data with conditions of heteroskedasticity and non-normality. Comparisons are made with other methods such as RWGM and also the conventional FGLS. Conclusion of the paper is presented in the Section 4.

2. Methodology

The newly proposed Two Step Heteroskedasticity-Outlier (TSHO) robust estimation involves two vital steps in which two different types of weights are determined to protect against the fatal effects of heteroskedasticity and outlying values. The first weight is evaluated by using F values (Djauhari, 2010) determined by the Robust Diagnostic-F from Midi and Abu Bakar (2015). On the other hand, the second weight is determined by log transformation of the residuals to dampen heteroskedasticity. The following steps describe the algorithm of TSHO and the derivation of the new weights.

Step 1: Transform panel data by robust MM-centering (Abu Bakar and Midi, 2015)

Step 2: Determine the first weights, by using the newly proposed Robust Diagnostic-F or RDF (Midi and Abu Bakar, 2015). Tukey's Biweight function is selected to determine ; meant to down weigh any observation with large residual. The tuning function of Biweight function is chosen to be 4.685 to provide a balance between efficiency and robustness (Wagenvoort and Waldmann, 2002). The diagonal elements of the second weighting matrix W_O is rewritten as

$$W_O = \min\left[1, \frac{\text{cutoff}_{RDF}}{F}\right]$$

with cut off point of RDF taken as $c\chi_r^2$ with

$$c = \frac{Tr(S_m^2)}{Tr(S_m)} \quad \text{and} \quad r = \frac{\{Tr(S_m)\}^2}{Tr(S_m^2)}$$

where S_m is the scatter matrix.

Step 3: Compute beta estimates, $\hat{\beta}_{LTS}$ and the residuals, $\hat{\varepsilon}_{it} = \tilde{y}_{it} - \tilde{x}_{it}^{'}\hat{\beta}_{LTS}$ based on initial estimates by Least Trimmed Square (LTS). It must be noted that the residuals derived in FGLS is based on OLS which is highly affected by outliers. LTS is used in our proposed method since the method is more useful and can provide robust estimates in a contaminated panel data.

Step 4: Determine the second set of weights, W_H by taking the log of squared residuals, $\ln(\hat{\varepsilon}_{it}^2)$ and regressed them on all of the fixed independent variables using Weighted Least Square (WLS) with W_0 as the robust weights. The fitted values, $\hat{g}$ are derived and the second weight, W_H for each data point is evaluated as $W_H = 1 \Big/ \sqrt{\exp(\hat{g})}$. This is a similar step taken in FGLS to protect against the heteroskedastic error terms.

Step 5: Perform WLS on $\tilde{y}_{it}$ and $\tilde{x}_{it}$ with combined weights, $W_{HO} = W_H \times W_O$. The WLS determines the estimates of β by minimizing the weighted sum of squares. Efficient estimates are expected to derive at the final iteration. Thus, the general solution of WLS to the newly proposed TSHO is formulated as

$$\hat{\beta}_{TSHO} = \left(\tilde{X}^{'} W_{HO} \tilde{X}\right)^{-1} \left(\tilde{X}^{'} W_{HO} \tilde{Y}\right).$$

The regression coefficients obtained from the newly proposed TSHO method are the desired estimates of the heteroskedastic multiple regression model in the presence of block HLPs.

3. Result

In this section, real world data set is considered to evaluate the performance of the newly proposed TSHO method. The Grunfeld data is a well-known, balanced panel data on 10 large United States (US) manufacturing firms over 20 years, for the years of 1935 until 1954. The data are taken from Baltagi (2013) and readily available in R programming (R Core Team, 2013) using plm package. There are various versions of the Grunfeld data which are circulated online. Various text books and articles in journals use different subsets of the original Grunfeld. Some of which contain errors in a few data points compared to the original data used by Grunfeld (1958) in his PhD thesis (Greene, 2016). The Grunfeld data consist of three variables and the model to be estimated is

$$I_{it} = \beta_0 + \beta_1 F_{it} + \beta_2 C_{it} + \varepsilon_{it}$$

where i indexes firms and t indexes years and variables
= Gross investment,
= Market value of the firm at the end of the previous year,
= Value of the stock of plant and equipment at the end of the previous year.

All figures in the Grunfeld data are in millions of dollars. The variables and reflect anticipated profit and the expected amount of replacement investment required (Greene, 2016). The Grunfeld data heavily suffers heteroskedasticity as indicated by the Figure 1 by which, the plot of residuals shows an apparent funnel shape form.

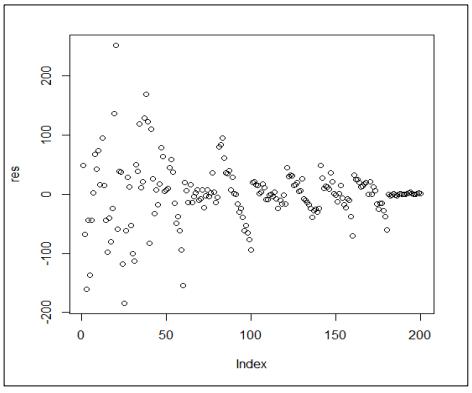


Figure 1: Plot of residuals for Grunfeld Data

The performance of the newly proposed TSHO method is compared to the existing RWGM-estimator under the robust data transformations of MM-centering. Its performance is also compared to the classical method of FGLS. Random block leverage contamination at 5% and 10% are presented in the heteroskedastic Grunfeld data by adding HLPs into the data set. Once contaminated, data are ready to be transformed by either the mean or MM-centering. The transformed data are then regressed by each of the method under study - either WG(OLS), FGLS or the newly proposed TSHO. It must be noted that the non-robust WG(OLS) and FGLS will only be applied to the mean-centered data. Otherwise, robust methods will be applied to the robustly transformed data. Results are reported in Table 1 for the beta estimates of the original and modified Grunfeld data. FGLS should provide more efficient estimates than WG(OLS) for the original, uncontaminated Grunfeld data.

Results show that both WG(OLS) and FGLS provide bias and wrong results when contamination is introduced into the panel data set. At 5% contamination level, for both WG(OLS) and FGLS bears a negative sign which

will give a completely different interpretation from the initial uncontaminated data. The results become worse as the intensity of contamination increases at 10% whereby negative signs are observed in for FGLS and for WG(OLS). The results show that both WG(OLS) and FGLS are highly influenced by the block HLPs which is due to their least square basis. Moreover, the mean-centering procedure may introduce more outlying values or HLPs into the data set and hence, the results become more distorted as the intensity of contamination increases. On the other hand, the robust estimators – RWGM and TSHO are able to give resistant and efficient beta estimates even though more HLPs are introduced into the data set. By being efficient means that they are able to provide similar results to the estimation by WG(OLS) in the original, uncontaminated data. It is also observed that the newly proposed TSHO is able to provide better results than RWGM as more HLPs are added into the data at 10% contamination.

Table 1: Beta estimates of the original and modified Grunfeld data with standard error in parentheses

Conta-mination	Estimate	WG(OLS)	FGLS	RWGM	TSHO
		Mean Centering		MM Centering	
0% (Original Data)	$\hat{\beta}_0$	3.541e-15 (3.6450)	-0.5573 (3.4372)	-0.4677 (0.8036)	1.9147 (1.3450)
	$\hat{\beta}_1$	0.1101 (0.0115)	0.1000 (0.0105)	0.0689 (0.0054)	0.0636 (0.0080)
	$\hat{\beta}_2$	0.3101 (0.0170)	0.2797 (0.0172)	0.1123 (0.0071)	0.1212 (0.0098)
5% (Modified Data)	$\hat{\beta}_0$	-2.219e-14 (7.4440)	2.5580 (6.8737)	-0.8041 (0.7898)	2.2533 (1.3819)
	$\hat{\beta}_1$	-0.0519 (0.0096)	-0.0604 (0.0107)	0.0605 (0.0051)	0.0600 (0.0061)
	$\hat{\beta}_2$	0.1175 (0.0225)	0.1498 (0.0255)	0.1167 (0.0067)	0.1242 (0.0092)
10% (Modified Data)	$\hat{\beta}_0$	5.265e-15 (8.7250)	2.0435 (8.3197)	-0.5590 (0.9403)	0.4358 (1.4389)
	$\hat{\beta}_1$	0.0007 (0.0072)	-0.0128 (0.0087)	0.0787 (0.0063)	0.0663 (0.0076)
	$\hat{\beta}_2$	-0.0086 (0.0160)	0.0269 (0.0199)	0.1081 (0.0082)	0.1358 (0.0109)

4. Discussion and Conclusion

Primarily, the least square fixed effect regression provides the best linear unbiased estimator (BLUE) under the assumptions of normally distributed, independent and identically distributed errors. However, the presence of simultaneous distortions towards normality and homoscedasticity of the error terms often lead to wrong statistical analysis and conclusions of the method. Thus, this study proposes heteroskedasticity and outlier-robust estimator and its algorithm are proposed to dampen the effects of heteroskedasticity and also high leverage values. In the first step of Two Step HO (TSHO) a procedure is taken to reduce the influence of heteroskedasticity by placing appropriate weights to the residuals. Consequently, the second step guards against the fatal effects of high leverage values by introducing robust weights. The TSHO uses residuals by RWGM(RDF) to warrant only true high leverage values to be given low robust weights. In this way, potential outliers or high leverage values are investigated and dealt with appropriately. The simulation and numerical studies reveal the reliability of the respective TSHO algorithm. Fixed effect data are completely distorted in the presence of the highly contagious block HLPs. The success of TSHO regression in providing efficient estimates under the condition of heteroskedastic and non-normal errors show that when weights can be estimated appropriately, weighted least squares becomes a superior least squares analysis. (Carroll and Ruppert, 1982; Ryan, 1997).

References

1. Abu Bakar, N.M. and Midi, H. (2015), Robust centering in the fixed effect panel data model, Pakistan Journal of Statistics, Vol. 31(1), 33-48.
2. Baltagi, B.H. (2013). The Econometrics of Panel Data. John Wiley and Sons, New York. ISBN: 978-1-118-67232-7.
3. Bramati, M. C. and Croux, C. (2007). Robust estimators for the fixed effects panel data model, Econometrics Journal. 10(3), 521–540.
4. Carroll, R.J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models, Annals of Statistics, 10(2), 429-4414
5. Djauhari, M. (2010). A multivariate process variability monitoring based on individual observations. Modern Applied Science. 4(10).
6. Greene, W. H. (2017). Econometric Analysis. 6th edition. Upper Saddle River. New Jersey: Prentice Hall.
7. Midi, H. and Abu Bakar, N.M. (2015). The performance of robust diagnostic-f in the identification of multiple high leverage points, Pakistan Journal of Statistics, Vol. 31(5), 461-472.
8. Miller, S. and Startz, R. (2018). Feasible generalized least squares using machine learning. SSRN Library.
9. R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
10. Ryan, T. P. (1997). Modern Regression Methods. New York: Wiley.
11. Stock, J. and Watson,M. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. Econometrica. 76(1):155–174.
12. Wagenvoort, R. and Waldmann, R. (2002). On B-robust instrumental variable estimation of the linear model with panel data. Journal of Econometrics. 106:297-324.

Household Income Survey 2019: The sampling methodology

Diyana Amalina Fadzil, Noor Masayu Mhd Khalili
Department of Statistics Malaysia, Putrajaya, Malaysia

Abstract

Household Income Survey (HIS) is one of surveys conducted by the Department of Statistics, Malaysia (DOSM). The survey was conducted for the first time in 1973 and subsequently carried out twice every five years, two surveys in each period of the Malaysian Development Plan. The main aims of HIS was to measure the economic well-being of the population, collecting information on household income distribution patterns according to various socio-economic features and providing basic data for calculating Poverty Line Income (PLI). Household income and poverty statistics are used for policy formulation and development plans especially for poverty eradication strategies and income distribution programs. This paper aimed to describe the sampling methodology used in designing HIS 2019 in Malaysia. In addition, this paper will emphasize on enhancement made in HIS 2019. A two stage stratified random sampling was adopted in the survey. The sampling frame was stratified by administrative district and urban/rural localities. The Primary Sampling Unit (PSU) is Enumeration Block (EBs) based on the information from 2010 Population Census. A total of 11,529 EBs were selected from the total EBs in Malaysia, where 7,557 and 3,962 EBs was selected from urban and rural areas respectively. The Secondary Sampling Unit (SSU) is Living Quarters (LQs) within the selected EBs. On the average, eight LQs were randomly selected from each selected EBs. All households within the selected LQs were included in the study. Sampling methodology is one of the critical success factors of a survey. Hence, a robust sampling methodology was used in executing this survey to ensure the representativeness of the study, thus will provide reliable findings for users particularly policy makers.

Keywords

Household Income Survey, HIS, income, sampling methodology

1. Introduction

Household Income Survey (HIS) is one of surveys conducted by Price, Income & Expenditure Statistics Division, Department of Statistics, Malaysia (DOSM) and is carried out twice in five years. The previous HIS was conducted in 2016 through 2017. HIS make available statistics concerning the household income and poverty to measure the economic well-being of the population.

These statistics is used primarily by the government as inputs for the formulation of national development plans and monitoring the Malaysia Plan. In addition, these statistics is also widely used by the researchers as well as individuals for further analysis and research purposes.

Survey Instruments and Data Collection Techniques

Personal interview approach is used in collecting HIS data. Training was given to the officers in DOSM state's office who were involved as enumerator in this survey. They will visit selected households to collect information on demography and income using a set of bi-lingual (Malay and English) structured questionnaires on the scopes of the survey.

There were several modules included in the questionnaire. The household income questionnaire contains identification particulars (would be head of household as any members whether male or female which is an income recipient and age 15 years and over), household member particulars,

individual and household income, annual household income payment during the last twelve (12) months and current transfer payment. To ensure the quality of the data, supervisor will perform certain checking procedures to detect and correct any error or omission during the survey.

Scope and Coverage

Target population for HIS 2019 is all households in Malaysia in both urban and rural areas. Individuals who live in residential institutions such as hostels, hotels, hospitals, old folk homes, prisons and welfare homes were excluded from the survey.

Sampling Frame

The sampling frame used for the HIS 2019 was based on the Household Sampling Frame i.e. a list of enumeration blocks (EBs) that were updated from time to time subsequent to the 2010 Population and Housing Census. Malaysia was geographically divided into 79,000 enumeration blocks (EBs). Furthermore, EBs were recognized as geographical contiguous areas of land with identified boundaries, contains about 80 to 120 living quarters with an average population of 500 to 600 people.

Basically, all EBs were based on the population size of the gazette boundaries. The EBs in the sampling frame was classified into either urban or rural areas. Urban areas were gazette area, with their adjoining built-up areas, which has a combined population of 10,000 or more as defined in the 2010 Population and Housing Census. Meanwhile, gazette area with population less than 10,000 could be classified as rural area. In year 2018, there were about 79,000 EBs in HIS sampling frame, with about 59,000 and 20,000 urban and rural EBs, respectively.

The sampling frame used to draw a HIS sample includes 78,976 enumeration blocks (EB) for urban and rural areas in strata 1 to strata 4 for Peninsular Malaysia as well as strata 1 to 0 in Sabah and Sarawak. The definition of strata is as below:

Strata Code	Strata	Description
1	Metropolitan	Population greater than 75,000
2	Large town	Population from 10,000 to 74,999
3	Small town	Population from 1,000 to 9,999
4	Rural	Population less than 9,999
5	Rural 1	Can reach from downtown in 30 minutes
6	Rural 2	Can reach from downtown in 30 minutes to one hour
7	Rural 3	Can reach downtown in one to two hours
8	Rural 4	Can reach from downtown in two to four hours
9	Rural 5	Can reach from downtown in four hours in one day
0	Rural 6	Can reach by helicopter, boat or other transportation and take more than one day.

Sampling Design

The two-stage stratified random sampling was adopted in this survey. The strata were the primary stratum, which made up of the states in Malaysia, including Federal Territories. The second stratum, which was made up of the administrative districts by state and the tertiary stratum was made up of the urban and rural formed within the second stratum.

The sampling involved two stages; the primary sampling unit (PSU) was the enumeration blocks (EBs) which the selections of the sample is done using Probability Proportionate to Size Sampling. Meanwhile, the secondary sampling unit (SSU) was living quarters (LQs) within each selected EB which the selections is done using Systematic Random Sampling. On an average, eight (8) LQs were randomly selected from each selected EBs. All households within the selected LQs were included in the study.

Sample Size

The optimum sample size estimation was based on several inputs i.e. total number of household 2018, average of household income, design effect and response rate from the previous HIS 2016/17. The estimation is calculated for each domain i.e. urban and rural at administrative district levels with several choices of Relative Standard Error (RSE) value (5.0%, 6.0%, 7.0% and 10.0%) to give an option of the sample size estimation. The final sample size for HIS 2019 is determined taking into consideration sample size and the RSE from previous survey (HIS 2016/17) as well as capability in terms of burden for data collection and cost involved. In general, sample size is increased for domain with

considerably high RSE. In contrast, the previous sample size is maintained for domain with low RSE. This is to ensure the reliability of the estimates produced.

Table 1: Sample size (EB) by state and strata for HIS 2019 and HIS 2016/2017

State	HIS 2019			HIS 2016/17
	Urban	Rural	Total	
Johor	643	325	968	936
Kedah	452	325	777	690
Kelantan	316	384	700	646
Melaka	286	46	332	329
N. Sembilan	224	185	409	395
Pahang	327	274	601	563
P. Pinang	565	69	634	638
Perak	594	300	894	794
Perlis	103	102	205	205
Selangor	1,136	197	1,333	1,192
Terengganu	316	235	551	489
Sabah	838	618	1,456	1,374
Sarawak	895	890	1,785	1,562
WP KL	702	-	702	702
WP Labuan	79	12	91	91
WP Putrajaya	81	-	81	66
Total	**7,557**	**3,962**	**11,519**	**10,672**

Improvements in HIS 2019 Sampling Design

Representative samples are important as they ensure that all relevant types of people are included in the sample and that the right mix of people is interviewed. If the sample isn't representative it will be subject to bias. Certain groups may be over-represented and their opinions magnified while others may be under-represented.

Generally, household's income has a correlation with the type of LQ and areas they are residing. Since the information of income for all household is not available in the sampling frame, this assumption is used as a proxy to ensure the sample of EBs and LQs consists of all income level in Malaysia. Hence, for HIS 2019, the number of sample size allocated for each urban and rural at administrative district level is proportionately distributed into type of LQ and details strata based on distribution of EB according to type of LQ as well as distribution of EB according to detail strata in the sampling frame.

Thus, would increase the representativeness of the samples. Table 2 below shows the EBs distribution by type of LQ and detail strata in Johor.

Table 2: Allocation of EBs by each strata and type of LQ for sample HIS 2019 in Johor

Types of Living Quarter	Population (No.)				
	Strata				Total
	1	2	3	4	
Apartment	231	9		18	258
Groups	63	7	4	1	75
Twinhouse	182	66	18	89	355
Condominium	93				93
Other	5	2			7
Townhouse	16	2		2	20
Shop office	92	21	6	5	124
Flat	616	69	8	74	767
Bungalow	805	464	207	1369	2845
Terrace	3143	1100	303	409	4955
Blank	2				2
Total	5248	1740	546	1967	9501

Types of Living Quarter	Sampel (No.)				
	Strata				Total
	1	2	3	4	
Apartment	10	19	10	23	62
Groups	1	2	1		4
Twinhouse	22	7	3	21	53
Condominium	2	8	1	12	23
Other	2	2	3	3	10
Townhouse	1			1	2
Shop office	3	7	2	5	17
Flat	18	21		13	52
Bungalow	129	93	31	97	350
Terrace	194	53	18	81	346
Blank	49				49
Total	431	212	69	256	968

Types of Living Quarter	Population (%)				
	Strata				Total
	1	2	3	4	
Apartment	89.5%	3.5%	0.0%	7.0%	100.0%
Groups	84.0%	9.3%	5.3%	1.3%	100.0%
Twinhouse	51.3%	18.6%	5.1%	25.1%	100.0%
Condominium	100.0%				100.0%
Other	71.4%	28.6%			100.0%
Townhouse	80.0%	10.0%		10.0%	100.0%
Shop office	74.2%	16.9%	4.8%	4.0%	100.0%
Flat	80.3%	9.0%	1.0%	9.6%	100.0%
Bungalow	28.3%	16.3%	7.3%	48.1%	100.0%
Terrace	63.4%	22.2%	6.1%	8.3%	100.0%
Blank	100.0%				100.0%

Types of Living Quarter	Sampel (%)				
	Strata				Total
	1	2	3	4	
Apartment	16.1%	30.6%	16.1%	37.1%	100.0%
Groups	25.0%	50.0%	25.0%		100.0%
Twinhouse	41.5%	13.2%	5.7%	39.6%	100.0%
Condominium	8.7%	34.8%	4.3%	52.2%	100.0%
Other	20.0%	20.0%	30.0%	30.0%	100.0%
Townhouse	50.0%			50.0%	100.0%
Shop office	17.6%	41.2%	11.8%	29.4%	100.0%
Flat	34.6%	40.4%		25.0%	100.0%
Bungalow	36.9%	26.6%	8.9%	27.7%	100.0%
Terrace	56.1%	15.3%	5.2%	23.4%	100.0%
Blank	100.0%				100.0%

Way Forward

Ideally, a good sampling design for HIS requires income information for each household in Malaysia particularly in the sampling frame. This is because based on this information strata according to income group can be formed to ensure that the selected sample covers all income groups. This will increase the representativeness which will ultimately produce the best estimate for the average household income in Malaysia. As an effort to realize this, starting from 2017, DOSM has begun collecting income information for every household. This information will also be collected during the 2020 Population Census. Hopefully with this effort, the sampling design for HIS will be improved in the near future.

References

1. Canberra Group (Expert Group on Household Income Statistics) (2001): Final Report and Recommendations, Ottawa. (Download under http://www.lipsroject.org/links/canberra/finalreport.pdf).
2. Department of Statistics, Malaysia. (2017). Household Income and Basic Amenities Survey Report 2016.
3. EPU (Economic Planning Unit) Malaysia. (2012). *Sosio-economic Statistics: Household Income and Poverty*. Retrieved November 20, 2012, from http://www.epu.gov.my/householdincome- poverty.
4. Khazanah Research Institute. (2018). The State of Households 2018: Different Realities (3rd ed.).
5. Malaysia. 2013c. *Household Income Survey Report*. Kuala Lumpur: Department of Statistics.

Using 'rmapshaper' to Modify Boundary Files for Use in Linked Micromap Plots

Braden Probst, Jürgen Symanzik
Utah State University, Department of Mathematics and Statistics, Logan, UT, USA

Abstract
Linked micromap plots have been in use since their creation in the 1990's. Initially, the underlying code was complex and the shapefiles used to represent the spatial boundaries were not easily obtained or efficient to use. Using modern software, the process of modifying and simplifying shapefiles has become more accessible, facilitating the ability to more easily create and analyze linked micromap plots --- and doing so on a larger scale.

Keywords
Spatial Data, Data Visualization, Map, Shapefile

1. Introduction

Over the last three decades, linked micromap plots (LMplots) have been developed as a way to visualize potential trends within some spatial geographic data and within some associated statistical variable(s) [6]. In the statistical software environment R [4], LMplots can be created via the 'micromap' R package [3] for example.

While there exist different plot types that tie statistical data to spatial data, such as choropleth maps, LMplots employ some visualization concepts that, in many cases, make the statistical trends and spatial patterns more visible and clear. The first of these concepts used by LMplots that becomes one of its greatest strengths, is the concept of small multiples. The benefit here is rather than having all of the data shown in a single (large) map, the data are spread into several smaller and comparable maps.

The basic structure of LMplots is an array of several vertical panels with the underlying data being oriented by row. Typically, although not exclusively, the first of these panels is used to plot spatial data in the form of a shapefile for the geographic regions of interest, along with the labels and names of these regions in panels three and four. One or more columns of statistical data corresponding to the regions in the spatial panel are included in the subsequent panels. The panels plotting the statistical variables can take on a variety of forms, although dot plots and line plots seem to be the most common choices. The entire LMplot is then sorted by one of the variables that were included. The sorting adjusts all of the rows according to the specific variable and direction of the sorting. The complete LMplot shows whether

there is any correlation among the selected statistical variables and whether or not the data are also spatially correlated, as reflected in the filled-in shapefiles found within the spatial panel.

In Section 2 of this article, we describe the Mapshaper and 'rmapshaper' software that form the basis for creating meaningful shapefiles for further use in LMplots. The motivation why such modified shapefiles are needed is given in Section 3. The methods used to modify shapefiles from within R are summarized in Section 4. An example for Canada follows in Section 5. We finish with a discussion and conclusion in Sections 6 and 7, respectively.

2. Methodology

Mapshaper, and its web browser version *mapshaper.org* [2], is a software tool introduced to the public by Matthew Bloch and Mark Harrower in 2006 as an open-source tool to modify and simplify shapefiles. Before the introduction of Mapshaper, there were a few software programs that allowed users to modify shapefiles as needed. However, the real strength of Mapshaper is that it was the first, free software program that provided a WYSIWYG (what you see is what you get) approach for users that are not trained or prepared in modifying shapefiles.

While most of the functionality of the Mapshaper software is still done manually through a command line, the syntax of the commands is easy to understand even for beginners. Further, with the shapefile being updated in close to real-time, users are able to see exactly how each modification would appear in the final product.

Andy Teucher, in 2016, brought even more accessibility to the modifying and simplifying of shapefiles when he released an R package titled 'rmapshaper' [7]. 'rmapshaper', at its core, provides an R wrapper to the functionality provided within *mapshaper.org*. As 'rmapshaper' is still fairly new, not every command in Mapshaper has been directly translated into a standalone R function. While 'rmapshaper' is still being updated to include more commands as R functions, Andy Teucher has provided access to the entirety of Mapshaper commands through the inclusion of the command *apply_ mapshaper_commmands*(). While not necessarily ideal, this still does allow us to use all of the commands to tailor our shapefiles to be exactly what we need.

3. Motivation

While shape files for various countries and other regions of interest are readily available through the internet, e.g., via the Database of Global Administrative Areas (GADM) [1], plotting shapefiles in their "raw" format, especially coastal regions, is a time consuming and computationally expensive process, especially when plotting several repeated plots. Further, many of the

polygons in an existing shapefile carry no meaning when filled with color due to a small size or the location of the polygon.

While there exist various software solutions for simplifying the boundaries of shapefiles, this has never been done on a large scale. Further, finding simplified boundaries for a given country that are ready to use poses a different challenge in that the simplified shapefiles available for public use are not available in a single location, but are scattered throughout various packages in R and publications.

The ultimate goal of this article is to provide access to shapefiles for as many countries as possible to allow for wider use of spatial maps, particularly linked micromap plots. In addition to the recommended modifications and shapefiles corresponding to them, there should be tools accessible to users to further modify shapefiles in a manner that would be more meaningful for them.

4. Methods Used in R:

There are several types of modifications that are available for addressing the problems caused by the underlying geography of a region. Some of these modifications will require the removal of certain polygons, enlarging areas that are otherwise difficult to see when compared to larger surrounding areas, or manipulating the shape altogether by moving a given area to a location that may be more meaningful.

4.1 Thinning Boundaries

In order to understand why the process of thinning needs to be applied to a shapefile, one needs to understand how a shapefile is constructed. At first glance, a shapefile is simply the outline of some region, often a geographic or political boundary. These outlines are created by connected line segments tracing around the borders. The lines themselves can be considered as a dense collection of points.

Shapefiles tend to be more complicated when the administrative boundaries are created to follow the geographic features in the area. This is especially true in coastal regions where the boundaries tend to not follow a straight line. When plotting such regions, this can be an intensive and time-consuming process for a computer. In order to simplify these boundaries, thinning is often the first step.

The process of thinning is as intuitive as it sounds. If we consider the many individual line segments that make up a shapefile's boundaries to be a dense collection of points lying next to each other, then the process of thinning is selecting a proportion of these points to keep in our final shapefile.

Shown in Figure 1, there are two outputs for Mexico using two different proportions of total points to keep. These shapefiles also act as an example of

how thinning can simplify both internal and external boundaries, as well as removing small islands that would hold no value in the final map.

4.2 Filter Islands

In most cases, the thinning step described in Section 4.1 would remove any regions that would otherwise be affected by this function. However, this step does exist as a valid option for cases where further thinning would remove regions that should be kept in the final shapefile. This option allows a user to retain more of the original geographically correct shapefile, but still filter out islands that carry no meaning in the context of LMplots.

(a) 10% of original points (b) .05% of original points

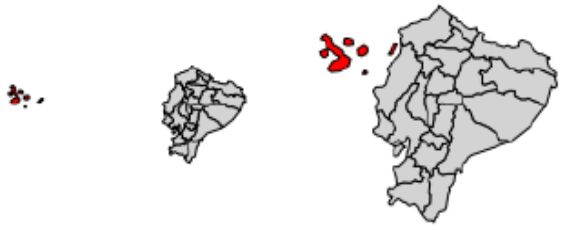
4.3 Moving Regions

There are two cases where we need to move regions around to be in a more meaningful location for interpreting the spatial trend shown by a linked micromap plot.

4.3.1 Moving Detached Regions

In cases where an island (island, here, refers to both the geographic definition of a body of land surrounded by water and to a region of a country that may completely be surrounded by a different country or other geographic feature) that lies far outside the mainland region, the plotting region is stretched and causes the entire region to appear smaller.

In Figure 2, the shapefile for Ecuador is plotted, after having been thinned previously. Due to the Galapagos Islands being so far from mainland Ecuador, the entire country appears to be smaller than it should be. While this may not be a problem for all spatial plots, in a LMplot, we would have several small repetitions of this shapefile making each appear even smaller. The effect of a single horizontal shift of the Galapagos Isalnds is that the entire region appears larger, even when nothing else was scaled differently.



(a) (b)

Figure 2: The Galapagos Islands are so far off the coast from mainland Ecuador that, upon plotting, the plotting window is filled with white space (a). Do doing nothing more than shifting the islands to be closer (b), the entire country appears enlarged.

4.3.2 Moving Small Regions Outside of the Boundaries

Another instance that may require us to move a region is where a region is so small compared to the regions surrounding it that in its unchanged state it is either hard to see or not visible at all. Such is the case of many city-states and federal districts within countries.

To fix this issue, a small region may be better represented as an "island" outside of a country's boundaries. In order to do this, it is likely not a fix that can use the same approach we used to fix the sizing issues that the Galapagos Islands caused in the Ecuador shapefile. In order to accomplish this in a meaningful way, we will not only have to enlarge a region but also change the coordinates of a region to have it show up adjacent to the main land.

In Figure 3, the region of Littoral is the main urban area in Benin, a country in Western Africa. When plotting without modifications, the region is barely visible along the southern border. Further, if we were to simply enlarge the region in this case, by the time is was large enough for plotting, it would be masking the surrounding regions. By representing the region as an "island", we are able to make the region much more visible without intruding on the visibility of surrounding regions. As this modification may not be ideal, particularly to locals in the region, this is meant to be viewed as one possible

(a) (b)

Figure 3: Modification done on the country of Benin in West Africa. (a) shows a thinned but otherwise unchanged shapefile, where along the southern border, the capitol region of Littoral is nearly unseen. By enlarging and shifting Littoral to be represented as an island it is now visible and interpretable.

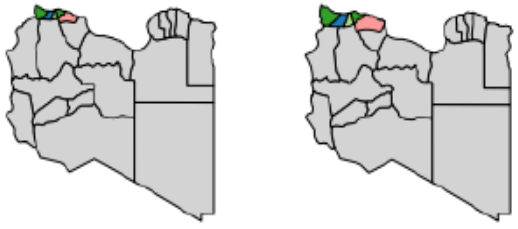
solution addressing the fact that simply enlarging the region would cause other issues that may not be as easy to fix.

4.4 Enlarging Regions

In some countries, we may have some regions that are hard to see when plotted in LMplots, but that are surrounded by large regions. In these types of cases, it is possible to address the small regions without having to change too much of the underlying shape of the shapefile. In these cases, we can enlarge the areas of the smaller regions into the larger surrounding regions.

When enlarging a region in Mapshaper, the scaling of a region is centered at the midpoint of the original region. This process does not save the existing borders between two neighboring regions and instead, the enlarged region and all of its neighbors now have overlapping borders. In order to correct this, we first shift the enlarged region to the location we desire and then cut out a hole in the existing region, effectively removing parts of the surrounding polygons and, by extension, the overlap caused by the enlarging of a region. We can then put our modified enlarged region in the hole that we had previously cut out.

Figure 4 shows a potential modification performed on the country of Libya, whose northwestern states are comparatively small when compared to the other states in the country.

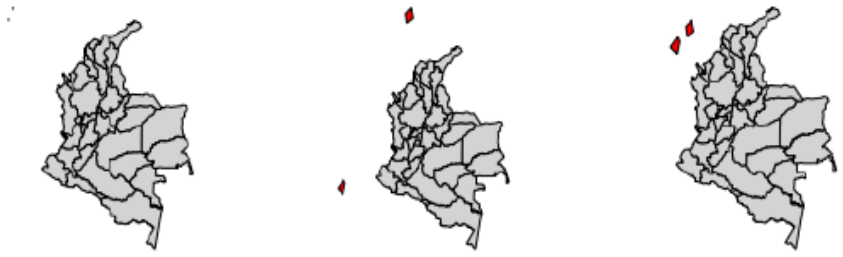


(a) (b)

Figure 4: Modifications done in Libya in northern Africa. (a) shows a thinned shapefile with no other changes. The five north eastern states are comparatively small and would not be meaningful in an LMplot. In (b) those five regions have all been enlarged to be more visible.

4.5 Moving Disjoint Areas of the Same Administrative District

In some cases, it may not be enough to employ any single method from the previous sections for a given subregion within the area of interest. These cases are rare, but employ multiple methods from the previous sections. Take the South American country of Columbia, for example, shown in Figure 5. No single transformation could have provided a representation that allows a user to see all regions within the country in a meaningful way. In fact, multiple steps have to be performed on a single region to place it in a meaningful location.



(a) (b) (c)

Figure 5: Modification on Columbia in South America. In (a) the two islands of San Andrès and Providencia are barely seen at all to the northwest of the mainland country. By enlarging, they are seen in (b), but not show together or located in meaningful locations. By splitting the polygons and shifting separately we obtain the representation shown in (c)

5. Example Using a Modified Canada Shapefile

Using the methods detailed in the previous section to create a more meaningful shapefile and using real data, Figure 6 shows a completed linked micromap plot for Canada comparing the statistical variables of GDP per capita [9] and theft rate per capita [8] for the 13 Canadian provinces and territories. Sorted by GDP per capita, the spatial data shows that the GDP share per individual in each province increases in the western parts of the country, the exception being British Columbia. This is due to the small populations in these provinces compared to the production and resources in each of these regions.

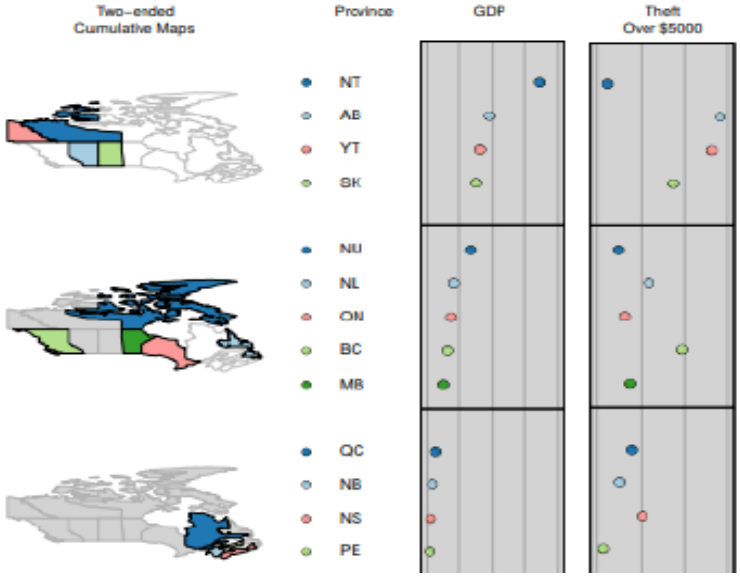


Figure 6: The geographic data is sorted by GDP per capita. The lower GDP areas are the more populous areas, while the less populous are more rich in natural resources. The theft rate follows the the same trend as the GDP with the correlation between the two being roughly 21%. The Northwest territories are clearly seen as an outlier with respect to the theft data.

While not a particularly strong correlation (r = 0.21), the rate of thefts totaling over $5,000 also increases with the GDP per capita. A strong outlier here are the Northwest Territories which have the highest GDP per capita and one of the lowest theft rates.

6. Discussion

We have created simplified ready-to-use shapefiles for the vast majority of countries in the world, including countries in North and South America, Africa, Europe, and Asia. It is our goal to create an R package that exists as a data repository for the storage on not only these proposed shapefiles, but the documentation for each country showing the exact changes that were made from the original shapefiles.

It is to be understood that the modifications made during the creation of our R package are recommendations and may not represent the best way to display the regions in the countries in question. For this purpose, we will include multiple ways for users to further modify the shapefiles we will provide, or start from scratch with their own shapefiles. The methods for further modification will be an approach using data tables in R, or through a Shiny app [5], included in our R package.

Due to some geographic limitations, not every country in the world will be included in our R package initially. Countries, such as Chile, that cover a larger range in the latitudinal direction than the longitudinal direction will be excluded as the current formatting for LMplots is not optimal for these shapes. Countries that are archipelagos consisting of many small islands will also be excluded. Countries with less than five administrative regions at the selected level will also be excluded as a LMplot would not be a meaningful way to display spatial patterns for so few regions. Lastly, it should be noted that due to the technical capabilities of data structures used in the R 'micromap' package, countries that contain one or more regions that are entirely surrounded by another single region will also be excluded until a solution is found for this limitation.

7. Conclusion

With new advancements, such as 'rmapshaper', the modification of shapefiles is more accessible than it ever has been. With tools such as these, shapefiles can be prepared on a large scale and can easily address issues that exist within the geography captured in the shapefile. As shapefiles are provided ready-to-use in our R package, linked micromap plots can be created and used to display spatial data alongside statistical data.

References

1. Database of Global Administrative Areas (GADM) (2018). *GADM Maps and Data*. URL: https://www.gadm.org.
2. Harrower, M. & Bloch, M. (2006). *MapShaper.org: A Map Generalization Web Service*. URL: https://www.mapshaper.org.
3. Payton, Q. C., McManus, M. G., Weber, M. H., Olson, A. R.& Kincaid, T. M. (2015). "micromap: A Package for Linked Micromaps". *Journal of Statistical Software* 63(2), pp. 1–16. URL: http://www.jstatsoft.org/v63/i02/.
4. R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.
5. RStudio, Inc (2013). *Easy Web Applications in R.* URL: http://www.rstudio.com/shiny/.
6. Symanzik, J. & Carr, D. B. (2008). "Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data". In: *Handbook of Data Visualization*. Ed. by C. Chen, W. Härdle, and A. Unwin. Berlin, Heidelberg: Springer, 267–294 & 2 Color Plates.
7. Teucher, A. & Russell, K. (2018). *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations*. R package version 0.4.0. URL: https://CRAN.R-project.org/package=rmapshaper.
8. Wikipedia, The Free Encyclopedia (2018). *Crime in Canada*. [Online; accessed 20-August-2018]. URL: https://en.wikipedia.org/wiki/Crime_in_Canada.
9. Wikipedia, The Free Encyclopedia (2018). *List of Canadian Provinces and Territories by Gross Domestic Product*. [Online; accessed 20-August-2018]. URL:
10. https://en.wikipedia.org/wiki/List_of_Canadian_provinces_and_territories_by_gross_domestic_product.

A hierarchical mixed effects model for batch cytometry data

Sharon X. Lee
School of Mathematical Sciences, University of Adelaide, South Australia, Australia

Abstract

Flow cytometry is an important tool in the diagnosis and monitoring of immunological diseases such as lymphomas, leukaemia, and AIDS. It is frequently used in immunological research, pre-clinical trials, and clinical diagnosis. However, these data are challenging to model and analyze due to the large number of observations and the inherent structure of the batch of samples. Moreover, it is known that they typically exhibit non-normal features such as asymmetry and heavy-tailedness. This paper considers the problem of jointly modelling multiple cytometry data that comes from the same batch. In particular, one of the aims is for the model to provide an automated segmentation of the data. To achieve this, we adopt a hierarchical mixture model approach to provide a probabilistic clustering of the data, together with skew component distributions to cater for non-normal clusters. Furthermore, our tool is designed to handle inter-data variations via the incorporation of a random effects model. Examples from real cytometry experiments will be used to demonstrate the effective od our approach.

Keywords

cytometry; mixed model; mixture model; clustering; skewness

1. Introduction

Flow cytometry is a powerful tool for characterizing single cell properties. It is routinely used in both clinical and research immunology. Its ability to study particles at the single-cell level renders it widely useful in many biomedical fields. After staining with fluorophore-conjugated antibodies (or markers), the sample is placed in a flow cytometer where cells are passed through a laser beam one at a time. The light emerging from each cell are captured and quantitated by different detectors. Modern cytometers can measure up to 30 markers simultaneously at a rate of 10,000 cells per second. This generates datasets of massive size in a high-throughput manner.

A critical part of the analysis of flow cytometry data is the segmentation of cells into different cell populations according to their properties. This task is currently carried out manually where an analyst would visually discriminate between different clusters or groups of points based on sequential bivariate projections of the data. Not only is this process laborious and error-prone, but

it is also limited by the non-reproducibility of results, the non-scalability to massive datasets, and the difficulty in detecting high-dimensional relationships from low-dimensional projected spaces. Due to this, recent efforts have turned to machine learning, computer science, and statistics to provide computational tools to analyse these data. Some reviews and comparisons of these methods can be found in Aghaeepour et al. (2013, 2016) and Weber et al. (2016).

Many of these methods employed a mixture model-based approach, whether explicitly or implicitly. This is because mixture model provides a convenient framework to characterize the heterogenous populations within the data. The task of cell segmentation then translates to the traditional problem of model-based clustering. However, it is well-known that cytometry data typically exhibit non-normal features including skewness and long-

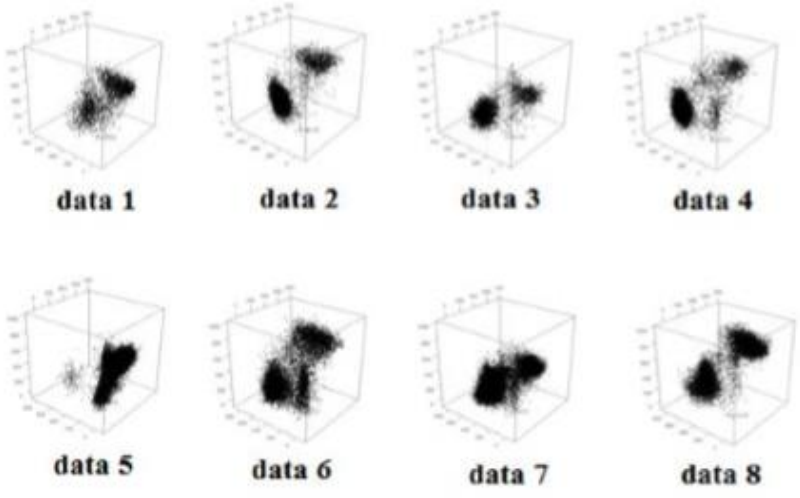


Figure 1. Eight examples from the batch of 16 DLBCL data. Large variations in the shape, size, and location of clusters be observed across the different data.

tailedness. Hence traditional mixture models find it challenging to handle the non-normal cluster shapes. To mitigate this, some methods attempt to normalize or transform the data (Lo et al, 2009) while some others considered merging multiple components from an overfitted model to allow for asymmetric clusters (Aghaeepour et al, 2011, Mosmann et al, 2014). While these approaches may alleviate the problem, it is ideal to have a flexible model that can directly handle non-normal clusters. We thus adopt skew mixture models for this task in this paper.

When analysing batch cytometry data, that is, a cohort of cytometry data with similar characteristics (for example, data from patients diagnosed with a certain disease or from the same individual across different time points), there

is the more challenging task of matching cell populations across the different data in the batch. An example is shown in Figure 1, where large variations in the size, shape, and location of the clusters of points can be observed across the data. This presents significant difficulties to automated methods as there can large variations between the data. Intuitive approaches such as fitting each data separately or pooling all data into an aggregate dataset fails to account for the inter-data relationship. Ad hoc approaches such as normalizing the data in a pre-processing step (Hahne et al, 2010) or matching the clusters in a post-hoc manner (Pyne et al., 2009) does not utilize all available and useful information. In particular, there is no information sharing between the data during the clustering step.

This paper presents Hcyto (Hierarchical model for cytometry data), a direct and automated approach for analysing batch cytometry data that inherently takes into the account variations between and within the data. We adopt a hierarchical approach to handle inter-data variations, where each data is conceptualized as an instance of a template mixture model. Under this framework, each data is modelling by an individual mixture model that is an affine transformation of the template model. An appealing advantage of this approach is that components of the individual mixture models are automatically aligned across the data. Another advantage is computational efficiency as clustering and aligning are performed at the same time without the need of additional pre-processing or postclustering steps. Furthermore, by adopting skew component densities, our approach can directly accommodate the non-normal features of the data. This avoid the need to search for a suitable transformation for each data or to determine how to merge components. To illustrate our approach, we apply Hcyto to real cytometry datasets, demonstrating favourable performance against other methods.

2. Methodology

The Hcyto model consists of two levels: the upper level for between-data variations and the lower level for within-data variations. The former is a mixed effects model whereas the latter is a finite mixture of skew distributions. The upper level model intrinsically links the lower level models to a batch template model - another finite mixture of skew distributions — that describes the overall characteristics of the batch. To facilitate discussion, we now introduce some notations. Let y_{jm} be a p-dimensional vector consisting the measurements of p markers on the j^{th} cell of data m, where $j = 1, \dots, n_k$ and $m = 1, \dots, M$. Here n_m denotes the total number of cells in data m, and M is the total number of data in the batch.

Lower level model

We adopt a finite mixture of skew t-distributions to model and cluster the cells in a data. Let there be g distinct cell populations in the data. Then the density of y_{jk} is given by

$$f(y_{jm};\ \Psi_m) = \sum_{i=1}^{g} \pi f(y_{jm};\theta_{im}), \qquad (1)$$

where Ψ_m denotes the vector containing all the unknown parameters of the mixture model for data m, θ_{im} denotes the vector containing the parameters for i^{th} component density, $f(y_{jm};\theta_{im}$, denotes the density of i^{th} component $i = 1, \dots, g$, and $\mu_{im}, \dots, \mu_{gm}$ are the mixing proportions. Here, the component density takes the form of a multivariate skew t (MST) distribution. More specifically, it can be expressed as

$$f(y_{jm}\ ;\ \mu_{im}, \Sigma_{im}, \delta_{im}, v_{im})$$
$$= 2\ t_p(y_{jm}\ ;\ \mu_{im}, \Omega_{im}, v_{im})\ T_1\left(\delta_{im}^T \Omega_{im}^{-1}(y_{jm} - \mu_{im})\sqrt{\frac{v_{im}+p}{v_{im}+d_{ijm}}}; 0, 1 - \delta_{im}^T \delta_{im}, v_{im} + p\right), \qquad (2)$$

where $t_p(\cdot\ ;\ \mu, \Omega, v)$ denotes the density of the p-variate t-distribution with mean μ, scale matrix Ω, and degrees of freedom v, and $t_p(\cdot\ ;\ \mu, \Omega, v)$ is the corresponding distribution function. In the above, we let $\Omega_{im} = \Sigma_{im} + \delta_{im}\delta_{im}^T$ and $d_{ijm} = (y_{jm} - \mu_{im})^T \Omega_{im}^{-1}(y_{jm} - \mu_{im})$. The parameter δ is a p-dimensional vector that regulates the skewness of the MST density. It is worth noting that there is currently no standard definition for a MST distribution. The formulation above follows the parameterization by Pyne et al. (2009) and is equivalent to the commonly used version proposed by Azzalini and Dalla Valle (1996) after re-parameterization; see Lee and McLachlan (2013) for a technical discussion. We can write $y_{jm} \sim MST_p(\mu_{im}, \Sigma_{im}, \delta_{im}, v_{im})$ When Y_{jm} has the distribution (2). From (1) and (2), each cluster in data m is characterized mathematically by a (data- and) cluster-specific MST distribution with parameters θ which consists of the elements of μ_{im}, the elements of δ_{im}, v_{im}, and the distinct elements of Σ_{im}.

Upper level model

To link the data-specific models (1) together, we first conceptualize these models as instances of a batch template model. This template is also characterized by a MST distribution and provides an 'overall' mathematical representation of the batch. The instances are then viewed as variations of the template. We adopt a random effects (RE) model to describe these inter-data variations. More specifically, we let the data-specific location vectors j4;, be affine transformation of the template location vector μ_{im} that is, we let

$$\mu_{im} = a_{im} \circ \mu_i + \beta_{im} \qquad (3)$$

Where a_{im} and β_{im} are independent random effects (RE) terms that govern the scaling and translation of μ_{im} from μ_i, respectively. These RE terms are independently distributed as

$$\alpha_{im} \sim N_p(1_p, A_i)$$

and

$$\beta_{im} \sim N(0, b_i)$$

Where $\beta_{im} = \beta_{im} 1_p$ and 1_p, is a p-dimensional vector with all elements being one. It follows that the batch template has a mixture of MST distributions with component distributions given by

$MST_p(\mu_i, \Sigma_i, \delta_i, v_i)$ for $i = 1, \dots, g$. This template is useful not only as a representative summary of the batch that facilitates visualization, but can be a powerful tool in downstream analyses such as across-batch comparisons and new sample classification. The latter can assist in clinical diagnosis of diseases.

Fitting the Hcyto model

The expectation-maximization (EM) algorithm (Dempster et al., 1977) has become a standard tool for carrying out maximum likelihood estimation of the parameters of finite mixture models. As both the lower- and upper- levels of Hcyto can be written as a mixture model, these models can be fitted via the EM algorithm. The technical details are omitted due to length restrictions, but the procedure for the lower-level models are similar to that for the MST mixture model by Pyne et al. (2009). They can be expressed in a hierarchical form involving a normal, a gamma, and a half normal random variable. With the upper-level, a further layer is added to this hierarchical form, leading to

$$\begin{aligned} Y_{jm}|\mu_{im}, u_{ijm}, w_{ijm}, z_{ijm} &\sim N_p(\mu_i + \delta_{im}|u_{ijm}|, \Sigma_{im}) \\ \mu_{im}|z_{ijm} &\sim N_p(\mu_i, M_i A_i M_i^T + B_i) \\ u_{ijm}|w_{ijm}, z_{ijm} &\sim HN(0,1) \\ w_{ijm}|z_{ijm} &\sim Gamma(\frac{v}{2}, \frac{v}{2}) \\ z_{ijm} &\sim Multi_g(\pi_m) \end{aligned}$$

where M is the diagonal matrix with diagonal elements given by μ_i, B_i is the diagonal matrix given by $b_i I_p$, HN denotes the (univariate) half-normal distribution, Gamma(·) denotes the gamma distribution, and $Multi_g(\pi_m)$ denotes the multinomial distribution with g categories and probabilities $\pi_m = (\pi_{m1}, \pi_{m2}, \dots, \pi_{mg})^T$. Note that although the Hcyto model consists of two levels, the model fitting procedure simultaneously estimate all parameters of the model (that is, including both the lower and upper levels) in a single step. No pre-processing or post-hoc steps are required.

3. Result

For demonstration, we consider the diffuse large B-cell lymphoma (DLBCL) dataset provided by the flowCAP I contest (Aghaeepour et al., 2013). It contains a collection of 30 data sampled from patients diagnosed with DLBCL. These data were gated manually by experts to provide a benchmark for evaluating the performance of computational methods. In this batch, there were 16 data that were determined to have three major cell populations. Thus, we will focus on these 16 data. As can be observed from Figure 1, there are substantial differences between the data. In particular, the location of the clusters varies considerably; see, for example, the upper right cluster in data 2 seems to have shifted vertically down in data 7. Another interesting observation from Figure 1 is that the changes in the abundance of the clusters is even more remarkable. The lower cluster in data 5 appears to be lightly populated whereas the same cluster in data 7 is densely populated. If we model the data individually, it is likely that it will misses these very small clusters in the data. On the other hand, if we pool all the data together and fitted a single model to it, the large variations in cluster locations will adversely affect the accuracy of the model, leading to high error rates in cell segmentation. This is also manifested in a contour plot (not shown) where the components show large contours in order to accommodate a wider range of data points. This is a scenario where Hcyto can provide more reasonable and accurate results.

Upon applying Hcyto to this batch, we obtain a parametric model for each data as well as an overall parametric template of the batch of 16 data. The three components of these mixture models are automatically matched across the data. It can be observed from Figure 2 that Hcyto can handle the inter-data variations quite well, as evident from the closed fitted contours. Although there are very few observations/cells in the red cluster of data 5, Hcyto was able to identify and model this cluster. Another remark from Figure 2 is that the cluster shapes differs between the data. For example, the blue cluster ranges from fairly spherical (data 8), elongated (data 5), to asymmetrical (data 1). It is also of interest to note Hcyto correctly matches all clusters across the data.

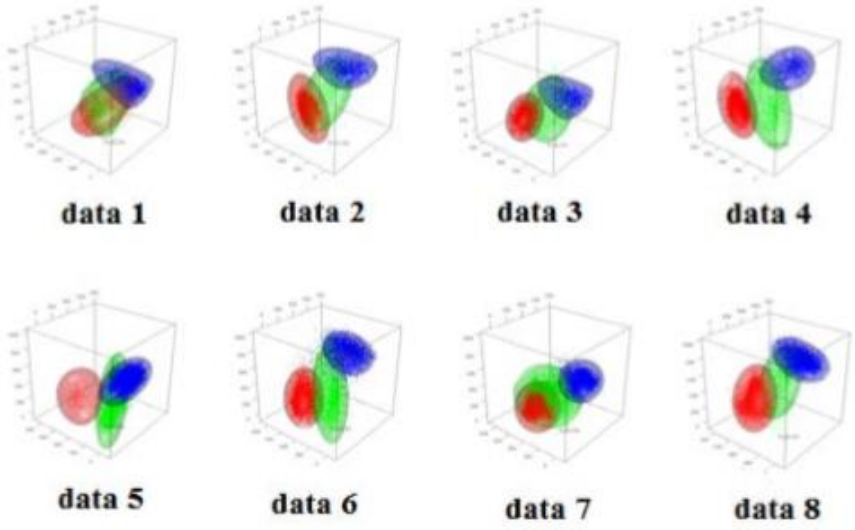


Figure 2. Contours of the fitted individual mixture models provided by Hcyto for the eight example data shown in Figure 1. Automatic clustering of the data are provided by Hcyto where clusters are automatically aligned across the data in the batch (shown as same colour across the batch).

Data	Hcyto	HDPGMM	FLAME
1	0.979	0.949	0.742
2	0.975	0.866	0.758
3	0.967	0.675	0.405
4	0.931	0.905	0.407
5	0.966	0.896	0.570
6	0.696	0.795	0.486
7	0.934	0.905	0.454
8	0.991	0.877	0.556
9	0.705	0.939	0.401
10	0.932	0.928	0.628
11	0.964	0.556	0.463
12	0.872	0.589	0.607
13	0.994	0.704	0.400
14	0.976	0.592	0.521
15	0.990	0.885	0.754
16	0.988	0.617	0.459
ACCR	0.929	0.796	0.538

Table 1. Segmentation performance of Hcyto compared to HDPGMM and FLAME. Hcyto obtained a higher correct classification rate (CCR) for many data in the DLBCL batch, achieving significantly higher average CCR (ACCR) than the other two methods considered.

To assess the performance of Hcyto in a quantitative way, we calculated the correct classification rate (CCR) which is the proportion of observations that were correctly classified according the benchmark gating results provided

by manual analysis. The CCR was calculated separately for each data and the results are shown in Table 1, together with the results by HDPGMM (Cron et al., 2013) and FLAME (Pyne et al., 2009). The later method adopts a cluster matching step in a post-hoc manner. It can be observed from Table 1 that Hcyto obtained a higher CCR than HDPGMM and FLAME for most of the 16 data. This is also supported by the average CCR across the batch, where Hcyto obtained 0.929 compared to 0.796 and 0.538 obtained by HDPGMM and FLAME, respectively.

4. Discussion and Conclusion

The clustering and alignment of cell populations across multiple data is an interesting and challenging problem. The proposed Hcyto method adopts a hierarchical approach to automatically segment and match these clusters, with implicit models that can directly handle non-normal distributional features. The methodology is motivated and demonstrated by cytometric data analysis, but is applicable to other types of data with similar structure. Results from the real example shows that Hcyto provides improved accuracy compared to other algorithms that adopt intuitive approaches such as pooling and post-hoc cluster matching. Future work may look at the scalability of the Hcyto framework for larger data and extend it for use in downstream analyses such as identification of discriminatory features, supervised classification of unlabelled data, and longitudinal modelling of batches.

References

1. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. (2011). Rapid cell population identification in flow cytometry data. Cytometry Part A 79A:6–13.
2. Aghaeepour, N., Finak, G., The FLOWCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann, T., Gottardo, R., Brinkman, R.R., Scheuermann, R.H. (2013). Critical assessment of automated ow cytometry analysis techniques. Nature Methods 10, 228-238
3. Aghaeepour, N., Chattopadhyay, P.K., Chikina, M., Van Gassen, S., Kurs, M., Malek, M., McLachlan, G.J., Qui, P., Saeys, Y., Stanton, R., Tong, D., Wang, K., Nolan, G., Finak, G., Gottardo, R., Mossman, T., Scheurmann R., and Brinkman, R. (2016). Benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. Cytometry: Part A 89A, 16-21
4. Azzalini, A., Dalla Valle, A.(1996). The multivariate skew-normal distribution. Biometrika 83, 715726 Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh SK, et al. (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. PLoS Computational Biology 9.

5. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1–38. Hahne F, Khodabakhshi AH, Bashashati A, Wong CJ, Gascoyne RD, Weng AP, Seyfert-Margolis V, Bourcier K, Asare A, et al. (2010). Per-channel basis normalization methods for flow cytometry data. Cytometry A 77:121–131.
6. Lee, S.X., McLachlan, G.J. (2013). On mixtures of skew-normal and skew t-distributions. Advances in Data Analysis and Classification 7, 241-266
7. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, et al. (2009). Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Science 106:8519–8524.
8. Lo K, Hahne F, Brinkman RR, Gottardo R. (2009). flowClust: A Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics10:145.
9. Mosmann TR, Naim I, Rebhahn J, Datta S, Cavenaugh JS, Weaver JM, et al. (2014). SWIFT— Scalable clustering for automated identification of rare cell populations in large, highdimensional flow cytometry datasets, Part 2: Biological evaluation. Cytometry Part A 85A:422– 433.
10. Weber, L.M., Robinson, M.D. (2016). Comparison of clustering methods for high-dimensional singlecell flow and mass cytometry data. Cytometry A 89A, 1084-1096

Incentives to improve response rate through electronic survey

Rafliza Ramli, Maslina Samsudin, Nor Rafidah Mat Hashim@Kasim
Department of Statistics, Malaysia

Abstract

In line with the recent technology development, Department of Statistics Malaysia (DOSM) has introduced electronic survey (e-survey) to expedite the collection process for Quarterly Construction Survey (QCS) since first quarter of 2015 known as e-QCS. Various approaches have been implemented to encourage respondents to use the e-QCS. However, the most significant and successful approach is the rewarding of additional Continuous Contractor Development (CCD) points to respondents who attended e-QCS Hands On Session conducted regularly by DOSM Perak and responded via e-QCS before the deadline. The 5 CCD points was awarded by Construction Industry Development Board (CIDB) for state of Perak as an incentive to respondents i.e. contractor for the purpose of renewal their CIDB's licence. Since the introduction of this award in early 2017, a significant increase of more than 30.0 percent has been recorded in e-QCS response rate for the state of Perak.

Keywords

Electronic survey; Quarterly Construction Survey (QCS); Continuous Contractor Development (CCD) points

1. Introduction

With nearly ubiquitous computer network access around the world, online data collection via e-survey are being made available to researchers. E-survey is a web-based survey instrument, constituting the questionnaire in the server network that can be accessed by other organisation through a web browser (Karen, J. J., Kevin, G. C., & Bernard, J. J., 2007; Habsah, S., 2014). E-survey provides a fast and easy alternative to hardcopy submission. Generally, the real time data is available through the esurvey and analysis can be accomplished through integrated system.

According to Karen, J. J., Kevin, G. C., & Bernard, J. J. (2007), three most common reasons for choosing an e-survey over conventional face-to-face interview are (1) decreased costs, (2) faster response times, and (3) increased response rates. Although previous studies has been mixed on the realization of these benefits, basically, researchers agree that faster response times and decreased costs are attainable benefits, while factor influencing the response rates vary based on variables beyond administration mode alone. Research also shows that the amount of incentives does not improve response rate in a

linear way (Fan, W., & Yan, Z., 2010). According to American Association for Public Opinion Research, the response rate is generally defines as the number of completed units divided by the number of eligible units in the sample (Fan, W., & Yan, Z., 2010).

Obtaining significant response rate via e-survey has been a major concern for survey researchers. Habsah, S. (2014) found that the response rate was very low in the first two years after the implementation of e-survey, mainly due to the inexperienced or disinclined of respondents to response to the new system. Various strategies have been introduced to increase participation of e-survey including organising a hands-on session with respondents (Jamaliah, J., 2012). Offering hands-on session can benefit the respondents as they will be guided on how to fill in and complete the questionnaire via e-survey system. The session can be held at the researcher's premise or respondent's premise.

Rewarding the respondent with an incentive is often used to increase the response rate of e-survey (Fan, W., & Yan, Z., 2010). Incentives for e-Survey normally in the form that can be easily transferred in the electronic environment such as redeemable loyalty points, gift certificates and provision of survey results. A combination of financial incentives, online and traditional advertising, public relations and marketing efforts might also be used to attract response rate via e-survey (Singer, E., & Ye, C., 2013).

In line with the modernisation of data collection, the Department of Statistics Malaysia (DOSM) has started using e-survey since 2008 for the International Trade-In Services Survey. The application of the method was then extended to various surveys including Quarterly Construction Survey (QCS). According to Habsah, S. (2014), the implementation of e-survey is to accomplish the DOSM's aspiration to have a better management of data collection operation, whereby: (1) data submission by the respondent become more efficient, (2) duration for data production become shorten, (3) operation cost are reducing, and (4) respondent's confidence level is improving.

In 2015, at the beginning of the introduction of e-QCS, the e-QCS response rate for the state of Perak was very low. The response rate was less than 4.0 percent in the first two years. Various approaches have been implemented to encourage respondents to use the e-QCS system. In this study, we shared the most significant and successful approach undertaken by DOSM Perak to increase the e-QCS responses using a reward system in collaboration with CIDB Perak.

2. Methodology

In 2006, DOSM conducted the Quarterly Construction Survey (QCS) which was based on the project approach for the reference period of first quarter 2006. The survey covers all main contractors with value of projects of

RM500,000 and above registered with the Construction Industry Development Board, Malaysia (CIDB). The main objective of the QCS is to collect and compile data on the value of construction work done for the purpose of publishing the Quarterly Construction Statistics and also being used in the compilation of Gross Domestic Product (GDP). The survey is conducted using mail questionnaire and e-survey. The respondents are given two weeks to complete and return the questionnaire to the Department. After the said period, field visits are conducted to obtain response from the establishments which have not returned the questionnaires.

As an incentive to improve response rates among these respondents, a strategic partnership between CIDB Headquarter in Kuala Lumpur and DOSM Headquarter was made. It was agreed that contractors selected for QCS and responded to the survey will be given 5 CCD points as a reward to them. The rewarding system which was introduced in 2010, has increased the overall QCS response rate tremendously. The response rate at the close of the survey was above 85 per cent.

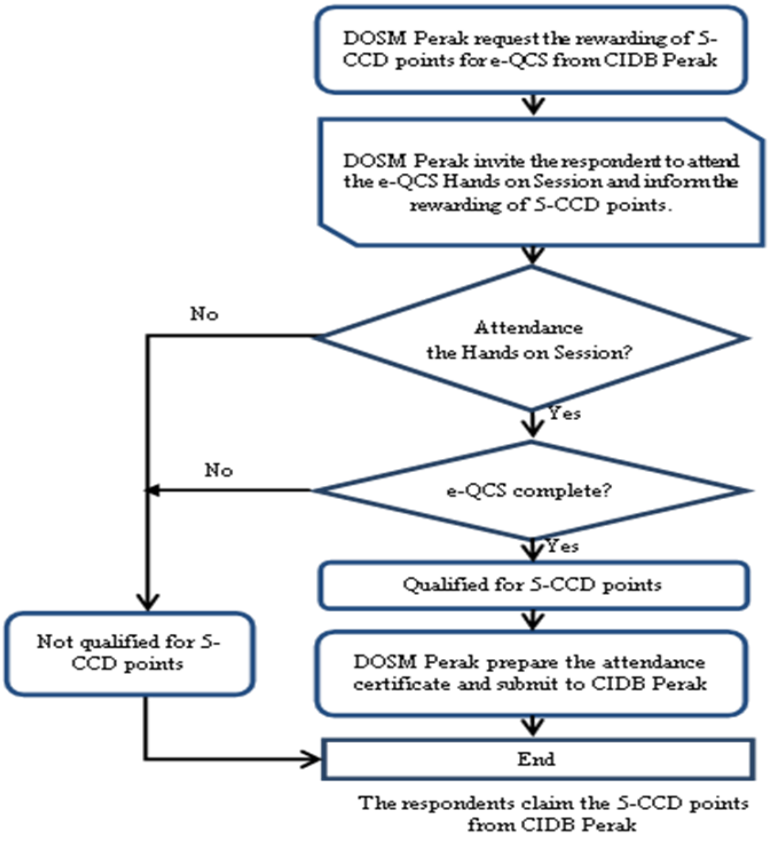
This reward is important for the respondent especially for the new contractors or small and medium contractors. The contractors are required to obtain certain number of the CCD points in order to renew contractor's registration. The CCD points will only be awarded once they attend and contribute to the programmes organised by CIDB with the objective to enhance the contractors' knowledge, professionalism, skills, expertise in the construction industy and to expand networking among them.

Emulating the same approach, DOSM Perak which is responsible for approximately 5 percent of the QCS total sample (average 300 contractors for state of Perak in every quarter) has engaged with CIDB Perak to overcome poor response of e-Survey amongst QCS respondents by rewarding additional 5 CCD points awarded by CIDB (state of Perak). This is an initiative by DOSM Perak to encourage e-survey responses through the e-QCS system. In order for the respondents to attain additional 5-CCD points, they are required to participate in the e-QCS hands on sessions conducted regularly by DOSM Perak and responded via e-QCS before the deadline. Since the introduction of this reward system, DOSM Perak has conducted six (6) hands on sessions for new contractors, which involved 88 new contractors.

Figure 1 shows the flow of the reward system applied in DOSM Perak. At the beginning of the year, DOSM Perak will request the rewarding of 5-CCD points for e-QCS from CIDB Perak. Upon approval of CIDB Perak, DOSM Perak will invite the respondents to attend the e-QCS hands on session conducted at DOSM's premise. The respondents will be informed about the rewarding of 5CCD points by the CIDB Perak and they will be asked to bring the information relevant to the questionnaire. During the hands on session, the respondents will be guided on the usage of e-QCS system and shown how to fill in the e-

QCS questionnaire by the trained field officers. The respondents should complete the e-QCS during the hands on session. If they do not bring enough information during the session, they are allowed to complete the e-QCS at their office before the end of the survey month. The completeness of the e-QCS will be verified by the state personnel as the first stage of quality checking. Upon completion of the e-QCS, the respondents are qualified for the 5CCD point. DOSM Perak will prepare the attendance certificates and submit to CIDB Perak as proof for the respondents to claim the 5-CCD points from the CIDB Perak.

Figure 1: Flow chart on rewarding the additional 5-CCD points to e-QCS respondents



3.

Result

The introduction of the reward system, in collaboration with CIDB Perak starting in the Q1 2017, has resulted in an increase of QCS responses via the e-QCS system. The e-QCS response rates for Perak increased drastically from 3.0 percent in Q4 2016 to 18.8 percent in Q1 2017 and the number continued to improve to record 36.5 percent in Q3 2018. The e-QCS response rates was also above the target of key performance indicator set by HQ i.e. 15 percent

responses for 2017 and 20 percent response for 2018. The e-QCS response rates for Perak was higher than the national level (25.6 percent). Distribution by state also shows the e-QCS response rate for Perak was among the highest and this reward system at state level was only introduced by Perak. The result illustrates that the incentives which directly benefited the respondent have positive effect on the response rates of eSurvey.

Figure 2: e-QCS response rate for Perak, Q1/2016-Q3/2018

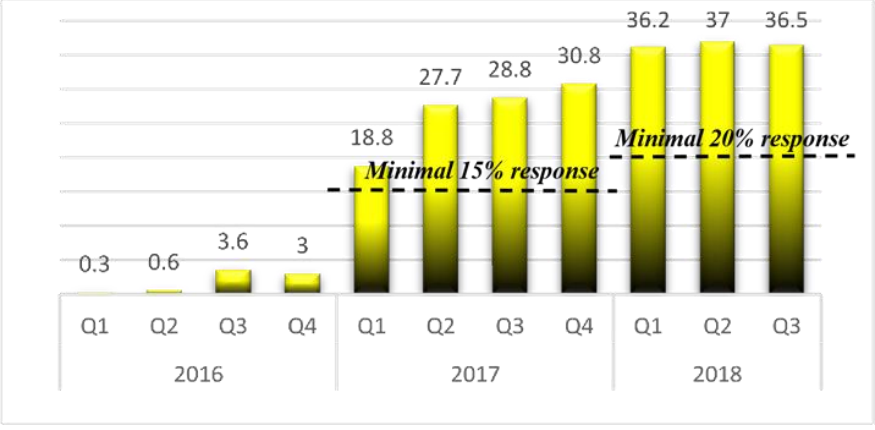
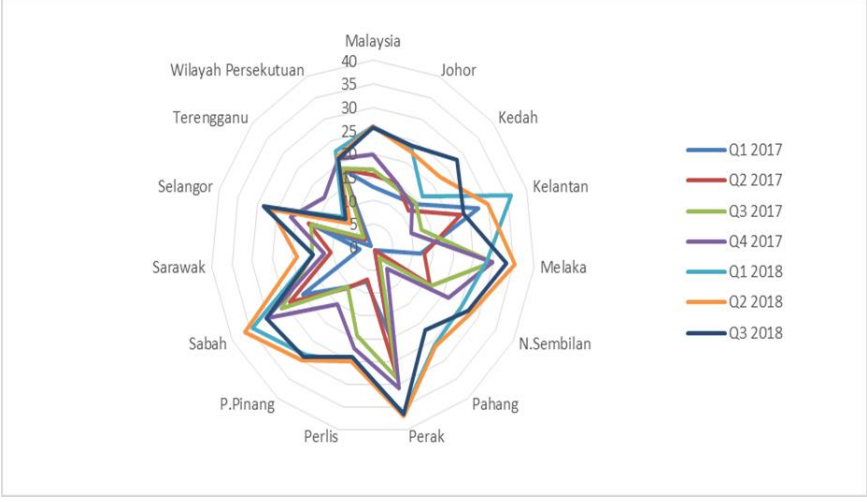


Figure 3: Distribution of e-QCS response rate by state, Q1/2017-Q3/2018

This achievement is also in line with six (6) hands on sessions conducted by DOSM Perak. Those who complete the e-QCS are qualified to receive additional 5 CCD points upon the DOSM Perak ratification to the CIDB Perak. To date, a total 88 contractors have benefited and received the additional 5 CCD points from CIDB Perak for their response and commitment in the e-QCS.

This types of incentive has created a good relationship and communication between the respondent, DOSM Perak and CIDB Perak which is a principle



factor that contributes to the success of e-QCS in a short period.

As part of continuous and future improvement, the respondents are required to provide feedback via the feedback form on the hands on session in terms of the e-QCS system and overall session. The response was positive with majority of the respondents were satisfied with the system which is user friendly and can be accessed easily. The respondents were willing to continue using the e-QCS if their company are still selected in the future. This feedback shows that how a survey is presented on the website can directly affect the response rate. The e-QCS was build using the scrolling designs, whereby all questions was display within one single webpage and allow the respondent to view the whole questionnaires and give answer.

In addition, DOSM Perak benefited a lot from the implementation of the reward system. The significant increase in e-QCS response rate has improved data management and quality, as the system helps to reduce errors that may be made by field enumerator or during data capture which may lead to non-sampling error. The system also increased efficiency by reducing time taken to produce data as the data processing was automatically done within the system, which can improve the reliability of data. The e-QCS system also guaranteed confidentiality, whereby only registered user will be given the access to system and the data given by the respondent shall not be disclosed or accessed without authorization. The increase in e-QCS response rates also contributed to the saving in the operation cost due to the reduction in paper printed, courier cost and field visit by the enumerators. In this regards, future research should examine the impact of implementing the e-Survey in reducing the operational cost and faster response times.

4. Discussion and Conclusion:

A strong collaboration with CIDB Perak allows attractive incentive to be offered to the respondents. Continuous efforts and engagement also play a pivotal role to sustaining the reward system. The rewarding of additional 5 - CCD points as an incentive to e-QCS respondents was proven to be effective and efficient to improve the response rate. The e-QCS response rate jumped above key performance indicator set by the HQ in the first quarter of its introduction and continued to show an increase ever since to record the highest e-QCS response rate of 36.5percent in Q3 2018. This shows a significant change in the culture of data collection in DOSM Perak and should be extended to the other states. We recommend other state offices to engage with CIDB of respective states and employ the incentive to improve e-QCS response rates. Furthermore, engagement and collaboration with related agency or associations in awarding certain amount of incentive should be considered to be implemented in other surveys.

References

1. Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review [Electronic Version]. Computer in Human Behavior 26 (2010) 132-139.
2. Habsah, S. (2014). The implementation of e-Survey in the Department of Statistics Malaysia [Electronic Version]. Meeting on the Management of Statistical Information Systems (MSIS 2014) Dublin, Ireland and Manila, Philippines 14-16 April 2014. Hassan, F., Samad, Z.A.,
3. Hassan, S., Che Mat, M., & Isnin, Z. (2010). Training the Construction Workforce: A Case Study of Malaysia. Proceedings W089 – Special Track 18th CIB World Building Congress, May 2010, Salford, United Kingdom, Page 230
4. Jamaliah, J (2012). eSurvey di Jabatan Perangkaan Malaysia. Journal of Department of Statistics Malaysia, Volume 1 2012, Page 41
5. Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. The ANNALS of the American Academy of Political and Social Science, 645(1), 112-141.
6. Karen, J. J., Kevin, G. C., & Bernard, J. J. (2007). E-survey methodology [Electronic Version]. Available from https://faculty.ist.psu.edu/jjansen/academic/pubs/esurvey_chapter_jasen.pdf Retrieved \ February 6, 2019
7. Department of Statistics Malaysia (Various series). Quarterly Construction Statistics Report. Available from https://newss.statistics.gov.my/newssportalx/ep/epProductFreeDownloadSearch.seam

The impact of Financial Sector Master Plan on cost efficiency of Malaysian Bank: an analysis of Stochastic Frontier Analysis

Azrie Tamjis
Financial Conglomerates Supervision Department, Bank Negara Malaysia

Abstract

The Asian financial crisis in 1997–98 left a severe impact on Malaysia's economy and banking system. This has forced the Malaysian government to undertake financial restructuring initiatives to restore market and public confidence, and to meet the ongoing challenges associated with market structure, financial innovation and globalisation. Therefore, Bank Negara Malaysia (BNM) introduced a ten-year Financial Sector Master Plan (FSMP) to strengthen domestic banks and the regulatory structure, and to promote the banks' efficiency by stimulating a competitive banking industry through financial liberalisation. The crisis for banks in Malaysia and the region has been extensively studied. However, empirical studies of the post-crisis period, and the implementation of the FSMP, remain limited. Hence, a data set of all banks in Malaysia, which covers the period 2000–2011, was employed to examine the effect of the FSMP's initiatives on Malaysian banks' efficiency between 2000 and 2011. To measure this efficiency, this study employs parametric model namely, stochastic frontier analysis (SFA) cost efficiency is used in a one-stage SFA model, which includes control variables (e.g. capital adequacy, asset quality and liquidity) and environmental variables (e.g. ownership, size, specialisation, deregulation periods and market structure) in the model specifications. The level of cost efficiency of Malaysian banks worsened over the years 2000–2011, with average cost efficiency during this period was at 76.5%. Despite the various liberalisation measures introduced to the banking industry – particularly during the three phases of the FSMP; 2000–2003; 2004–2007; 2008–2011 – cost efficiency trended downward, due to the effects of consolidation by domestic banks, deregulation of interest rates, the introduction of foreign Islamic banks, and the global credit crisis. Banks in Malaysia were forced to adjust their inputs and outputs to the rapid changes in the banking industry, which might have made a negative impact on cost efficiency.

Keywords

Financial Sector Master Plan; Cost Efficiency; Stochastic Frontier Analysis; Malaysian Banks

1. Introduction

Malaysia was hit by the Asian financial crisis in 1997–98. The Malaysian Ringgit fell by 40% against the US Dollar; the stock market plunged by over 70%, resulting in extreme volatility in financial markets and the country's sovereign rating being downgraded (Jomo and Chin, 2001). The scenario worsened as economic activity declined: GDP contracted by 7.5% with weak regional export demand; companies were in distress and unable to service debt and over leveraging. In the banking system, the number of non-performing loans (NPLs) increased sharply, which caused capital erosion due to over-concentration of risk (mainly in the large corporate sector). At the same time, the intermediation process was also inefficient due to tight liquidity and loan growth moderated sharply.

The FSMP initiatives changed the financial landscape of the banking industry. As of 2011, the banking industry was consolidated and rationalised, from 33 domestic financial institutions into eight banking groups (Abdul Majid et al., 2011). Banks were also found to be diversifying and improved their efficiency in delivery channels for financial products and services by enhancing access to financing, particularly for SMEs and consumers. These changes diversified the financial sector, with a deep and liquid debt securities market and a better focus on investment banks assisting corporations to get alternative finance in the bond market. Banks are now more focused on corporate governance and risk management, particularly with the implementation of principles-based regulations, coupled with an adequate supervisory and surveillance framework. Moreover, the market structure has improved, with an increased emphasis on market orientation, supported by greater regional cooperation, increased competitive pressure from new and current foreign banks, and freedom in the pricing of lending and deposits. Similar to initiatives in other developing countries, the objectives of these reforms and liberalisations, via the FSMP, are to promote diversity, efficiency and productivity, and to facilitate a competitive banking system by improving resource allocation and building a stronger economy (Fry, 1995). As a consequence, banking efficiency received even more attention in the aftermath of the financial crisis, with structural reforms and liberalisations, rendering the examination of this banking efficiency an important issue for both the public and policymakers alike (Berger and Mester, 1997).[1]

[1] Improved banking efficiency could result in better resource allocation, which benefits society by intermediating greater amounts of funds, providing more products with better prices and service quality for customers, improving bank profitability and achieving greater safety and soundness in banking sector (Berger and Mester, 1997). Therefore, the study of efficiency could assist banking regulators to design policies by evaluating the impacts of financial liberalisation, consolidation and market structure on efficiency.

The aims of this study are: first, to carry out a cost-efficiency analysis of Malaysian banks for the years 2000– 2011 using stochastic frontier approach and examining how changes in the financial services affected efficiency, productivity and the market structure of the banking industry in Malaysia. Second, to examine the impact of market liberalisation initiatives, via the FSMP, on efficiency and productivity in Malaysian banks.

2. Methodology

In this study, frontier measurement is employed to measure the efficiency of Malaysian banks for the years 2000– 2011. For better estimation of cost-efficiency, and taking into account the effect of heterogeneity (e.g. ownership structure, banks specialisation, inherent risks, and size), this study uses Battese and Coelli's (1995) one-stage approach, which may have an impact on the efficiencies. In this one-stage approach, a set of control variables (e.g. capital adequacy, asset quality and liquidity) and environmental variables (e.g. ownership, specialisation, financial liberalisation and size) are included into the specification of cost- and profit-efficiency functions. These different sets of control and environmental variables are tested in several stages using statistical testing (i.e. the log-likelihood ratio test), searching for the best fitting model that is later utilised for the estimation of efficiency scores in Malaysian banks.

The SFA model assumes that in producing a certain level of output, firms face various technical inefficiencies and a given combination of input levels. The firm's production is influenced by the sum of a parametric function of known inputs, with unknown parameters, and a random error (associated with the measurement error of the level of production and inefficiency). SFA requires a functional form, such as cost or profit, with a two-component error terms: random error and inefficiency. By way of illustration, the single-equation stochastic cost function model is shown below:

$$\square\square\square_{\square\square} = \square_{\square\square\square} \square\square\square_{\square\square} + \square\square\square_{\square\square} \quad (1)$$

where lnY_{it} is the natural logarithm of output for the i-th bank at time t, lnX_{it} is a vector of inputs of i-th bank at time t, β_{xit} is a vector of unknown parameters to be estimated and $ln\varepsilon_{it}$ is the error term. Following Aigner et al. (1977), the assumption of the composed error term is:

$$\varepsilon_{it} = V_{it} + U_{it} \quad (2)$$

where V_{it} and U_{it} are independently distributed; V_{it} represents random uncontrollable error and is assumed to be normally distributed with zero mean and variance σ_v^2 is drawn from a one-sided distribution that is assumed to

capture inefficiency. U_{it} is assumed to be drawn from a half-normal distribution with mean zero and variance σ_u^2. U_{it}, can be estimated by using the conditional mean of inefficiency term, given the composed error term, as proposed by Jondrow et al. (1982) and derive the log-likelihood for inefficiency, which is expressed in terms of the two variance parameters, $\sigma^2 = \sigma_V{}^2 + \sigma_U{}^2$ capturing the variance of the composed error term $\lambda = \sigma_u{}^2/\sigma_V{}^2$, which measures the fraction of inefficiency relative to statistical noise. Moreover, U_{it} can be used to measure the level of inefficiency of banks. For instance, if U_{it} is equal to 0, it indicates that there is no inefficiency based on the production function imposed. On the other hand, if U_{it} is more than 0, it indicates that inefficiency is present. In the past, most studies using SFA were directed towards inefficiency prediction and this inefficiency is commonly measured using technical efficiency (*TE*). Equation 3 exhibits the common output-orientated measure of *TE* using the ratio of observed output to corresponding frontier output, which can be written as (Coelli et al., 2005):

$$TE_{it} = \frac{y_{it}}{\exp(\beta_{x_{it}} + v_{it})} = \frac{\exp(\beta_{xit} + v_{it} - u_{it})}{\exp(\beta_{x_{it}} + v_{it})} = \exp(-u_{it}) \qquad (3)$$

where *TE* is technical efficiency of *i*-th bank at time *t*, y_{it} is the observed output and exp (β_{xi} + v_i) is the corresponding frontier output. As mentioned earlier, *TE* has a value between 0 and 1, in which *TE* derives from the output of *i*-th bank relative to a fully-efficient bank's output, located on the estimated frontier curve that utilises the same input vector (Coelli et al, 2005).

Cost-efficiency indicates how close a bank's cost is to that of a best-practice bank, which produces the same outputs using the same technology. The variable costs of the cost function rely on: the prices of variable inputs, the amount of variable outputs, fixed netputs (or exogenous factors) (if any), random errors, and inefficiency (Berger and Mester 1997). The cost-efficiency of a bank is measured using the observed bank's total cost, relative to the total cost of a bank on the estimated frontier. Hence, the cost function is described as:

$$\ln \square\square_{\square\square} = \square(\square_{\square\square}, \square_{\square\square}, \square_{\square\square}, \square_{\square\square}, \square) + \ln \square_{\square\square} + \ln \square_{\square\square} \qquad (4)$$

where ln TC_{it} is the natural logarithm of the observed bank's total cost, $f()$ is the cost frontier's functional form, w_{it} is the vector of input prices, y_{it} is the vector of outputs, c_{it} denotes the vector of control variables (if any) and z_{it} represents the vector of environmental variables (if any). These control c_{it} and environmental z_{it} variables are included in the cost function to capture the heterogeneity effects of cost-efficiency. β is the parameters to be estimated. The term ln v_{it} + $ln\ u_{it}$ is treated as a composite error term, where ln u_{it} is the

inefficiency term: a non-negative and one-sided error component that follows an asymmetric half-normal distribution. ln v_{it} is the random error term that permits the random variation of the frontier across banks and captures the effects of measurement error, other statistical noise and random shocks outside the bank's control. This error term is assumed to consist of independently and identically distributed normal random variables, with zero mean and variance (Coelli et al., 2005).

The cost-efficiency of the *i*-th bank is the estimated cost needed to produce bank *i*'s output vector if the bank were as efficient as the best-practice bank (on the frontier curve) in the sample facing the same inputs and outputs, control, and environmental variables(w, y, c, z), divided by the actual cost of *i*-th bank, and adjusted for random error. It can be written as:

$$\exp(\square(\square_{\square\square}, \square_{\square\square}, \square_{\square\square}, \square_{\square\square}, \square) + \square_{\square\square} \qquad 1 \qquad (5)$$

$$CE_{it} = \frac{}{\exp(f(w_{it}, y_{it}, c_{it}, z_{it}, \beta) + v_{it} + u_{it}} = \frac{}{\exp(u_{it})}$$

where CE_{it} is the cost-efficiency of *i*-th bank. The numerator in equation 5 indicates the minimum cost that can be incurred by the best practice banks and the denominator in equation 5 denotes the actual cost incurred by *i*-th bank at time *t*. Hence, cost-efficiency CE_{it} is measured against the ratio of minimum cost banks (best-practice banks on the frontier) and the actual cost of *i*-th bank. Cost-efficiency CE_{it} could also be seen as a proportion of cost that is either being used efficiently or being wasted. For example, if CE_{it} of *i*-th bank is 0.60, it indicates that *i*-th bank is 60.0% efficient and 40.0% of its cost is being wasted when compared to the best-practice bank. Costefficiency ranges between 0 and 1. Banks with a cost-efficiency of 1 are considered to be best-practice banks within the observed data.

3. Result

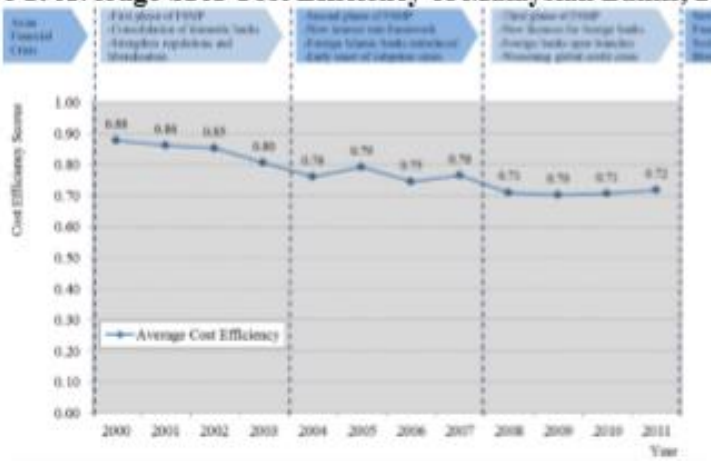
The cost efficiency scores for the Malaysian banks for the years 2000–2011 are shown in Table 1. The average cost efficiency score for Malaysian banks between 2000 and 2011 is 82.7%. The average score of cost efficiency was at 85.1% in 2000 and ended with 75.0% in 2011. The average cost efficiency scores suggest that Malaysian banks wasted around 20.0% of their input to produce the same level of outputs of the best performing banks. This finding (approximately 20% inefficiency) is consistent with past findings in the literature, where SFA was performed for the cost efficiency function (e.g De Young, 1997; Berger and De Young, 2001; Bonin et al., 2005).

Table 1: Average SFA Cost Efficiency Scores, 2000–2011

Year	Count	Mean	Standard Deviation	Minimum	Maximum
2000	24	0.8760	0.1519	0.2600	0.9721
2001	26	0.8614	0.1537	0.2698	0.9787
2002	24	0.8523	0.1524	0.3253	0.9725
2003	24	0.8103	0.1528	0.3731	0.9582
2004	25	0.7596	0.1734	0.2404	0.9270
2005	26	0.7913	0.1554	0.2088	0.9418
2006	29	0.7450	0.2095	0.2033	0.9317
2007	31	0.7648	0.1440	0.2704	0.9402
2008	36	0.7098	0.1783	0.0509	0.9324
2009	37	0.7025	0.1877	0.0296	0.9271
2010	37	0.7065	0.1884	0.0738	0.9478
2011	35	0.7177	0.2157	0.0848	0.9702
2000-2003	98	0.8502	0.1523	0.2600	0.9787
2004-2007	111	0.7647	0.1707	0.2033	0.9418
2008-2011	145	0.7090	0.1909	0.0296	0.9702
2000-2011	354	0.7655	0.1834	0.0296	0.9787

From Figure 1, based on the three different phases of the FSMP for the years 2000–2003, 2004–2007 and 20082011, the cost efficiency average scores were 85.0%, 76.5% and 70.9% respectively. From the result, the overall trend of the cost efficiency for the years 2000 to 2011 was on a declining trend. From previous literature, many found that deregulation of the banking sector resulted in greater efficiency (e.g Tortosa-Austina, 2003; Cuesta and Orea, 2002). On the other hand, deregulation of the banking sector could also result in lower efficiency as evidenced from past studies. For instance, Berger and Humphrey (1992) found that financial liberalisation forces banks to cut costs substantially in a very short period of time. However, this has not been the case as the banks displayed a higher level of cost inefficiency and a slow response in adjusting to changes to minimise their costs); and, based on observations by Humphrey and Pulley (1997), they suggested that banks' adjustments following deregulation can take up to four years to complete. Additionally, Girardone et al. (2004) state that banks may not be efficient when deregulatory initiatives take place at the same time as a macroeconomic downturn. During this period, Malaysian banks exhibited lower cost efficiency during the third phase of the FSMP as a result of the global credit crisis.

Figure 1: Average SFA Cost Efficiency of Malaysian Banks, 2000–2011



The first phase of the FSMP (2000–2003) is regarded as the period of initial reform in the banking industry. The banking sector in particular witnessed the emergence of large domestic banks from a guided consolidation exercise. During this period, cost efficiency scores were trending downward. This can be potentially explained by various recovery measures taken by the government of Malaysia and BNM. After the financial crisis in 1997–98, the Malaysian government took various drastic measures to improve the banking sector. Against the backdrop of that crisis, there were significant structural changes in the Malaysian banking sector (Sufian, 2004). With the severe losses faced by the Malaysian banks, and to maintain the integrity of public savings and the stability of financial system, the Malaysian government introduced a rescue scheme to acquire shares in some of the ailing commercial banks and absorb problem assets in distressed banks. Malaysia did not rely on assistance from the International Monetary Fund (IMF) after the financial crisis, unlike some other Association of South East Asia Nations (ASEAN) member countries. Under the IMF programme, insolvent banks were forced to close down, but Malaysia did not take this path as the social cost involved, in terms of dislocation of resources, would have been high. Malaysia took a different approach by introducing a guided consolidation of fragmented banking institutions, in which BNM played an intermediary role, solving issues of fairness to all parties involved in the merger.[2] The effect of consolidation of domestic banks resulted in declining cost efficiency because they were forced

[2] This consolidation programme was also in line with the requirement in having stronger domestic banks to compete regionally when opening its financial industry to the international players in 2003 under the World Trade Organisation (WTO). As a result, in 2001, the consolidation had successfully merged 54 Malaysian banks and financial institutions into 10 anchor banking groups.

to implement various rationalisation programmes including: restructuring of duplicated branch networks, managing staff redundancy, synchronising technology with the acquiring partner and implementation of internet banking services had resulted in decreasing cost efficiency during the first phase of the FSMP (Sufian, 2004).

During the FSMP's second phase (2004–2007), BNM introduced a new interest rate framework (NIRF) that aimed to facilitate more efficient pricing of financial products. Following the removal of BNM's intervention rate, the banking institutions were given the flexibility to determine their BLR based on their own cost structure and lending strategies. The deregulation of interest rates was intended to increase efficiency, productivity, innovation and profitability in the banking system (Leightner and Lovell, 1998; Berger and Mester, 2003). As a result of the deregulation of interest rates, banks were forced to adjust their inputs and outputs to remain competitive (Hao et al., 2001).[3] With the introduction of NIRF, Malaysian banks were able to price their funding costs and revenues based on their own cost structure and compete for customers using their own interest pricing structure.[4] There was a slight increase in cost efficiency scores in 2005, following the liberalisation of BLR, indicating some increasing level of competition in terms of pricing among Malaysian banks. Towards the end of the second phase of the FSMP, the cost efficiency scores dropped marginally due to a significant loss faced by full-fledged Islamic banks and the inception of new foreign Islamic banks. The overall average cost efficiency scores worsened because these new or de-novo foreign Islamic banks inherently faced higher operational costs during their early phase of operations.

In the third phase of the FSMP (2008–2011), the Malaysian banks were not affected at the initial stage of the subprime crisis in the US. Malaysian banks were prudent in their investments, particularly relating to derivatives products originated in the US and Europe: only a small portion of these instruments were held by them. However, in 2008, as the global economy deteriorated, demand for Malaysian exports declined, which affected the real sector. Malaysian GDP contracted by 1.7% in 2009. The NPLs of banks increased slightly in 2009, reflecting the contraction experienced by the economy.

[3] In terms of adjustments made to inputs and outputs, Humphrey and Pulley (1997) found that during the deregulation of interest rate, banks tend to respond in three ways. First, to offset higher deposit interest cost with higher explicit and implicit for small deposits. Second, to transfer the higher funding cost to borrowers. And third, to invest in risky assets to obtain higher yield.

[4] The interest rate deregulation program generally increases competitive pressure in the market and forced banks to reduce their cost (Mester, 1993).

Therefore, the declining trend of cost efficiency scores is driven by greater operating costs when managing excess liquidity from large inflows of foreign funds and placing greater resources into managing potentially delinquent loans. At the same time, BNM reduced its policy interest rate (OPR) from 3.5% to 2.0% in November 2008 to February 2009. The policy interest rates required banks to drastically adjust their input prices and outputs according to the indicative market interest rate (i.e., the OPR). Similar to Berger and Humphrey (1992), a slower response in adjusting to changes resulted in Malaysian banks experiencing lower cost efficiencies during the third phase of the FSMP. In addition, Basel II was also implemented between 2008 and 2010 (Standard Approach and Internal Ratings Based Approach) and Malaysian banks invested heavily in technology, physical assets, external consultants and specialised labour to comply with the new capital regulation.

4. Discussion and Conclusion

Before the Asian crisis, banks were found to lack effective risk management and corporate governance, resulting in a high level of fragility. Furthermore, the market conditions were very rigid, where prescriptive rules-based regulation and supervision was implemented in the financial sector. The pricing mechanism of banking products and services was also rigid, which did not encourage competition among financial players. Islamic finance was limited and not given full attention as to its possibilities and potentials. These conditions made Malaysian banks more vulnerable to macroeconomic distress and the inability to withstand these pressures during the Asian financial crisis. A major response by the Malaysian government to these pressures has been a substantial consolidation measure, resulting in a reduction in the total number of banks, in which a fragmented banking system of 33 domestic banking groups was reduced to 10 anchor domestic banking groups. At the same time, the FSMP was introduced to strengthen and further liberalise the Malaysian banking industry following the Asian financial crisis in 1997–98.

The implementation of the FSMP changed the financial landscape of the banking industry. Within the FSMP period, Malaysia witnessed a series of financial liberalisation measures, including; liberalisation of interest rate to market players, introduction of new foreign banks (both Islamic and conventional banks), branch liberalisation by allowing foreign banks to increase their branches, de-pegging of Malaysian Ringgit to US Dollars, simplified product approval process, lifting of wage moratorium, and allowing of outsourcing of banks' non-core activities. With these initiatives, excessive government intervention in banks' operations (particularly in relation to interest rates (e.g. BLR)) was reduced. One of the key objectives of the FSMP was to improve competition, which at the same time promotes the banking industry's resilience and soundness.

This study found that the cost inefficiencies in Malaysian banks are substantial, inefficient banks being approximately 20.0% less cost-efficient when compared to best-practice banks. Therefore, in order to produce the same level of outputs of the best-practice banks, these inefficient banks should improve their cost by approximately 20.0% respectively. Despite various initiatives introduced to improve the degree of competition in the market (e.g. liberalising controlled interest rates regime, allowing foreign banks to increase branches) and reduce market concentration (e.g. introducing new foreign banks), these measures have yet to show any improvements due to their nascent or growing stages of implementation, particularly during the post-consolidation period of domestic banks. A small number of large domestic banks could lead to collusive strategies, anticompetitive behaviour; and hence, can result in greater risks towards public welfare. Furthermore, market power may lead to lower efficiency in large banks, with managers enjoying the 'quiet life', and earning higher interest rates on loans and deposits. Therefore, regulators ought to accelerate their liberalisation initiatives to ensure adequate competitive pressures on large domestic banks. Probably, greater participation of foreign ownership through equity can be exploited as a catalyst for more efficiency.

References

1. Abdul Majid, M., Saal, D.S. & Battisti, G. 2011, "The impact of Islamic banking on the cost efficiency and productivity change of Malaysian commercial banks", Applied Economics, vol. 43, no. 16, pp. 2033-2054.
2. Aigner, D., Lovell, C.A.K. & Schmidt, P. 1977, "Formulation and estimation of stochastic frontier production function models", Journal of Econometrics, vol. 6, no. 1, pp. 21-37.
3. Battese, G.E. & Coelli, T.J. 1995, "A model for technical inefficiency effects in a stochastic frontier production function for panel data", Empirical Economics, vol. 20, no. 2, pp. 325-332.
4. Berger, A.N. & De Young, R. 1997, "Problem loans and cost efficiency in commercial banks", Journal of Banking & Finance, vol. 21, no. 6, pp. 849-870.
5. Berger, A.N. & Humphrey, D.B. 1992, "Measurement and Efficiency Issues in Commercial Banking" in Output Measurement in the Service Sectors, ed. Z. Griliches, National Bureau of Economic Research Vol. 56 edition, University of Chicago Press, Chicago, pp. 245-300
6. Berger, A.N. & Mester, L.J. 1997, "Inside the black box: What explains differences in the efficiencies of financial institutions?", Journal of Banking & Finance, vol. 21, no. 7, pp. 895-947.
7. Berger, A.N. & Mester, L.J. 2003, "Explaining the dramatic changes in performance of US banks: technological change, deregulation, and

dynamic changes in competition", Journal of Financial Intermediation, vol. 12, no. 1, pp. 57-95.
8. Bonin, J.P., Hasan, I. & Wachtel, P. 2005, "Bank performance, efficiency and ownership in transition countries", Journal of Banking & Finance, vol. 29, no. 1, pp. 31-53.
9. Coelli, T.J., Rao, D.S.P., O'Donnell, C.J. & Battese, G.E. 2005, "An Introduction to Efficiency and Productivity Analysis", Springer, New York.
10. Cuesta, R.A. & Orea, L. 2002, "Mergers and technical efficiency in Spanish savings banks: A stochastic distance function approach", Journal of Banking & Finance, vol. 26, no. 12, pp. 2231-2247.
11. DeYoung, R. 1997, "A diagnostic test for the distribution-free efficiency estimator: An example using U.S. commercial bank data", European Journal of Operational Research, vol. 98, no. 2, pp. 243-249.
12. Fry, M.J. 1995, Money, Interest and Banking in Economic Development, 2nd edn, John Hopkins University Press, London.
13. Girardone, C., Molyneux, P. & Gardener, E.P. 2004, "Analysing the determinants of bank efficiency: the case of Italian banks", Applied Economics, vol. 36, no. 3, pp. 215-227.
14. Hao, J., Hunter, W.C. & Yang, W.K. 2001, "Deregulation and efficiency: the case of private Korean banks", Journal of economics and business, vol. 53, no. 2-3, pp. 237-254.
15. Humphrey, D.B. & Pulley, L.B. 1997, "Banks' Responses to Deregulation: Profits, Technology, and Efficiency", Journal of Money, Credit and Banking, vol. 29, no. 1, pp. pp. 73-93.
16. Jomo, K.S. & Chin, K.F. 2001, "Financial reform and crisis in Malaysia", Financial big bang in Asia, edited by Masayoshi Tsurumi, pp. 225-250
17. Jondrow, J., Knox Lovell, C.A., Materov, I.S. & Schmidt, P. 1982, "On the estimation of technical inefficiency in the stochastic frontier production function model", Journal of Econometrics, vol. 19, no. 2-3, pp. 233-238.
18. Leightner, J.E. & Lovell, C.A.K. 1998, "The Impact of Financial Liberalization on the Performance of Thai Banks", Journal of economics and business, vol. 50, no. 2, pp. 115-131.
19. Sufian, F. 2004, "The Efficiency Effect of Bank Mergers and Acquisition in Developing Economy: Evidence from Malaysia", International Journal of Applied Econometrics and Quantitative Studies, vol. 1, no. 4, pp. 53-74.
20. Tortosa-Ausina, E. 2003, "Nontraditional activities and bank efficiency revisited: a distributional analysis for Spanish financial institutions", Journal of economics and business, vol. 55, no. 4, pp. 371-395.

Microdata dissemination at DOSM: Challenges, potential and implementation

Azrin Ahmad, Rosnah Muhamad Ali, Tuan Noraida Tuan Hamzah, Siti Haslinda Mohd Din
Department of Statistics Malaysia

Abstract

The dynamic data behaviour of microdata dissemination has inflicted Department of Statistics Malaysia (DOSM) as data producer facing the expanding demand for microdata. Determining the best way to disseminate these data is a challenge for DOSM. According to a study by Li on the Microdata Dissemination, microdata often remain inaccessible to the community, due to technical, financial, legal even political obstacles. Thus, the main objective of this paper is to assess the best option in accessing microdata while protecting its confidentiality. As the national data producers, DOSM have to implement procedures for the documentation, cataloguing and dissemination of the data. This is in line with the aim of data dissemination in maximizing the usage of data and to reassure the data being used in an optimum manner. Referring to the Committee for the Coordination of Statistical Activities Report on Microdata Dissemination Best Practices, the establishment of the core principles of micro data procedures and dissemination should include openness, transparency, legal conformity and protection of privacy, protection of intellectual property, interoperability, quality, security and accountability. Apart from this, the limitation of the developed policies and procedures will also be discovered through the identification of the best practices of data dissemination. Subsequently, the microdata confidentiality and disclosure protection policy developed will impose the balance between demand and the need to keep respondent information confidential. At the same time, access to microdata by the research community would foster diversity and quality of the analyses. This will broaden the use of existing data, and increase the return on data collection investment indirectly may enhance the data quality of DOSM. Moreover, it will lead to a better measurement and impact towards the value and respond to the demand of data. At the end of the study, we are expecting to summarize the best practices in developing policies and methodology on the access to microdata in DOSM.

Keywords

Microdata; dissemination; confidentiality; policy

1. Introduction

In 1949, the Department of Statistics, Malaysia (DOSM) was established under the Statistics Ordinance 1949, which then be known as Bureau of Statistics. Statistics Ordinance was then revoked by Statistics Act 1965 that has revamped The Bureau of Statistics into Department of Statistics, led by the Chief Statistician. Statistics Act 1965 was more thorough and has strengthened the authority for DOSM in collecting data. Department of Statistics, Malaysia acts as the main official statistical agency that is responsible for the country's official statistics in collecting, processing, interpreting and disseminating data. In fulfilling the user's demand for the data, DOSM is facing its own challenges in disseminating the acquired data in a way that the confidentiality of the data maintained safeguarded. This problem arised not only in DOSM, but also at global such as in China, Sri Lanka, Indonesia and others. In protecting the data confidentiality and to meet user's demands for microdata, agencies and researchers have developed an array of Statistical Disclosure Limitation (SDL) strategies (Duncan, de Wolf, Jabine and Straf, 1993).

Through time, the demand for microdata is growing and becoming more diverse. Microdata referred to records that were collected as input data for the surveys. These records are confidential subjected to law restrictions and ethical standards set in place. However, under special conditions, this microdata could be disseminated to special groups or users. These users may include various type of people such as government officials, academic researchers, policymakers, and the general public. Data may be disseminated publicly without any restrictions or specifically to certain users under specific conditions. The availability of microdata is often dependent on national laws and regulations. Special consideration were required in making data and documentation files available to users. More is involved in the dissemination procedure than merely providing data access to the users. In disseminating the microdata, data provider shall assure the users that the provided data is trustworthy, fully documented, has no confidentiality concerns, and is securely preserved for future use. An additional aspect of dissemination is the method or alternative way to share research findings with those interested parties. It is vital to consider who is using the data and the purpose of the data being used as part of a comprehensive dissemination strategy. These objectives were not embraced not only by DOSM, but also by many international organizations, social science data archives, and survey research projects.

DOSM aims to maximize the data usage by disseminating data and to reassure the data being used in an optimum manner. If it is regarding aggregated data, there is no critical issue on the dissemination part. How about microdata?

Since it is inaccessible to the community, there is possibility on misinterpreting the released statistics by the media and may lead to

misleading implications to the community. This will jeopardize DOSM's image as the main national data producer. Hence, allowing access on the microdata will avoid the possibility of misinterpreting the aggregated data or the released statistics if the research community makes use of the data and come out with a high quality analysis and various statistical reports to support the official statistics.

2. Current Practice at DOSM

The demand or data request is increasing every single day. DOSM disseminate the microdata based on the demand and has been disseminating the microdata via various platforms over the years. There are a lot of benefit of the microdata dissemination such as reducing the duplication of data collection activities and it will broaden the use of existing data. However data dissemination is also entails risks and challenges.

The main challenge is disclosure risk. As a data disseminator, it is DOSM priority to prevent any disclosures of identities or sensitive information to the public. If this happen, DOSM may lose the trust from the public and the most serious consequence is violation of the statistics act.

2.1 Microdata in DOSM

Microdata is the unpublished data and to obtain the microdata, public or user have two options. The first one, public or user should request and apply to DOSM via e-Statistics. Secondly, if the user interested to have the feel and look of the microdata, they may come to DOSM headquartes to access it via one of our data dissemination sub-product; StatsDW MyLab. StatsDW is DOSM datawarehouse and there are 3 external modules and 2 internal modules. StatsDW MyLab is the internal module and only can be access in DOSM. Currently we have 215 of datasets in our StatsDW comprises of 172 Economic datasets, 31 Social datasets and 12 Compilation datasets.

2.2 Core Principles of Microdata Dissemination

The National Statistics Offices of other countries have their own guideline and principles in the dissemination of microdata. Each NSO has different approach and practice in establishing the key principles it should comply to specifically in ensuring the safety and security of microdata. As for the Organisation for Economic Co-operation and Development (OECD), the core principles defined in the microdata dissemination are as follows:

#	Key principles	Details	DOSM's practice
1	Openness	• Not referring to unrestricted access or open data • Data provided at minimal cost	• Microdata disseminate based on request and objective of the data usage • Minimal administrative charge
2	Transparency	• Metadata to be provided • Specification of conditions of the use of the data should be made available	• Microdata metadata provided is incomprehensible • Conditions of the use of the data is made available at the website
3	Legal conformity and protection of privacy	• Protection of privacy and practices	• Microdata disseminated bound to the Statistical Act 1965 (revised 1989)
4	Protection of intellectual property	• Applicability of copyright • Ownership of data	• Copyright published on the website • Microdata wholly owned by DOSM
5	Interoperability	• Use of international standard of data documentation, particularly in access arrangement	• Microdata access practiced by DOSM is compliant to the international standard
6	Quality	• Compliance to the process and methods of data quality assurance	• DOSM is in the process of establishing the quality assurance framework
7	Security	• Integrity and security of the data	• Integrity and security of the data is guaranteed in accordance to the DOSM Microdata Dissemination Policy
8	Accountability	• Periodic evaluation of the data access arrangement to be conducted	• DOSM conducted pre, on-going and post process evaluation

Based on the table above, apparently, the practice of microdata dissemination practiced by DOSM is compliant to the key principles defined by the OECD.

2.3 Statistical Disclosure Limitation (SDL)

Nowadays, the emerging of new technologies in the world of data dissemination and access has put the data producer and disseminators in a difficult position. They are pressured by users who required everything about the data but at the same time presurred by the limit and restriction on what to be released. The conflict between the accuracy of the dissemination and the risk of disclosing respondent information also have to be put into consideration and overcomed with the most appropriate disclosure procedure or implementation. The key aspect also is how to trade off these two elements whilst guaranteeing the users requirement are being fulfilled. Statistical disclosure limitation divides into strategies based on restricted data and those based on restricted access.

Restricted data SDL strategies referred to masking and modifiying the data in ways that limit potential for disclosure. The modifications includes simple thing such as removing variables and records. However, in most cases, this is not enough which required a more complex alterarion such as swapping (Dalenius and Reiss, 1982; Gomatam, Karr and Sanil, 2003), adding random noise to units' values (Fuller, 1993), microaggregation, and other forms of data pertubation (Gomatam et al., 2016). For example, the first step in preventing identity disclosure is by removing explicit identifiers such as HIV status, address, identification card number, as well as implicit identifiers, such as "Occupation = Chief Statistician of Malaysia." Another example of this strategy is to protect units with high incomes, income is frequently "top coded," so that one category is "More than $X." Resctricted data SDl strategies can be applied with varying intensity. Generally, the higher the SDL intensity, the greater the protection against disclosure risk, but the less the utility of the released data. At least implicitly, agencies choose SDL strategies by balancing confidentiality protection and utility of the released information.

While for the restricted access SDL strategies, the mechanisms include data centers, licensing, and vetting of researchers and their research plan. This strategy allows users to perform analyses directly on the underlying data. The specific analyses may be suppressed, if the analysis is known to threaten confidentiality, or a posteriori, the output reveals a threat. According to the confidential level, there are four types of files of dissemination: public use files, licensed files; data enclave; remote data access.These centers rely on the honesty of researcher to protect confidentiality, and can be expensive for the agencies and inconvenient for researcher.

3. Conclusions

Basically, DOSM has implemented the core principles of microdata dissemination accordingly. However, it will need to strengthen the metadata dissemination of which it should be more comprehensive and can assist users to understand better the data provided and produce an accurate output from the analyses conducted. The metadata provided should explain in details the definitions, concepts, coverage, variables description, history of the data, statistical methods used, and etc. However, in terms of the implementation of statistical disclosure, DOSM is preferable to widen the microdata dissemination by giving access to users via remote data access. Currently, DOSM offers a microdata access via a paltform namely StatsDW MyLab to users. However, users are only allowed to access the platform at the DOSM premise. In order for DOSM to open the access via a remote data access, DOSM must put security as the highest aspects to focused on. Apart from this, DOSM might have to empower and strengthen the Microdata Dissemination Policy especially in defining the variables to be disseminate via the remote data access platform. In maintaining the microdata confidentiality, it is advisable to apply the SDL strategies accordingly which will require continued and dedicated effort or even budgetary planning and funding. As a result, DOSM will be able to fullfill the user's needs, at the same time, the information confidentiality remained safeguarded.

References

1. Committee for the Coordination of Statistical Activities (2014). Microdata Dissemination Best Practices.
2. Dalenius, T. and REISS, S. P. (1982). Data-swapping: A tech nique for disclosure control. /. Statist. Plann. Inference 6 73-85.
3. Duncan, G. T., de Wolf, V. A., Jabine, T. B. and Straf, M. L. (1993). Report of the panel on confidentiality and data access. J. Official Statistics 9 271-274.
4. Fuller, W. A. (1993). Masking procedures for microdata disclo sure limitation. /. Official Statistics 9 383-406.
5. Gomatam, S., Karr, A. F., Reiter, J. P., & Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. Statistical Science, 163-177.
6. Gomatam, S., Karr, A. F. and Sanil, A. P. (2003). Data swap ping as a decision problem. Unpublished manuscript.
7. Granda, P., & Blasczyk, E. (2010). Data Dissemination. Cross-cultural survey guidelines.
8. Li Li (2006). Microdata Dissemination: What shouldn't be ignored to improve the statistical capacity of China.

Weighting Longitudinal School Surveys with population changes: The case of Geres

Gabrielle Palermo
University of Southampton (UK)

Abstract

The Longitudinal Study of the 2005 School Generation (Geres) is the first survey of its kind successfully achieved and remain the only one in Brazil. It aimed to observe pupils' achievements in mathematics and reading, and changes in individual performance and school characteristics over junior school years, from 2005 to 2008 in five cities: Rio de Janeiro, Belo Horizonte, Campinas, Campo Grande and Salvador. All pupils that were registered in the 1st grade of the selected schools made up the Geres sample for the first wave. They were followed by five waves, two in 2005 and one every year from 2006 to 2008. Also, the survey tracked additional pupils that joined the main grade currently in each wave in the selected schools. Each city represents one stratum, and they were divided into three to four strata according to the schools' administration system: state, municipal, private and exceptional schools. The survey had 17 strata in total, and they were called explicit strata. The published weights include weights for the explicit strata at wave 1. Furthermore, the survey report (Brooke and Bonamino, 2011) explained the implicit stratification, where each explicit stratum was divided by up to eight groups, according to the school size and socio-economic levels. This report defined weights for the first wave only, for schools and pupils separately, though incomplete for pupils. There was not a subsample into the selected schools, then the inclusion probability of the selected pupils was the same as their school. The municipal schools of Rio de Janeiro were chosen to be studied. Following the weights' report of the survey, we computed the weights for all pupils from the first wave to estimate the total of pupils and compared the same estimation but using different estimators, considering only the schools' weights for explicit and implicit strata. We proposed to compute the total of pupils that repeated one or more grades considering this observation at wave 4. Different cross-sectional weights were proposed, according to the population of interested. Longitudinal weights for pupils observed at wave 4, considering some of the combinations of the previous waves were also estimated.

Keywords

Educational survey; Mobility; Weight share method; Cross-sectional estimates.

1. Introduction

Geres was designed to follow the imitial sampled pupils along five waves, from 2005 to 2008, and the new pupils that joined the classes of original pupils, since they continued studying at a sampled school. The surveys additions could be from the same initial population or a different population rather than the one at wave 1. Additional pupils from the same population can have their selection probability computed if the sampling frame is available. However, it would be necessary to add all students studying with them in 2005 (1° grade), and track them over the waves. As a consequence, the $\pi_{hk}^{(1)}$ would change for all selected schools, and the sample representation for the study population could fail.

On the other hand, some of the new students could be studying in a grade above, and repeat the year at some point, joining a surveyed classes. Likewise, an additional student could be studying in another place that is not part of the area covered by the survey. Thus, pupils that are from a different population at wave 1 do not have an inclusion probability.

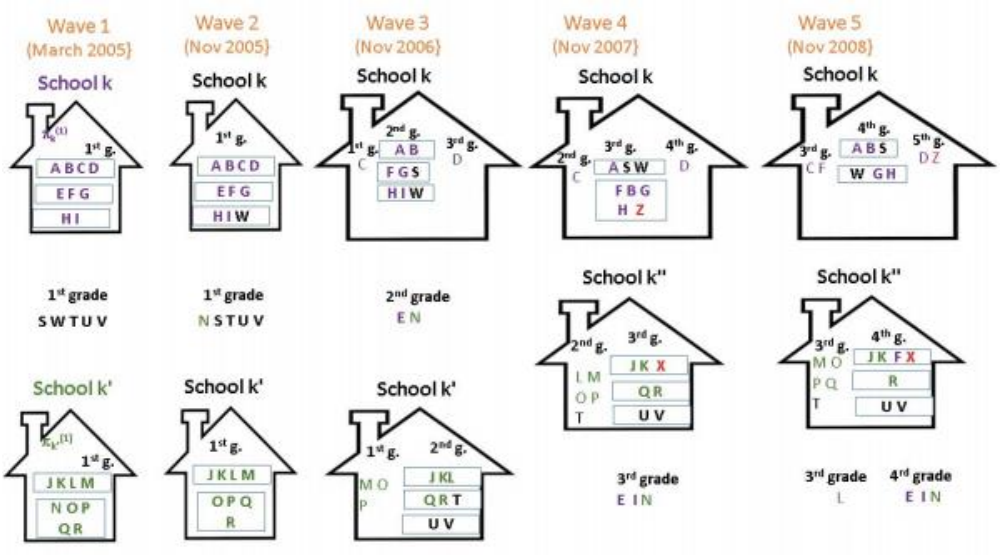
Indirect sampling is crucial when one wants to study rare populations that are difficult to identify or populations of interested that are not listed; since a probability sample cannot be selected directly. The idea of indirect sampling happens when a sample from a given target population B can be reached through links with another sample of a population A, then, these two populations are somehow connected. Indirect sampling is a form of probability sampling because it is based on a sampling frame (Lavallée, 2007), and it started to be used in cross-sectional surveys to make inference about the population of interest.

The figure 1 shows the different ways that pupils could be included or followed by Geres. The pupils that are outside school were not selected then. And pupils could repeat or skip a grade more than once, but the figure only shows pupils that were attending up to one year less or more than the main grade, in other to make it simpler.

The main grade is the one initial pupils are expected to be studying at a given wave. The first and the second waves happened at the same academic year, and the further ones after one year. Thus, the main grade for each wave is:

Wave 1: 1st grade;
Wave 2: 1st grade;
Wave 3: 2nd grade;
Wave 4: 3th grade;
Wave 5: 4th grade.

Figure 1: Summary of pupils moviments over time – Geres



Once observed, the additional pupils became members of the survey. From wave 3 onward, the survey members that repeated the grade and continued studying at a sampled school were tracked either, though their classmates were not added to the survey. The members that moved to another school which was not in the initial sample were not followed. Except for the case of Rio de Janeiro, where 600 Geres pupils[1] were transferred to non-sampled schools in 2007, due to a local administrative decision (Brooke and Bonamino, 2011), so they were followed, and their new class peers were surveyed as well.

2. Methodology

Generalised Weight Share Method (GWSM) presented by Lavall´ee (1995, 2007) can be applied for either longitudinal and cross-sectional surveys. Longitudinal surveys have their sample designed for the first wave. Some of these surveys add new units to the original sample, and a few of them can be from another population than the initial one, or it is quite complex to compute their initial inclusion probability, considering the initial survey design.

A similar problem happens with some of the cross-sectional surveys, they do not have a frame list to select a probabilistic sample, but their correspondent target population has links with another framed population. Thus, the subsequent waves population or unframed cross-sectional population could be sampled through an indirect sampling when they are somehow linked with a weighted sample.

[1] Pupils from municipal schools.

The GWSM can be applied at Geres's data to estimate weights for all waves. The links between the base wave b and a given next wave t could be represented by $l_{i,kj}$, where pupil $i \in U^{(b)}$ and pupils j of school $k \in U_k^{(t)}$. The GWSM steps are:

Stage 1: Compute the initial weight w'_k, given school $k \in \Omega^{(t)}$.
And $\Omega^{(t)} = \{j \in U^{(t)} \mid \exists i \in s^b \text{ and } L_{i,k} > 0\}$:

$$w'_k = \sum_{j=1}^{M_k^{(t)}} \sum_{i=1}^{M^{(b)}} l_{i,kj} \times t_i \times w_i \tag{0.1}$$

Where:

$t_i = 1$ if $i \in s^{(b)}$, and 0 otherwise. $l_{i,kj} = 1$ if pupil $i \in U^{(b)}$ corresponds to pupil $j \in U_k^{(t)}$, and 0 otherwise.

Stage 2: Sum of the links into each school $k \in \Omega^{(t)}$:

$$L_k^{(t)} = \sum_{j=1}^{M_k^{(t)}} \sum_{i=1}^{M^{(b)}} l_{i,kj} \tag{0.2}$$

Stage 3: The final weight $w_k^{(t)}$ is:

$$w_k^{(t)} = \frac{w'_k}{L_k} \tag{0.3}$$

Where: $w_{kj}^{(t)} = w_k^{(t)}$ for all pupil $j \in U_k^{(t)}$.

Therefore, the estimated total of pupils $Y^{(t)}$ with a given characteristics is:

$$\begin{aligned} \hat{Y}^{(t)} &= \sum_{k=1}^{n^{(t)}} \sum_{j=1}^{M_k^{(t)}} w_{kj} y_{kj} \\ &= \sum_{k=1}^{n^{(t)}} w_k^{(t)} Y_k^{(t)} \end{aligned} \tag{0.4}$$

3. Result

The estimations weights for wave 4 were computed for schools, having the population $U1$ as the base. The same calculations were done, but now having as base the estimation weights of wave 3.

The figures 2 and 3 show the first results. The sampling weights are quite different from the estimation weights.

Figure 2: Estimation weights for wave 4 - using GWSM from estimation weights of wave 1, - RJ,municipal schools

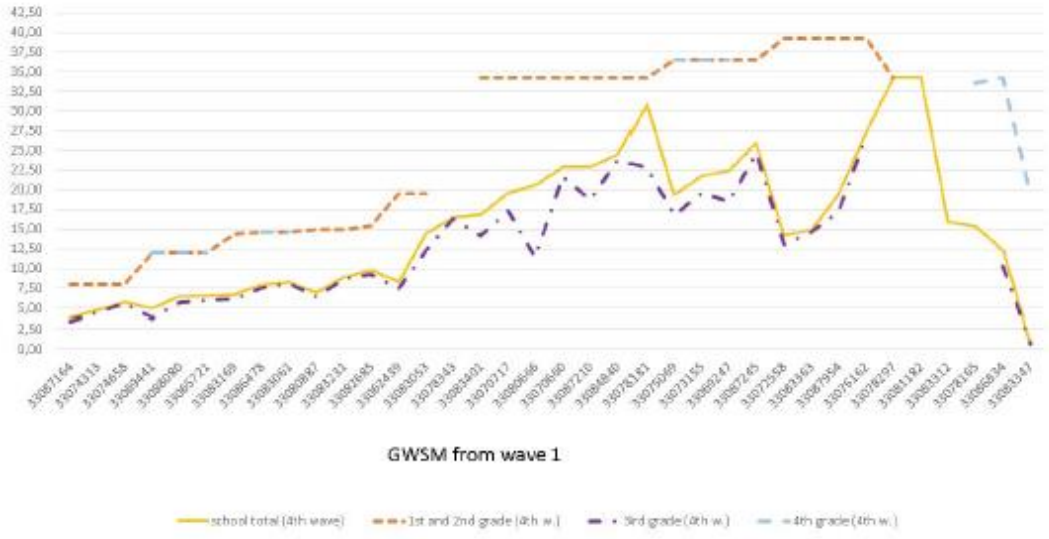
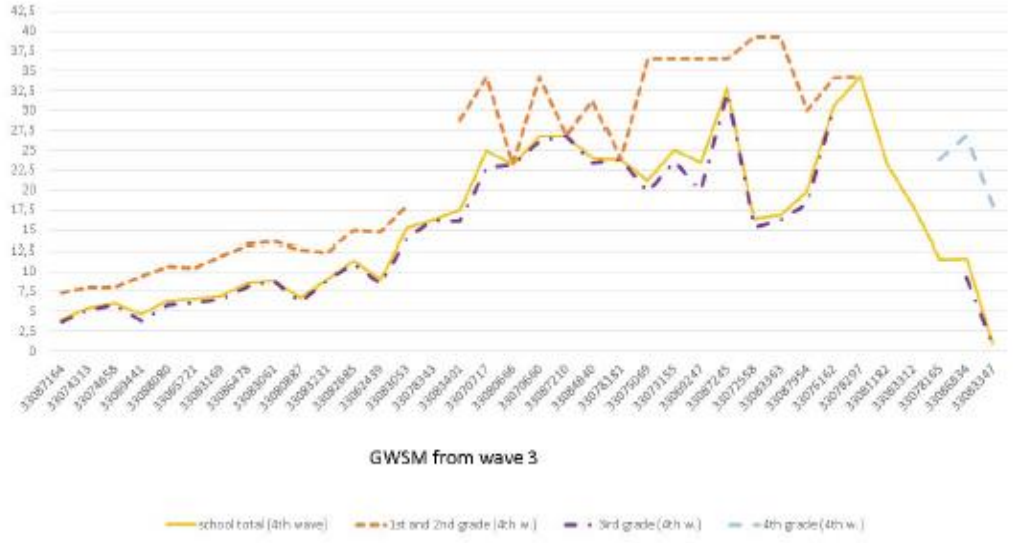


Figure 3: Estimation weights for wave 4 - using GWSM from estimation weights of wave 3, - RJ, municipal schools



4. Discussion and Conclusion

At the moment, we are running a small simulation to study the performance of the estimators for the total of repeaters in wave 4, estimating the variances via Bootstrap. Also, the estimation weights were calculated previous any adjustment, as non-response, calibration or poststratification.

Then, we are also working on the calibration of the population estimates given the survey did not follow all initial pupils.

References

1. Brooke, N. and Bonamino, A. (2011). Raz˜oes e resultados de uma pesquisa longitudinal sobre efic´acia escolar. Rio de Janeiro: Walprint gr´afica e editora.
2. Lavall´ee, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. Survey Methodology, 21(1):25-32.
3. Lavall´ee, P. (1997). Indirect Sampling. Springer Science and Business Media.

Health Tourism in Malaysia: Does this industry provide a catalyst for economic growth?

Siti Norfadillah Md Saat, Zainuddin Ahmad
Department of Statistics Malaysia

Abstract

Medical or health tourism is one of the new growing industry in global market and Malaysia is heading towards positioning the country to be an attractive destination for health tourism. In 2017, the number of healthcare travellers visit to Malaysia posted a growth of 12.7 per cent compared to 2016. Whilst, Malaysia's Gross Domestic Product (GDP) recorded a growth of 5.9 per cent in 2017. The objective of this study is to examine the relationship between real Malaysia's GDP and healthcare travellers that visit to Malaysia for seeking the treatment. This study covers the quarterly data from Q1 2012 to Q4 2017. Unrestricted VAR model is used to determine short term relationship between the variables. Granger causality test is employed to investigate the causality between economic growth and the healthcare travellers. The results show that healthcare traveller's lag one is significant to explain healthcare travellers. However, there is no statistical evidence that healthcare travellers influence real GDP of Malaysia. In term of causality, healthcare travellers Granger cause real GDP.

Keywords

Health tourism; medical tourism; healthcare; economic growth

1. Introduction

The terms of health tourism is not new in healthcare industry. It is often used synonymously with the term "medical tourism," which is a part of the same spectrum. The Medical Tourism Association's defines medical tourism as when people who live in one country, travel to another country to receive medical, dental and surgical care while at the same time receiving equal to or greater care than they would have in their own country, and are traveling for medical care because of affordability, better access to care or a higher level of quality of care.

Globalisation has made health tourism possible and continues to flourish. The encouraging development of this industry can provide significant opportunity through a combination of key services sector such as health, accommodation, travels/leisure, food & beverages and transportation. Furthermore, health tourism will provide more job opportunities which may

help improve the people's standard of living, especially the younger generation.

The health tourism industry is also dynamic and volatile due to various factors such as economic climate, travel restrictions, geopolitical shifts, domestic policy changes, political instability, advertising practices and innovative and new treatment options contribute towards a shift in the pattern of consumption and production of health services in a domestic and overseas market (Zion Market Research, 2018).

Health tourism in Malaysia has emerged as one of the fastest growing segments over the last few years. The intensification of the industry can be seen from the increasing number of international patients into the country. In 2017, Malaysia received 1,038,632 healthcare travellers with revenue of RM1,265.8 million. In 2018, Malaysia Healthcare Travel Council (MHTC) aims to achieve RM1.3 billion in revenue, and potentially contribute RM5 billion to the nation's gross domestic product. Table 1 indicates the number of healthcare travellers visit Malaysia and the real GDP for year 2012 to 2017.

Table 1: Number of Healthcare Travellers Visit Malaysia and GDP, 2012 – 2017

Year	2012	2013	2014	2015	2016	2017
Healthcare Travellers (Persons)	648,132	770,134	837,718	853,875	921,481	1,038,632
Real GDP (RM' million)	912,261	955,080	1,012,448	1,063,998	1,108,311	1,173,177

Source: Department of Statistics Malaysia (DOSM) and Malaysia Healthcare Travel Council (MHTC)

This paper intends to study the relationship between real GDP of Malaysia and healthcare travellers to Malaysia.

2. Literature Review

A review of the literature indicates that health tourism has usually been considered a positive contribution to economic growth. In recent years, many countries have been actively promoting health tourism to stimulate economic growth. Chor Foon Tang (2015) examines the effect of medical tourism on economic growth in Malaysia. The cointegration, Granger causality and also the Generalised Variance Decomposition are applied. The results indicate that economic growth, medical tourism and other determinants in Malaysia are cointegrated. Moreover, the empirical results suggest that medical tourism Granger-cause economic growth in Malaysia, regardless of short or long run.

Harun Uçak (2016) investigates the effect of health and social service sector growth on the flow of inbound health tourism in Turkey by employing Granger causality and Johansen cointegration approaches. The findings suggested that there is a long-run Granger causality from domestic health and social work expenditures to health tourism income whereas this is non-existence in the opposite direction. Another finding of the study is that there

is no causality from number of health sector tourists to health sector expenditure in Turkey.

In another study, Zahra Pourkhaghan (2013) examines the interaction of economic indicators and medical tourism using descriptive and qualitative content analysis. The results show that medical tourism is helping the sustainable development and economy dynamism through exchange gaining, creating jobs and etc. This type of tourism is more profitable than the other subdirectories of tourism and considers the sustainable development of tourism destinations in a more appropriate manner.

Wang Liangju, Zhang Huihui and Li Wanlian (2012) investigate the causal relationship between China's domestic tourism and economic growth by using cointegration analysis and Granger causality test. Cointegration analysis indicates that there are long-term and stable equilibrium relationships between the development of China's domestic tourism and economic growth. The results from the Error Correction Model indicate that there are short-term disequilibrium relationship between the development of China's domestic tourism and economic growth. The development of China's domestic tourism is the Granger cause of economic growth, and China's economic growth is the Granger cause of development of domestic tourism as well. The findings imply that China may enhance its economic growth by strategically strengthening the tourism industry while not neglecting the other sectors which also promote growth.

In a study on the relationship between healthcare and tourism sectors to economic growth in Malaysia, Singapore and Thailand, Chan-Fatt Cheah and A. S. Abdul-Rahim (2018) employed ARDL test. The results show a significant positive short and long run relationship between development of healthcare and tourism sectors to economic growth in Malaysia, Singapore and Thailand.

3. Methodology

The aim of this study is to examine relationship between Malaysia real GDP and healthcare travellers who come to Malaysia for treatment. This study covers the quarterly time series data obtained from DOSM and MHTC that covers the period from quarter 1 2012 to quarter 4 2017. The data used are Malaysia real GDP and the number of healthcare travellers' visits to Malaysia for the purpose of health or for medical reasons. In this study, it is found that Malaysia real GDP has seasonality. Hence, the seasonality is removed before conducting relationship study.

In this study, general model is as follows:

$$Y_t = b_{10} - b_{12}X_t + \gamma_{11}Y_{t-1} + \gamma_{12}X_{t-1} + \varepsilon_{yt}$$
$$X_t = b_{20} - b_{21}Y_t + \gamma_{21}Y_{t-1} + \gamma_{22}X_{t-1} + \varepsilon_{xt}$$

Where:

Y = real GDP
X = number of healthcare travellers visit to Malaysia
b and γ = constant term
t = time trend
ε = error term.

We begin the analysis by investigate the stationarity of variables using the Augmented Dickey-Fuller (ADF) unit root test. In addition, the optimal lag is chosen carefully using the Akaike Information Criterion (AIC).

Unrestricted Vector Autoregressive (VAR) model is employed to determine short run relationship between the variables before the causality test. In this study, the Granger causality test is employed to investigate causal relationship between economic growth and the number of healthcare travellers. Granger introduced the concept of Granger causality in 1969 and it has been widely used in econometrics studies to test availability and the direction of the causality (Granger, 1969). It is also necessary to do model diagnostics, in order to check whether the fitted model is appropriate.

4. Empirical Result and Discussion

Correlation analysis between Malaysia real GDP and healthcare travellers shows a very strong positive relationship (r = 0.88).

4.1 Stationarity test

The ADF test for stationarity shows that healthcare travellers is stationary at level. Meanwhile, real GDP is stationary after it is converted into the first difference. The null hypothesis of non-stationary can be rejected when the p-value is less than a significant level of 5 per cent. The summary of ADF is in Table 2.

Table 2: Augmented Dickey Fuller Test Result

Variable	Stationary	t-stat	p-value
GDP	First Difference	-5.094924	0.0026
Healthcare travellers	Level	-5.375531	0.0013

Source: Author computation

4.2 Optimal lag

Optimal number of lags is conducted using appropriate lag length selection criteria. The results of AIC show that optimal lag is three. The summary is in Table 3.

Table 3: Optimal Lag: Akaike Information Criterion (AIC)

Lag	AIC
2	66.25504
3	66.13122*
4	66.43926

Source: Author computation

4.3 Unrestricted VAR Model

After stationary test of the variables and lag selection, the Unrestricted VAR model is employed to study short run relationship between real GDP and number of healthcare travellers. The results show that healthcare travellers do not influence real GDP of Malaysia. However, for healthcare travellers as dependent variable, there is positive relationship for healthcare travellers lag one and healthcare traveller (Table 4). One per cent increase in healthcare traveller lag one will result in 0.68 per cent increase in healthcare. However, healthcare travellers do not influence real GDP of Malaysia.

Table 4: Unrestricted VAR Model Result

Variables	Coefficient	Std. Error	t-Statistic	Prob.
Dependent variable (Healthcare Travellers)				
Healthcare Travellers lag 1	0.680065	0.281419	2.416556	0.0224

Source: Author computation

4.4 Granger Causality Test

The Granger causality test result (Table 5) indicates that healthcare travellers Granger cause real GDP at one per cent significant level. Meanwhile, the real GDP doesn't Granger cause healthcare travellers.

Table 5: Granger Causality Tests Result

Null Hypothesis:	Chi-sq	Prob.
Healthcare Travellers does not Granger Cause GDP	15.35687	0.0015
GDP does not Granger Cause Healthcare Travellers	3.541802	0.3154

Source: Author computation

4.5 Diagnostic Test

In this Unrestricted VAR model, R-square and F-statistics are significant. Serial correlation test is also conducted and the results show that there is no serial correlation (prob. 0.1177). Check on residual normality which use Jarque-Bera shows that the residual is normal (prob. 0.6235). Finally, Breush-Pagan-Godfrey result shows that there is no residual heteroskedasticity (prob. 0.3326).

5. Conclusion

The objective of this study is to examine the relationship between real Malaysia's GDP and healthcare travellers that visit to Malaysia. Unrestricted VAR model and Granger causality test technique are used in this study. The results show that healthcare traveller's lag one is significant to explain healthcare travellers. This study also found there is no statistical evidence that healthcare travellers influence real GDP of Malaysia. In terms of causality,

healthcare travellers Granger cause real GDP. From this study, Malaysia government can take measures in enhancing and promoting the healthcare travel industry of Malaysia. The strong government support and involvement collaborations between public-private partnerships domestically and abroad will lead towards strengthening this industry.

In this study, there is only one independent variable examined. For further research, other variables that can be included are real exchange rate, investment or the domestic tourism arrival.

References

1. Chan-Fatt Cheah and A. S. Abdul-Rahim, (2018). Relationship between Health Care and Tourism Sectors to Economic Growth: The Case of Malaysia, Singapore and Thailand, Pertanika J. Soc. Sci. & Hum. 26 (2), 1203-1214.
2. Chor Foon Tang, (2015). Medical Tourism and Its Implication on Malaysia's Economic Growth. MPRA Paper No. 63365, 3-7.
3. Granger C. W. J., (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica. Vol. 37, No. 424-438.
4. Harun Ucak, (2016). The Relationship between the growth in the health sector and inbound health tourism: the case of Turkey. Springer Plus, Pg. 4-8.
5. I Aniza , M Aidalina, R Nirmalini, MCH Inggit, TE Ajeng, (2009). Health Tourism in Malaysia: The Strength and Weaknesses. Jurnal of Community Health 2009, Vol. 15 Number 1.
6. Markéta Arltová and Darina Fedorová, (2016). Selection of Unit Root Test on the Basis of Length of the Time Series and Value of AR(1) Parameter. Pg. 48.
7. Osaro Aigbogun, Sadoun Naser Yassin, Zeeshan Shoukat, (2013). A Model for Accelerating the Growth of Health Care Tourism in Malaysia. Journal of Business and Economics, ISSN 2155-7950. Vol. 4, No. 2, 169-179.
8. Wang Liangju, Zhang Huihui and Li Wanlian, (2012). Analysis of Causality between Tourism and Economic Growth Based on Computational Econometrics. Journal of Computers, Vol. 7, No. 9, 2155-2158.
9. Zahra Pourkhaghan, (2013). Interaction of Economic Indicators and Medical Tourism Industry. International Journal of Travel Medicine & Global Health. Vol. 1, Issue 3, 134-138.
10. Zion Market Research, (2018, June 22), Global Medical Tourism Market Set for Rapid Growth, to reach Value around USD 28.0 Billion by 2024. Retrieved February 4, 2019 from https://www.zionmarketresearch.com/news/medical-tourism-market

Functional Areas: opportunities for "Leave no one behind' agenda and statistical challenges

Florabela Carausu[1], Ibtissam Sahir[1]
1 GOPA Luxembourg S.a r.l

Abstract

It is widely recognised that reliable data are the cornerstone of evidence-based decision making, and in particular it is at the local or regional level where concrete and reliable data remain scarce. The call for more focus on the geographical breakdown can serve – among others - to monitor Sustainable Development Goals (SDGs) at sub-national level, in line with the commitment to Leave no one behind, envisaged by the SDG Agenda 2030.

If it is commonly agreed that the effectiveness of policy making requires a lower level of analysis and intervention, the selection of the sub-national level is key for its efficiency. A lower level is needed for policy making, and data is needed for analysing and interpreting adequately the social and economic interactions within the territory, so as to ensure that those staying behind do not get left behind. An analogy in the approach to development planning and monitoring, with the shift from regional policy to territorial cohesion, and the current targeting of functional areas instead of the classical regions, in the European context, is done with the aim to promote this approach in different contexts.

The territorial interdependencies and interaction imply that almost any development issue has to extend beyond the administrative borders. The examples are diverse, from social inclusion to industrial and economic synergies, to environmental challenges or infrastructure provision, etc. This improved orientation to policy making establishes a real challenge from the statistical point of view. On one side, data should be available for the definition or delineation of functional areas, and furthermore, the methods for the delineation of functional areas, as the Labour Market Areas case study showed, are sensitive to the quality of the input data. On the other side, once defined the functional areas, data and indicators at their level should be produced with certain regularity in order to allow for the monitoring and evaluation of the targeted measures. The necessity of continuation makes the difference between action oriented analysis and decision oriented analysis.

Keywords

functional areas; Leave no one behind; regional and territorial statistics; official statistics; SDGs

1. Introduction

It is widely recognised that reliable data are the cornerstone of evidence-based decision making, and in particular it is at the local or regional level where concrete and reliable data remain scarce.

As highlighted by the Paris 21 PRESS 2017[1] report: data and statistics are attracting more resources and new donors, but the support remains insufficient. More and better-quality financial support to data and statistics is vital to ensure robust SDG monitoring at national level. In the same line, the UN Data Revolution for Sustainable Development report 'A World that Counts' emphasises that "data are the lifeblood of decision-making and the raw material for accountability"[2].

The increased demand for regional and local statistics asks for the modernisation of statistical processes. The call for more focus on the geographical breakdown can serve – among others - to monitor Sustainable Development Goals (SDGs) at sub-national level, in line with the commitment to Leave no one behind, envisaged by the SDG Agenda 2030.

While the Leave no one behind[3] slogan has been widely adopted, there are different interpretations. In this paper the following features or readings of the slogan are considered: a lower level is needed for policy making, and data is needed for analysing and interpreting adequately the social and economic interactions within the territory, so as to ensure that those staying behind do not get left behind. An analogy in the approach to development planning and monitoring, with the shift from regional policy to territorial cohesion, and the current targeting of functional areas instead of the classical regions, in the European context, is done with the aim to promote this approach in different contexts.

2. Methodology

At the European level the regional policy has experienced peaks and falls, and has been criticised that it favoured most developed regions, especially in the context of the entry of the new Member States (MS) after the 2004 enlargement.

As an alternative, the Barca report[4] proposed a place-based (development) policy, understood as a long-term strategy aimed at tackling persistent underutilisation of potential and reducing persistent social exclusion in

[1] Partnership in Statistics for Development in the 21st century, Partner Report on Support to Statistics

[2] UN Data Revolution for Sustainable Development Independent Expert Advisory Group : "A World that Counts"

[3] T. German; J. Randel (2017): 'Delivering results to Leave no one behind' – discussion paper for the Results Community OECD workshop 'What Results – Who Counts?'

[4] Fabrizio Barca (2019): 'An Agenda for a Reformed Cohesion Policy – A place-based approach to meeting European Union challenges and expectations'

specific places through external interventions and multilevel governance. The strategy proposed by Barca makes accountable the territorial focus.

The territorial cohesion entered in the European context as a consequence of the limitations of the European regional policy, criticised for placing more emphasis on the potential (capabilities) of the region, than on financial redistribution, and also for the lack of coherence between the actions financed. This understanding has promoted the territorial cohesion, as a territorial development approach characterised by a focus on the use of endogenous potentials, promoting the creation of functional and spatial structures such as functional areas.

If it is commonly agreed that the effectiveness of policy making requires a lower level of analysis and intervention, the selection of the sub-national level is key for its efficiency. Traditionally, the region is a sub-national level formed by the grouping of adjacent administrative units (e.g. municipalities). Nevertheless, the European context has shown that, the single most important policy indicator in regional policy, the per capita average value of Gross Domestic Product (GDP) is frequently distorted[5] when using administrative units. The reason is that the income generated in one municipality, such as a city, may be largely consumed by households in other municipalities. And this may turn into questioning the credibility of (some important) official statistics, because people in their day-to-day life do not identify with "averages" (i.e. they don't feel official statistics capture the correct image).

The limitation of the administrative units approach to region definition helps arguing that what is needed in the socio-economic policy field is a set of areas[6] which are equivalent to 'consumption areas' (Coombes et all, 2012). Pursing this further, an analogy can be established with the standard set by Leave no one behind; i.e. averages and generalised progress are not enough because they do not reveal who is missing. In the context of SDGs Agenda for 2030, the Leave no one behind puts as much emphasis on 'who' benefits as on 'what' has been delivered[7]. The approach aims for a strategy ensuring that those staying behind do not get left behind. 'This requires data and different benchmarks of progress. Data needs to be disaggregated to show how different parts of the population are faring. Benchmarks cannot rest on a specified line or target, they have to measure the extent to which different people are being included in the rate of progress'[8]

[5]ESPON (2007) Final Report of the ESPON Project 1.4.4 "Preparatory Study on Feasibility of Flows Analysis Final Report"

[6] Coombes M., Wymer C., Casado J.M., Martinez L., Carausu F. (2012): 'Study on comparable Labour Market Areas' prepared for Eurostat

[7] 7T. German; J. Randel (2017): 'Delivering results to Leave no one behind' – discussion paper for the Results Community OECD workshop 'What Results – Who Counts?'

[8] idem

Leave no one behind slogan links to the local community (the individuals), but as well to territorial capital. Therefore, the approach should be the same as when shifting from regional policy to territorial cohesion; i.e. the development approach should be characterised by a focus on the use of endogenous potentials and territorial targeting, promoting the creating of functional and spatial structures (i.e. functional areas).

The concept of functional areas may be ambiguous, and it appeared rather as a targeted alternative, than as concrete solution. Maybe the most well-known conceptualised functional areas are the Functional Urban Areas (FUA) and Labour Market Areas (LMA).

FUA have been developed by OECD in collaboration with the European Commission (DG REGIO and Eurostat). A harmonised definition of the urban areas as 'functional economic units', overcoming the limitation linked to the administrative units, has been developed by the two institutions. The methodology used to identify the FUAs chooses as building blocks the smallest administrative units for which national commuting data are available. The methodology is applied to 29 OECD countries. The OECD metropolitan database[9] publishes a set of annual variables related to some of the OECD functional urban areas with a population above 500 000.

Labour Market Areas are defined as functional areas defined based on the patterns of commuting, similar to FUA, but not limited to major cities and their immediate regions.

The LMA approach has been investigated by Eurostat[10] with the support of researchers, and the proposed LMA methodology has been tested by several EU Member States through grants. The initiative targeted the definition of consistently defined LMAs covering the entire territory of the EU. Preliminary initiatives at national level showed that some EU MS had defined LMAs for guiding policy actions to improve regional economic structures, for defining industrial districts, for the dissemination of socio-economic statistics at lower spatial scales or for improving public transport provision, among others.

3. Result

The fact that the selection of the functional area is key for the efficiency of the policy in question, and sensitive for the comparison of results, allows arguing that there isn't any standard set of areas that is ideal for all type of analyses. The most appropriate set of areas will depend on the purpose of the

[9] http://measuringurban.oecd.org/

[10] https://ec.europa.eu/eurostat/cros/content/labour-market-areas_en

analysis concerned[11]. It is important to emphasise that location and place are vital components of effective decision making. Actually, the spatial / territorial level which is chosen for describing a phenomenon plays the role of a filter. The change of zoning leads automatically to a change in the results[12].

Reliable data are the cornerstone of evidence-based decision making, but at the same time making proper use of the data is critical as well, especially given that data is not 'neutral'. Evidence-based decision making would not deliver either the anticipated results, even though quality data would be available, if not accompanied by a sound decision-making approach or mechanism.

Recently it has been argued (Radermacher, 2018) that "just as the map is not the territory, so (official) statistics will never be as accurate and complete as the reality they represent. There will always be more or less significant differences, omissions, generalisations and distortions between statistics (the map) and the field in question (the territory). Indeed,

statistics are only a partial representation at a given moment in time of a reality which is not static, but in constant motion and of a complexity which is impossible to portray precisely and exhaustively". As emphasised by the author, the distinction between 'map' and 'territory' is needed for explaining and attempting to draw the boundaries between objective truths and subjective reality.

The need for more disaggregated data becomes of utmost importance, but not the final aim. Though, the production or access to lower level data can be assigned as a priority, improving its quality should go in parallel with offering the support for evidence-based policy making. The high demand for disaggregated data may not be in most of the cases satisfied by the official statistics producers, the data at local level not being collected through most frequent surveys, but only through census. Nevertheless, the official statistics producers can substantially contribute by controlling or advising on the quality of data coming from other sources (e.g. administrative data, Big Data) or by supporting specialised statistical analyses for estimating data at lower levels.

BigData, administrative data, and Small Area Estimation (SAE), but also the upcoming census round offer opportunities supporting the definition of functional areas and following up on the implementation of policies at their level.

[11]Coombes M., Wymer C., Casado J.M., Martinez L., Carausu F. (2012): 'Study on comparable Labour Market Areas' prepared for Eurostat

[12]For more details, see ESPON (2006): The Modifiable Areas Unit Problems

Big Data can directly or indirectly benefit policy making, if the associated risks are adequately prevented, by:

- answering new questions and producing new indicators;
- bridging time lags in the availability of official statistics and supporting the timelier forecasting of indicators;
- providing an innovative data source in the production of official statistics.[13]

Mobile positioning data, a Big Data source, can offer the possibility to detect daily commuting flows, allowing for the delineation of LMAs or offers the possibility for population estimates.

Other statistical techniques may include the downscaling of socio economic indicators based on grid cells and GIS opportunities; see ESPON(2011): Disaggregation of socioeconomic data into a regular grid and combination with other types of data.

At European level, Eurostat promotes and finances exploratory analyses of functional areas, such as the Labour Market Areas (LMA); see https://ec.europa.eu/eurostat/cros/content/labour-market-areas_en

More recently, Eurostat has entrusted to the consulting company GOPA Luxembourg S.a r.l in collaboration with the University of Trier, a study on 'Small Area Estimation (SAE) for city statistics and other functional areas (part II)'. The objective of the study is to test more sophisticated Small Area Estimation (SAE) methods and to produce guidelines to Eurostat and the NSIs on how to estimate data from social surveys such as the Survey on Income and Living Conditions (SILC) and the Labour Force Survey (LFS) at city and Functional Urban Area (FUA) level. The main outcome will be a set of guidelines proposing a sound methodology using SAE (and other method such as cluster analysis and probability statistics) to calculate variables and indicators of interest. The guidelines should be applicable to all kind of social surveys, not only the SILC. The guidelines will be finalised by mid-2019.

The study is a continuation of 'Small Area Estimation (SAE) for city statistics and other functional areas (part I)'[14] which tested the application of SAE methods on the Urban Audit city data collection on the basis of the indicator Share of Persons at Risk of Poverty or Social Exclusion.

[13] See IMF Staff Discussion Note: 'Big Data: Potential Challenges and Statistical Implications', SDN/17/06, September 2017

[14] https://circabc.europa.eu/webdav/CircaBC/ESTAT/regstat/Library/Working%20Group%20Meeting%202017/Documents/9.3%20E4_REG_2017_93_Annex_SAE%20for%20City%20statistics.pdf

4. Discussion and Conclusion

The focus on the geographical breakdown has raised expectations from the users of statistics. Nevertheless, not always the hyper-local or hyper-detailed approach is possible or relevant. Understanding the trade-off between accuracy and relevance is important. The close and regular dialogue between (official) statistics producers and users is needed in order to bring in line the demand and the supply for local and regional statistics.

The territorial interdependencies and interaction imply that almost any development issue has to extend beyond the administrative borders. The examples are diverse, from social inclusion to industrial and economic synergies, to environmental challenges or infrastructure provision, etc. This improved orientation to policy making establishes a real challenge from the statistical point of view. On one side, data should be available for the definition or delineation of functional areas, and furthermore, the methods for the delineation of functional areas, as the Labour Market Areas case study showed, are sensitive to the quality of the input data.

On the other side, once defined the functional areas, data and indicators at their level should be produced with certain regularity in order toallow for the monitoring and evaluation of the targeted measures. The necessity of continuation makes the difference between actionoriented analysis and decision oriented analysis.

In the European context, there is a wide consensus that functional areas are the preferred level for policy making, though there is not any ideal functional area, they all depend on the purpose of the analysis concerned. Functional areas can be seen as an approach and opportunity to target the Leave no one behind objective. At European level there are several initiatives that may serve as an example.

References

1. Partnership in Statistics for Development in the 21st century, Partner Report on Support to Statistics
2. UN Data Revolution for Sustainable Development Independent Expert Advisory Group: "A World that Counts"
3. Coombes M., Casado J.M, Martinez L, Wymer C.: 'Labour Market Areas (LMAs): the challenge of meeting policy and statistical requirements', SCORUS 2018 Warsaw, Poland
4. T. German; J. Randel (2017): 'Delivering results to Leave no one behind' – discussion paper for the Results Community OECD workshop 'What Results – Who Counts?'
5. Fabrizio Barca (2019): 'An Agenda for a Reformed Cohesion Policy – A place-based approach to meeting European Union challenges and expectations'
6. ESPON (2007) Final Report of the ESPON Project 1.4.4 "Preparatory Study on Feasibility of Flows Analysis Final Report"
8. Coombes M., Wymer C., Casado J.M., Martinez L., Carausu F. (2012): 'Study on comparable Labour Market Areas' prepared for Eurostat
10. Cervera, J. & Carausu, F. (2018): "Regional statistics in transition and developing countries: lessons learnt from technical assistance", SCORUS 2018 Conference, Warsaw June 2018 (SCORUS – Standing Committee of Regional and Urban Statistics)
11. Radermacher W. (2018): 'Official Statistics in the era of big data opportunities and threats', International Journal of Data Science and Analytics, November 2018, Volume6, Issue 3 pp225-231
12. OECD (2013): Definition of Functional Urban Areas (FUA) for the OECD metropolitan database, OECD, Paris

A study of the factors affecting Hong Kong residents' willingness to pay for waste disposal by Logit Models

Iris M H Yeung, William Chung
Department of Management Sciences, City University of Hong Kong

Abstract

A survey was conducted to study the knowledge and attitude of residents towards three government proposed policies (waste charge, landfill extension and building a new incinerator) for waste management in Hong Kong. About one third of the respondents are willing to pay (WTP) less than HK$30 (36.1%), exactly HK$30 (32.4%) and more than HK$30 (31.4%) for waste disposal respectively. The average WTP amount is HK$38.4 per month. Logit models indicate that the degree of support for waste charge and new incinerator policies, daily waste disposal amount, age and income significantly affect WTP for the contrast between "above HK$30" and "below HK$30". Multinomial logit model suggests that only degree of support for waste charge policy, knowledge on landfill fullness and construction time for new incinerator significantly affect WTP for the contrast between "exactly HK$30" and "below HK$30". Implications of these findings will be discussed.

Keywords

Environment; Multinomial and Ordinal Logit Models; Solid Waste; Willingness to pay; Hong Kong

1. Introduction

Solid waste management is a big challenge to the government authorities throughout the world (Ma and Hipel 2016, Vergara and Tchobanoglous 2012). Since 2005, the Hong Kong government has launched various waste reductions at source programs, such as source separation of domestic, commercial and industrial waste; the implementation of construction waste disposal charging scheme; producer responsibility scheme for local e-waste reduction and environmental levy scheme on plastic shopping bags. However, the amount of waste in Hong Kong is expected to increase along with the population and economic growth, and consumption patterns of people. But the three landfills that are mainly used for waste disposal are nearly filled up. To solve the waste management problem, the Hong Kong government proposed three policies in May 2013, namely waste charge, landfill extension and building a new incinerator.

A questionnaire survey was carried out by the authors to understand the knowledge and attitude of residents towards the three proposed policies.

Yeung and Chung (2018) examined the effect of knowledge and attitude of local residents on three government proposals, daily waste amount, and socio-economic characteristics on willingness to pay (WTP) for waste disposal using binary logit model. In this paper, we shall use multinomial and ordinal logistic regression to revisit the problem. The results help to understand more clearly about the factors affecting residents' WTP and are useful to the government and related parties in policy making and implementation. The succeeding sections are organised as follows: Section 2 describes the methodology. Section 3 presents the results. Section 4 draws the conclusion.

2. Methodology

Sampling and data collection

The survey targeted all Hong Kong residents aged at least 18. Simple random sampling method and computer-assisted telephone interview (CATI) system were used to select the households. The household member with the nearest birthday were asked to complete the questionnaire.

1005 persons were contacted. After eliminating questionnaires which refuse to answer three key attitude questions (the degree of support towards waste charge, landfill extension and building a new incinerator respectively), WTP questions and more than half of the socio-economic questions, the final sample data set used in this study consisted of 753 responses.

Variables used for logit models

The dependent variable used for multinomial and ordinal logit models is WTP for waste disposal. Contingent valuation method was used to estimate WTP amount. Following researchers such as Zhang et al (2012), the respondents were first asked whether they were willing to pay HK$30 each month if they disposed two waste bags daily (i.e. around 60 bags per month) with sizes the same as an 8-kg rice bag in supermarkets. Then, they were asked to give the maximum WTP in an open-ended question. The reason for using the amount of HK$30 in the first WTP question was that according to the consultation paper of The Council for Sustainable Development in Hong Kong, the waste disposal charge for each household was estimated to range from HK$30-HK$60 (roughly US$3.85-7.69) per month (The Standard, 2013). The second WTP question was used to estimate the maximum amount that the respondent would pay. Furthermore, this second question gave respondents who did not answer the first WTP question a second chance to express their WTP amount. In this paper, we code WTP into three groups (below HK$30, exactly HK$30, and above HK$30).

The independent variables include knowledge of the respondents on waste charge method, landfill fullness and construction time for new incinerator (coded to 2 categories: 1= correct knowledge, 0 = not correct), degree of

support towards the three government proposed policies, which are measured by a five-point Likert scale (from 1 = strongly oppose to 5 = strongly support), the daily number of waste disposal bags with sizes the same an 8-kg rice bag in supermarkets), gender (male or female), residential district (coded to 2 categories: 1 = live near landfill and new incinerator site, 0 = not near), education (coded to 3 categories: primary and below, secondary or post-secondary), age (coded to 6 categories: 18–24, 25–34, 35–44, 45–54, 55–64 and ≥65 years), household size (coded to 6 categories: 1, 2, 3, 4, 5, 6 and above), type of living quarters (coded to 3 categories: "public rental housing", "subsidized home ownership housing", or "private housing and others"), monthly personal income (coded to 3 categories: "under HK$15K," "HK$15K and under HK$30K," and "HK$30K and above").

3. Result

Characteristics of the respondents

The sample consisted of 40% males and 60% females. A majority of the respondents (at least 81%) did not live near the landfill sites and the new incinerator. Around 80% of the respondents received secondary school or above education. The average age and monthly personal income of the respondents were 47.07 years and HK$13, 151 respectively. More than 60% of the respondents lived in subsidized home ownership housing and private housing. As the effect size (Cohen 1992) of gender, residential district, education, age, type of living quarters and monthly personal income fell between 0.01 and 0.24 based on the population data in 2013 (Census and Statistics Department 2014), the sample was generally representative of the population of the residents in Hong Kong.

Approximately one third of the respondents were willing to pay less than HK$30 (36.1%), exactly HK$30 (32.4%) and more than HK$30 (31.4%) for waste disposal respectively. The monthly average WTP was HK$38.4, which is slightly above the minimum waste charge amount of HK$30-$60 estimated by the Council for Sustainable Development in Hong Kong.

More than half of the respondents (51.1%) were able to answer correctly that the government preferred quantity-based waste charge method. Few respondents knew that one of the landfills would be full around 2015 (22.7%) and it would take 7-9 years to build new incinerator (7.3%). More than half of the respondents indicated support for waste charge, landfill extension and building a new incinerator proposals (52%, 54% and 69% respectively). About 60% of the respondents disposed one waste bag each day.

Factors affecting three WTP groups

Both multinomial and ordinal logit models are statistically significant (likelihood-ratio chi-square test statistic = 216.04; p value < 0.0001and

likelihood-ratio chi-square test statistic = 194.31; p value < 0.0001 respectively). However, multinomial logit model has slightly bigger generalized R-square value (0.281 versus 0.256) and the score test of the proportional odds assumption shows that the ordinal logit model does not meet the assumption (p value of 0.018). Despite this, the signs of the parameter estimates for almost all the variables in the ordinal logit model (column 8) are the same as those for multinomial logit model (columns 2 and 5), and the magnitude of the parameter estimates for most variables in the ordinal logit model lie between those of the two contrasts in multinomial logit model except for three knowledge variables. It suggests that the effects of most variables on WTP follow the ordered sequence of WTP except the three knowledge variables. Due to better model performance, the results of multinomial logit models are discussed. The following observations are drawn for the contrast between "WTP above HK$30" and "WTP below HK$30" being given in columns 2-4 of Table 1:

1. The degree of support for waste charge policy (b = 1.045; odds ratio = 2.843) has the greatest positive impact on the contrast. However, the degree of support for building a new incinerator (b = 0.226; odds ratio = 1.253) has the smallest positive impact on the contrast. Obviously the waste charge policy is more relevant to WTP. The odds of WTP above HK$30 increased by a greater factor (2.843) per unit increase in the degree of support for the waste charge policy.
2. Income (b = 0.591; odds ratio = 1.806) and daily waste disposal (b = 0.504; odds ratio = 1.655) have medium positive impact on the contrast. Residents who are richer and have more waste disposal are 81% and 66% more likely to be WTP above HK$30 per unit increase in income and waste disposal. These residents can afford to pay more and they have greater needs for waste disposal.
3. Age has a negative significant effect on the contrast (b = -0.031; odds ratio = 0.969). The older the person is, the less likely they are WTP above HK$30.

The parameter estimation results for comparing "WTP exactly HK$30" versus "WTP below HK$30" cases are given in columns 5-7 of Table 1.

The following observations were drawn:

1. Like "WTP above HK$30" versus "WTP below HK$30" contrast, the degree of support for waste charge policy (b = 0.611; odds ratio = 1.842) has the greatest positive impact. However, the magnitude is smaller perhaps because the WTP amount being compared ("exactly HK$30" vs. "below HK$30") is less extreme. This may explain why the degree of support for building a new incinerator, age, income and daily waste disposal amount have no significant positive impact on this contrast.

2. It is interesting to note that knowledge on landfill fullness and construction time for new incinerator has significant but opposite effects on the contrast. The positive effect for knowledge of landfill fullness (b = 0.464; odds ratio = 1.591) suggests that a person with this knowledge is 59% more likely to be willing to pay HK$30 rather than below HK$30. On the other hand, a person possessing knowledge of long construction time for new incinerator is 2.33 times (b = -0.843; odds ratio = 0.430) more likely to be willing to pay below HK$30 rather than exactly HK$30.

   Table 2 shows the percentage distribution of 3 WTP groups for all three knowledge variables. Based on the chi-square tests, all three knowledge variables affect WTP significantly. But respondents possessing knowledge on waste charge method and landfill fullness tend to be willing to pay HK$30 or above, whereas respondents possessing knowledge on construction time for new incinerator are willing to pay either below or above HK$30 rather than exactly HK$30. This may explain why knowledge on construction time has negative effect on this contrast.

Table 1. Parameter Estimates, Standard Errors (SE), and Odds Ratio (OR) for Multinomial and Ordinal Logit Models of the Willingness to Pay (n=753)

	Above HK$30 vs. Below HK$30			Exactly HK$30 vs. Below HK$30			Ordinal Logit Model		
Variable	Estimate	SE	OR	Estimate	SE	OR	Estimate	SE	OR
Intercept	-5.566***	0.967		2.500***	0.846		-4.848*** -3.193***	0.672 0.658	
Know waste charge	0.209	0.205	1.233	0.055	0.188	1.057	0.225	0.171	1.129
Know landfill	0.352	0.249	1.422	0.464**	0.232	1.591	0.122	0.144	1.252
Know incinerator	-0.169	0.372	0.844	-0.843**	0.413	0.430	-0.179	0.277	0.836
Support waste charge	1.045***	0.120	2.843	0.611**	0.103	1.842	0.794***	0.082	2.213
Support landfill	0.151	0.121	1.163	0.156	0.110	1.169	0.088	0.085	1.092
Support incinerator	0.226*	0.126	1.253	0.040	0.116	1.040	0.165*	0.089	1.179
Daily waste amount	0.504***	0.169	1.655	0.134	0.154	1.143	0.333***	0.118	1.395
Gender (male)	0.141	0.210	1.152	-0.176	0.199	0.839	0.089	0.149	1.093
Area (near landfill & incin)	-0.234	0.265	0.792	-0.024	0.233	0.977	-0.132	0.184	0.877
Edu (secondary)	0.012	0.336	1.012	0.195	0.272	1.216	0.027	0.225	1.028

Edu (post-secondary)	0.193	0.406	1.213	-0.136	0.359	0.873	0.167	0.281	1.182
Age	-0.031***	0.008	0.969	-0.011	0.008	0.990	-0.022***	0.006	0.978
Household size	-0.099	0.088	0.906	-0.107	0.079	0.898	-0.071	0.061	0.931
House type (HOS)	0.121	0.296	1.129	0.130	0.269	1.139	0.118	0.208	1.125
House type (private)	0.286	0.245	1.331	0.331	0.222	1.393	0.184	0.171	1.202
Income	0.591***	0.170	1.806	0.108	0.170	1.114	0.441***	0.121	1.554

Notes:
1. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.
2. Overall model evaluation criteria
Multinomial: Likelihood ratio statistic = 216.04 (p value < 0.0001), Generalized R-square = 0.281
Ordinal: Likelihood ratio statistic = 194.305 (p value < 0.0001), Generalized R-square = 0.256

Table 2: Distribution of Three WTP Groups for Knowledge on Waste Charge Method, Landfill Fullness and Construction Time for New Incinerator (n=753)

		Sample size	Below HK$30 (n=272)	Exactly HK$30 (n=244)	Above HK$30 (n=237)	Test results	p-value
Know waste charge	Yes	385	31.95%	32.47%	35.58%	$\chi2 = 8.030$**	0.018
	No	368	40.49%	32.34%	27.17%		
Know landfill	Yes	171	26.90%	36.26%	36.84%	$\chi2 = 8.248$**	0.016
	No	582	38.83%	31.27%	29.90%		
Know incinerator	Yes	55	41.82%	18.18%	40.00%	$\chi2 = 5.607$*	0.061
	No	698	35.67%	33.52%	30.80%		

Note: ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

4. Discussion and Conclusion

The current paper identifies the factors affecting residents' WTP for waste disposal in Hong Kong using multinomial and ordinal logit models. Both models suggest that higher degree of support for waste charge and building new incinerator policies, more daily waste disposal amount, younger age and higher income increase the likelihood of WTP for the contrast between "above HK$30" and "below HK$30". Multinomial logit model further shows that only higher degree of support for waste charge, knowledge on landfill fullness and construction time for new incinerator affect the likelihood of WTP for the contrast between "exactly HK$30" and "below HK$30".

The multinomial and ordinal logit models provide more information on the factors affecting residents' WTP as compared with binary logit models (Yeung and Chung 2018). So, the government might try to launch promotional and educational programs on both waste charge and new incinerator policies, and information on landfill fullness as they all have positive effects on WTP.

However, the educational program on construction time for new incinerator has to be designed carefully in order to bring positive rather than negative effect on WTP.

Apart from the above knowledge and attitude variables, the models suggest that people who are younger, richer and have more waste disposal are willing to pay more than HK$30. Since the government policies aim to reduce waste at source, these people might be encouraged to reduce waste even though they afford to do so. Also, government should consider the financial burden and fair waste charge calculation for people who are older, poorer and have less waste.

References

1. Census & Statistics Department (2014) Quarterly Report on General Household Survey (Fourth Quarter 2013), Government of Hong Kong
2. Chong W (2013) $30-$60 urged for rubbish disposal. The Standard. http://www.thestandard.com.hk/news_detail.asp?pp_cat=30&art_id=137938&sid=40476062&con_type=1. Accessed 26 November 2013
3. Cohen J (1992) A Power primer. Psychology Bulletin 112:155-159
4. Ma J, Hipel K W (2016) "Exploring social dimensions of municipal solid waste management around the globe – A systematic literature review", Waste Management 56: 3-12
5. Vergara S E and Tchobanoglous G (2012) "Municipal solid waste and the environment: a global perspective", Annual Review of Environment and Resources, 37: 277-309
6. Yeung I M H, Chung W (2018) "Factors that Affect the Willingness of Residents to Pay for Solid Waste Management in Hong Kong", Environmental Science and Pollution Research 25:7504-7517
7. Zhang W, Che Y, Yang K, Ren X, Tai J (2012) Public Opinion about the source separation of municipal solid waste in Shanghai, China. Waste Management and Research 30(12):1261-1271

Modernization in statistical training management system

Wan Azhar Wan Mokhtar, Mohd Ridauddin Masud, Siti Kartini Salim
Department of Statistics, Malaysia

Abstract

The development of in-house training programmes in the Department of Statistics, Malaysia (DOSM) has started since 1987, with the formation of a Working Group of Human Resource Development. The development of training landscape was uplifted in 2012 with the establishment of Institut Latihan Statistik Malaysia (ILSM). This was mainly to facilitate the need for development of human capital in DOSM. Information and Communication Technology (ICT) can be an important component in enhancing training management process. In this regard, this paper discusses two (2) aspects of modernization in statistical training management system which been innovated by DOSM. Firstly, DOSM Training Information & Management System (DTIMS) was initiated in 2017 to overcome challenges that faced by ILSM in planning and managing the training throughout the years. The system is also able to produce a comprehensive training report based on the records either by individual or group. Secondly, transformation in training mechanism was innovated through exploration of DOSM e-learning platform which known as Malaysia Statistical Ubiquitous Learning (MySUL). Along with the rapid development of ICT as well as changes of the statistical global is one of the factors that necessitated creation of online training solutions.

Keywords

Training Information & Management System; e-learning; Ubiquitous Learning

1. Introduction

The development of in-house training programmes in the Department of Statistics, Malaysia (DOSM) has started since 1987, with the formation of a Working Group of Human Resource Development. In 1990, the Statistical Development Division has been established with responsibility of providing in-house training programmes. Due to highly demand in conducting in-house and international training, the division has been further enhanced with the establishment of the Statistical Training Division in 2003.

The development of training landscape was uplifted in 2012 with the establishment of ILSM. This was mainly to facilitate the need for development of human capital in the DOSM following to the large scale of DOSM's restructuring in 2007 with additional of more than 1,000 recruitments. ILSM is

also responsible to enhance the statistical literacy and understanding of statistics among members of the public service and the community at large. During those years, method of training was through traditional ways which were mainly conducted in classroom and face-to-face mode.

In order to enhance the training management following to the larger numbers of staff and training volume, DOSM initiated the development of DTIMS in 2017. The system was inspired mainly to solve and overcome problems and challenges faced by ILSM in planning and managing the training throughout the years. It is also a need to produce a comprehensive report based on the records either by individual or group.

Transformation in training mechanism was innovated through exploration of DOSM e-learning platform which is known as MySUL. Rapid development of ICT as well as changes of the statistical global is the factors that necessitated creation of online training solutions. DOSM's staffs are located at different geographical locations and training them the traditional way is time consuming and costly. Due to these reasons, it became imperative for DOSM's staffs to be constantly updated on their knowledge and skills. This means training has to be continual and on-going. That is how e-Learning came into picture.

2. DOSM Training Information & Management System (DTIMS):

The development of DTIMS is in line with the establishment of ILSM, DOSM's centre of excellence which is responsible to provide a comprehensive and quality learning package as well as high impact outcome delivery to the stakeholders. Due to this, the development of DTIMS was initiated in 2017 in order to enhance the training management following to the larger numbers of staff and training volume. The system will also assist ILSM to overcome problems and challenges faced in planning and managing the training throughout the years.

DTIMS was developing with the objectives:

i. To manage competency profiling for DOSM staff including cadres to determine the competency level and training needed; internal; external or international;
ii. To enable and manage DOSM's staff an instructors' talent;
iii. To provide competency indicator for staff and Panel Pembangunan Sumber Manusia in managing expertise, career and promotion;
iv. To improve efficiency and flexibility of DOSM's training management including a comprehensive facility management at ILSM; and
v. To provide a budget management platform to monitor the return of investment (ROI) in DOSM's training management.

DTIMS structure was designed based on all elements involved in the training cycle. There are eight modules in DTIMS (Figure 1).

Figure 1: DTIMS's module

There are two main parties that have been taken into account in the development of the system i.e: system user (DOSM' staff, other agencies, trainers and cadres) and subject matter division (ILSM, BKS, BKP, BPM and BPID).

The architecture of the system was designed based on five components as shown in Figure 2.

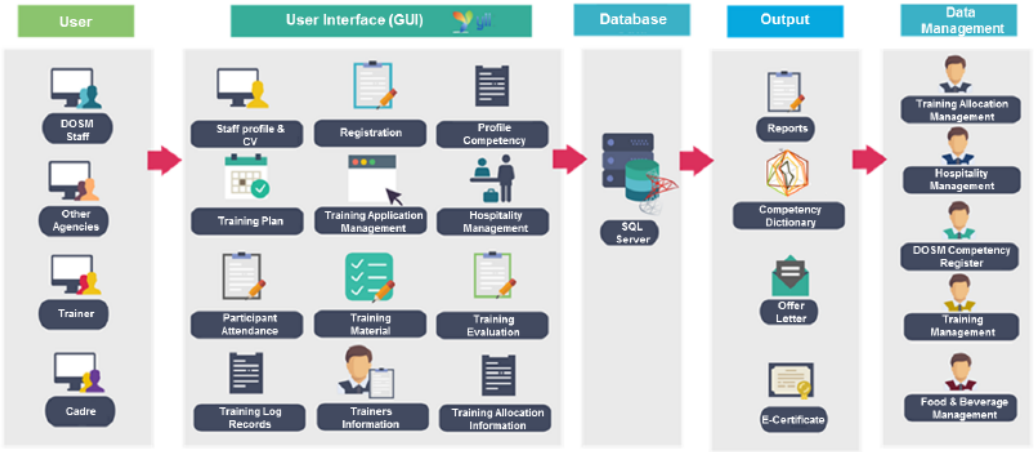


Figure 2: System Architecture

Users for the system consists of DOSM's staff, Other Agencies, Trainer and Cadre. The first sub module in **user interface (GUI)** is **staff profile & CV** which contain personal information for system users. **Registration** sub module is a space for the system users to fill out the information for registration. The next sub module is **Profile competency** which displays the user competency level throughout their services period. In sub module **Training plan,** users can identify training to be participated based on Training Schedule for the year. Sub module **Training Application Management** is a module controlled by

ILSM for training operation management. Meanwhile for the **Hospitality Management**, the sub module is used by the users to manage their hospitality needs during training in ILSM. **Participant attendance** is a sub module used by the ILSM and training secretariat to capture the participant attendance during training. The system also allows the users to upload and download materials required for training in the **Training material** sub module such as slide, brochure etc. In order for ILSM to monitor the participant achievement and future training improvement, sub module **Training Evaluation** has been formed to allow the participant to do online evaluation. **Training log records** is used to record all training attended by users throughout the services period. **Trainers Information** sub module is a platform to provide trainers profile for training secretariat reference. **Training Allocation Information** is very important as it will provide information on annual operational training cost.

As for **Database** component, Microsoft SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications. There are four (4) sub module under **Output** component i.e **Reports, Competency dictionary, Offer Letter** and **e-Certificate. Reports** refer to a document that presents information in an organized format for a specific audience and purpose. Although summaries of reports may be delivered orally, complete reports are almost always in the form of written documents. **Competency dictionary** is a tool or data structure that includes all or most of the general competencies needed to cover all job families and competencies that are core or common to all jobs within an organization. They may also include competencies that are more closely related to theknowledge and skills needed for specific jobs or functions. The training **Offer letter** and **e-Certificate** can be downloaded from DTIMS by participants.

Data management component is controlled by ILSM which covered training operations management. The sub module in this component comprise of **Training Allocation Management, Hospitality Management, DOSM Competency Register, Training Management and Food & Beverage Management.**

3. Malaysia Statistical Ubiquitous Learning (MySUL):

The rapid changes in a used of ICT in learning process has widely applied across the globe including Malaysia. The highly demand on more flexible and easier to reach learning method has become priority. The blended learning **(Figure 3)** approach was come into consideration in a way to harmonize the human capital development needs.

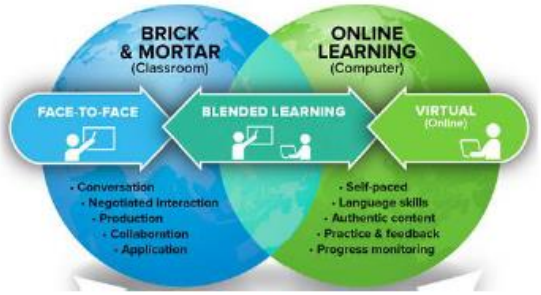


Figure 3: Blended Learning

Virtual or E-learning has been recognised as a vital tool by higher learning institution in Malaysia. In fulfilling the department's aspiration of producing competent human capital, ILSM has developed the e-learning platform namely as MySUL in 2018. The development and implementation of the system is using open source software which allows the users to exchange of information among users geographically dispersed, through mechanisms of synchronous (chats) and asynchronous communication (discussion forums). MySUL is web and mobile application which allow the department to:

i. Provide an online learning platform;
ii. Provide an online learning management;
iii. Uplift staff competency quality through online training evaluation; and
iv. Provide an integration platform with DTIMS.

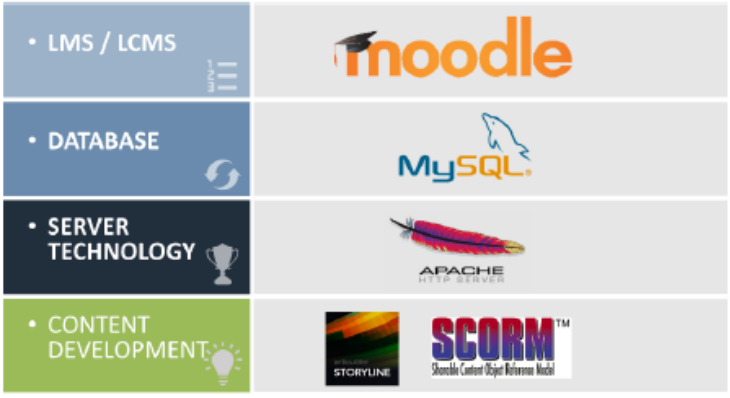


Figure 4: Technology Development

Figure 4 shows technology aspects involved in the development of the system. Modular Object-Oriented Dynamic Learning Environment (Moodle), open source learning management system (LMS) was used as a platform in developing the system. In term of database and server technology, the system

is using MySQL and Apache HTTP Server. In developing interactive training material, Sharable Content Object Reference Model (SCORM) and Articulate

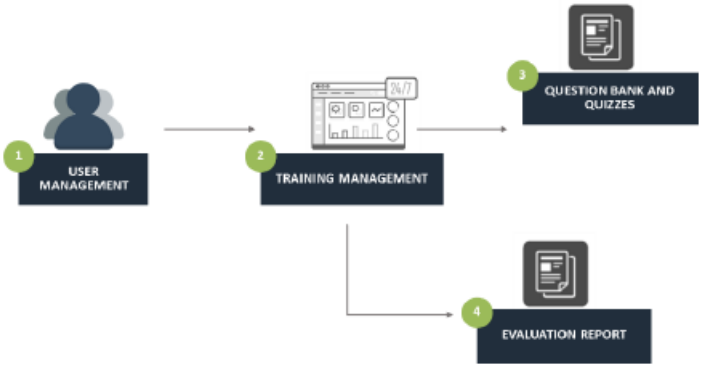


Figure 5: System Components

Storyline package were used to support the system operation.

The system was designed to cater the needs of DOSM's staff who will be the user for the system. In fulfilling this aspiration, the system comprises four main components such as **User Management, Training Management, Evaluation Report, Question Bank and Quizzes. (Figure 5)**

Figure 6: Ubiquitous Learning

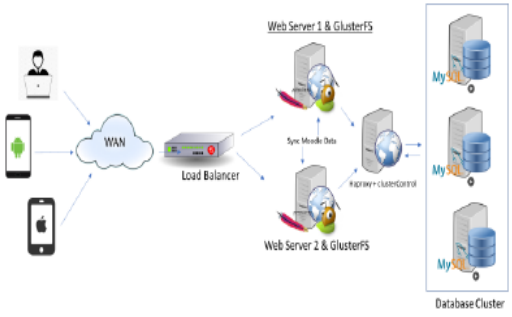


Figure 7: System Architecture

Figure 6 shows the ubiquitous learning approach was adopted in the system which means the system can be accessed in present, appearing or can be access everywhere through their device as long the user has a network (Figure 4). The system provide an interaction learning activities such as, discussion with an expert through forum or chat, interactive learning material and be a one stop centre for training material repository.

In the Figure 7 also shows the MySQL was used in the system as an open source relational database management system (RDBMS) associated with web applications and online publishing.

4. Conclusion

As numbers of staff and training volume have been increased in recent years, ILSM has made use of ICT to deliver its training programme through the development of DTIMS and MySUL. Both systems are useful and effectively supported the training management system by ILSM. The enhancement and improvement of both systems is an on-going process through system evaluation by stakeholders.

References

1. Middle-East Journal of Scientific Research 14 (11): 1471-1479, 2013. Ibrahim Almarashdeh, Nur Fazidah Elias, Noraidah Sahari and Nor Azan Mat Zain, Development of an Interactive Learning Management System for Malaysian Distance Learning Institutions.
2. CENTERIS 2012 - Conference on ENTERprise Information Systems. Carolina Costa, Helena Alvelos, Leonor Teixeira, The use of Moodle e-learning platform: a study in a Portuguese University.
3. Procedia - Social and Behavioral Sciences 191 (2015) 872 – 877. Nadire Cavus, *Department of Computer Information Systems, Near East University, Lefkosa 98010, Cyprus,* Distance Learning And Learning Management Systems.
4. User Requirement Specification (URS) 2017, DOSM Training Information Management System (DTIMS).
5. Function Design Specification (FDS) 2017, DOSM Training Information Management System (DTIMS).
6. User Requirement Specification (URS) 2018, Malaysia Statistical Ubiquitous Learning (MySUL).
7. Function Design Specification (FDS) 2018, Malaysia Statistical Ubiquitous Learning (MySUL).

Instrumental Variable Approach to Estimating the Scalar-on-Function Regression Model with Measurement Error with Application to Energy Expenditure Assessment in Childhood Obesity

Carmen D. Tekwe[1], Roger S. Zoh[1]*, Lan Xue[2]
[1] Department of Epidemiology and Biostatistics, Indiana University, Bloomington, IN, USA
[2] Department of Statistics, Oregon State University, Corvallis, OR, USA

Abstract
Wearable device technology allows continuous monitoring of biological markers and thereby enables study of time-dependent relationships. For example, in this paper, we are interested in the impact of daily energy expenditure over a period of time on subsequent progression toward obesity among children. Data from these devices appear as either sparsely or densely observed functional data and methods of functional regression are often used for their statistical analyses. We study the scalar-on function regression model with imprecisely measured values of the predictor function. In this setting, we have a scalar-valued response and a function-valued covariate that are both collected at a single time period. We propose a generalized method of moments-based approach for estimation while an instrumental variable belonging in the same time space as the imprecisely measured covariate is used for model identification. Additionally, no distributional assumptions regarding the measurement errors are assumed, while complex covariance structures are allowed for the measurement errors in the implementation of our proposed methods. We demonstrate that our proposed estimator is L2 consistent and enjoys the optimal rate of convergence for univariate nonparametric functions. In a simulation study, we illustrate that ignoring measurement error leads to biased estimations of the functional coefficient. The simulation studies also confirm our ability to consistently estimate the function-valued coefficient when compared to approaches that ignore potential measurement errors. Our proposed methods are applied to our motivating example to assess the impact of baseline levels of energy expenditure on BMI among elementary school-aged children.

Keywords
Childhood obesity; Energy expenditure; Functional data; Measurement error; Instrumental variable

1. Introduction

It is estimated that about 20% of the U.S. child population suffer from obesity and the percentage of childhood obesity has more than tripled in the last 40 years (CDC, 2017). The consequences of childhood obesity include reduced healthy physiological, behavioural and psychological development during childhood. Obesity in children and adolescents also leads to adverse health outcomes such as type 2 diabetes and cardiovascular diseases in adulthood. To combat this epidemic, targeted environmental and behavioural school-based interventions designed to increase physical activity among school-aged children have gained widespread interest. Examples of these school-based interventions include activity permissive learning environments and the use of stand-biased desks in classrooms (Lanningham, 2008; Benden,2011).

In a recent study, stand-biased desks were introduced to a Texas school district as a means of increasing school day physical activity. A research question of interest was to quantify the association between daily energy expenditure and subsequent progression toward obesity among children. The children were given accelerometer armbands to approximate their daily energy expenditure. Since the levels of true daily energy expenditure is not directly observable, it is calculated as a function of the observed physical activity behaviour from the devices. In this manuscript, we assume that the objective measures of energy expenditure obtained from physical activity monitors are prone to measurement error and develop a method of analysis that calibrates the measurement error and is easily applicable for assessing the effects of daily energy expenditure on 18-month change in BMI.

In determining the role of energy expenditure in obesity development among children, we consider the linear scalar-on-function regression model with a scalar-valued outcome Y and an imprecisely observed function-valued covariate, X(t). In this setting, X(t) is a latent function-valued covariate that is not directly observable. Instead, it is unbiasedly measured by W(t) prone to some measurement error. Linear scalar-on-function regression models extend classical regression methods to allow function-valued covariates with scalar-valued outcomes in regression settings and many statistical methods have been proposed to estimate the model (Silverman, 2005) when the covariate is measured with no or negligible error. In this paper, we propose a different approach to incorporate measurement errors and allow unspecified error structures. A function-valued instrumental variable belonging in the same parameter space as X(t) is used for model identification, and the generalized method of moments-based approach is proposed to consistently estimate the functional coefficient, $\beta(t)$, in the presence of functional measurement errors. Our proposed method for functional measurement errors do not treat the imprecisely observed function-valued covariate as longitudinal or time

series data. Rather, we consider the functional covariate as a single function that is used to estimate a latent variable such as true energy expenditure. Under our newly developed methods, estimation of the measurement error covariance is not required for parameter estimation. To the best of our knowledge, the use of function-valued instrumental variables in the functional linear regression model is novel. We illustrate the impacts of measurement error and covariance structures on the estimated parameters through simulation studies. With the increasing use of wearable or activity monitoring devices to study biological phenomenon in biomedical research, it is critical that statistical methods that allow their accurate and unbiased assessments be developed.

2. Methodology

We propose a generalized method of moments based estimator to estimate the function-valued coefficient of the functional linear regression model. In this setting, the outcome is scalar-valued, while the covariate, X(t) is a function. Our proposed method requires no distributional assumptions for the measurement errors. However, the estimation of the function-valued coefficient depends on the assumption that an instrumental variable exists in the data. Additionally, estimation of the covariance matrix for the measurement error is not required for the successful implementation of our proposed methodology. Under current functional data methodology, a naive estimator of the coefficient would be based on the observed measures and the outcomes, where the observed measures are treated as the true measures for the unobservable latent covariate. The strength of our proposed estimator is that while the function-value covariate might not be directly observed, estimation of its effect on the response is based on its unbiased measure as well as additional information provided in the data in the form of the instrumental variable.

3. Result

In this section, we describe the application of our methods to the motivating example. Students enrolled in the study were followed over an 18-month period. The study design was a cluster randomized trial where teachers within three schools in the College Station Independent School District were randomly assigned to receive either the treatment (stand-biased desks) or control (traditional desks) (Benden, 2011). The data contain measurements obtained at baseline and at the beginning of each semester over two academic years. An objective of the study was to investigate the relationship between energy expenditure behaviour at baseline and the 18-month change in body mass index (BMI) from baseline among the students. Thus, an outcome of interest was the difference or change in BMI values from baseline to 18 months

post follow up. The count of steps represents the number of steps taken over a given period of time and is an indicator of a subject's physical activity levels. Current guidelines for recommended daily physical activity levels are based on the duration of time spent in either moderate or vigorous intensity activity levels and number of steps per day (Matthews, 2012). For example, Tudor et al. (2004) indicated that activity levels of 12,000 steps/day and 15,000 steps/day for boys and girls, respectively were recommended for maintenance of healthy body composition for children between the ages of 6-12 years. While daily energy expenditure is defined as the total number of calories or energy used by the body to perform daily bodily functions.

Table 1 Descriptive statistics for the study sample at baseline (n=255). "Other"=Asians/Native Americans, EE= energy expenditure, SD=standard deviation.

Variable	Mean(SD)/ N(%)
BMI at baseline (kg_m2)	17.40(2.98)
BMI in Spring Year 2 (kg_m2)	17.55(3.18)
Average Step Counts/hour	13.16(11.51)
Average EE (kcal/hour)	1.2(0.41)
Age (years)	8.79(0.76)
Whites	174(68.24 %)
Blacks	34(13.33 %)
Hispanics	25(9.80 %)
Other	22(8.63 %)
Boys	132(51.76 %)
Girls	123(48.24 %)
Treatment	148(58.04 %)
Control	107(41.96 %)

In our application, energy expenditure and step counts were both collected per minute from the SenseWear Armband® (BodyMedia, Pittsburgh, PA) among the 374 children enrolled in the study who wore accelerometers while in school for one week at baseline. The children's body weight, height, age, and sex were all collected at baseline, while their BMI's were calculated at the beginning of each semester over the study period. True daily energy expenditure behaviour, X(t), was considered the latent covariate. The surrogate measure for $X(t)$ was the energy expenditure taken per hour obtained from the device, W(t). Step counts measured by the device was treated as the instrumental variable in this application, M(t). We assume that $cov\{X(t),M(t)\} \neq 0$ and $cov\{M(t),U(t)\} = 0$. Justification of the use of instrumental variables is challenging in practice. However, an instrumental variable may be based on a separate independent measure of X(t). In our application, both M(t) and W(t) were obtained from the same device. But their measured or calculated measures were obtained separately. The SenseWear Armband® obtained the step count based on a 3-axis accelerometer and pattern

recognition. While the calculation of total energy expenditure was based on heat flux, skin temperature, galvanic skin response, and anthropometrics (Lee, 2015).

To assess impacts of energy expenditure obtained at baseline on the difference in BMI values among the enrolled students, we first assumed that both W and M were discretely observed on a time interval [0, T]. On average, the students wore the devices for six hours on each school day during the week it was worn at baseline. Since the accelerometry data were collected per minute, we combined all the data for the week the device was worn and averaged all the minute-level data collected within the week to hourly-level data to reduce any potential noise associated with the data collection. Figure 1 provides the plot of Wi(t) and Mi(t) against time for all subjects included in the study. The grey lines illustrate the individual trajectories while the blue solid line is the smoothed mean for the observed energy expenditure and step counts among all the subjects. Two sets of analyses were performed to illustrate our developed methods. We first assessed the relationship between energy expenditure and BMI at baseline. The second analysis involved investigating the impact of energy expenditure at baseline on changes in BMI values at 18 months follow up. Due to loss of follow up or missing data, 255 and 156 students contributed to the baseline and the 18-month follow up analyses, respectively. The average BMI values at baseline was 17.4 kg/m2 (SD = 2.98) and 17.6 kg/m2 (SD = 3.2) during the spring semester of the second academic year. The mean step counts per hour at baseline was 13.16 (SD = 11.5) and the mean energy expenditure at baseline was 1.21 kcal/hour $(SD = 0.41), while the average age of the children at baseline was 7.9 years (SD = 0.80). About n=174 (68.24%) were whites, blacks n=34 (13.33%), Hispanics n=25 (9.8%) and others n=22 (8.63%). See Table 1 for details.

We provide the results from the baseline analyses and the follow up analyses in Figure 2. Plots of the estimated functional coefficient and the estimated 95% point-wise confidence intervals are provided in the figure. For assessments of the impact of energy expenditure on BMI at baseline, the bootstrap confidence intervals did not contain the zero line completely, indicating that the functional coefficient was not zero across the whole time space. Similarly, in determining the impacts of baseline measures of energy expenditure on the 18-month change in BMI over the study period, the estimated bootstrap confidence intervals did not contain the zero line completely. Because the function-valued coefficient was not completely zero across time, there was some statistical evidence of a relationship between baseline measures of energy expenditure and BMI values obtained at a future time, such as 18 months post baseline. Additionally, the relationship observed depended on both the level of energy expenditure and time.

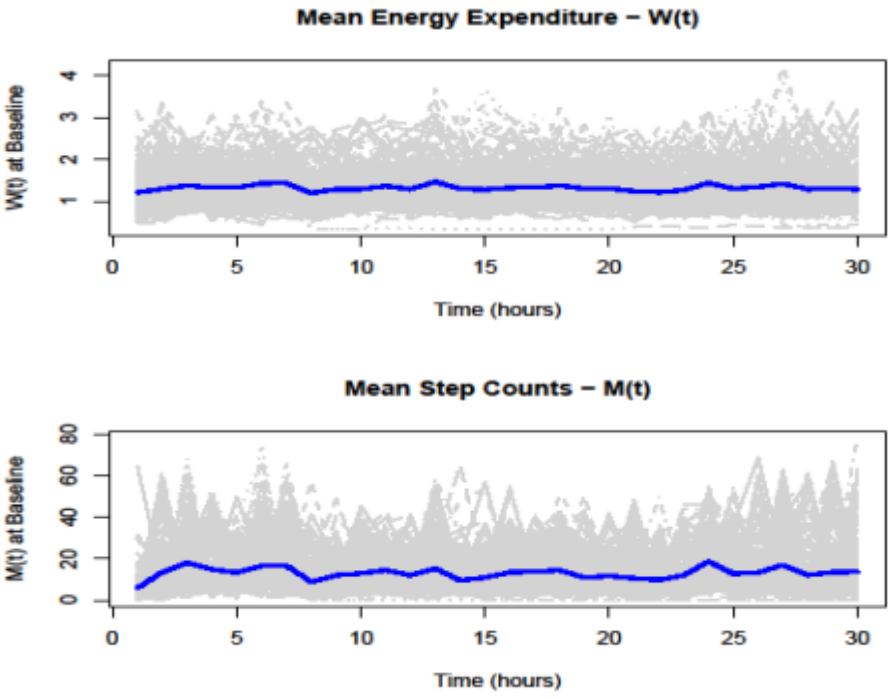


Figure 1: Plots of observed energy expenditure W(t) and mean step counts M(t) vs. time for all subjects at baseline from our motivating example. The figure confirms that the relationship between W(t) with time is nonlinear. In this setting W(t) is assumed to be an unbiased measure of X(t), while M(t) is an instrumental variable for X(t).

Impact of measurement error on the analyses

In addition to our method of moments-based instrumental variable estimator, we also obtained naive estimators of the effects of energy expenditure on BMI see Figure 2. As illustrated in both sets of analyses, the approaches obtained without accounting for measurement error appeared notably different from the estimators obtained from the instrumental variable based approaches. Based on Figure \ref{fig3}, the impacts of measurement error on both sets of analyses depended on time. While it is well known in simple linear regression models that the effects of measurement on estimation is to attenuate its effects towards zero, its impact in this functional linear regression setting is more complex. For both sets of analyses, we found that the measurement error adjusted function-valued coefficients tended to be larger than the naive coefficient. However, the naive estimate of $\beta(t)$ at baseline was found to be larger than the measurement error adjusted at the beginning and the end of the observational period.

4. Discussion and Conclusion

Trinh et al. (2013) recently studied the relationship between baseline energy expenditure and the threeyear change in BMI among 182 five to ten year old children with overweight and obesity health conditions in Australia. Using regression analysis and change in BMI Z-scores, the authors concluded that baseline measures of energy expenditure significantly impacted the three-year change in BMI among the children. However, our current results indicated that baseline levels of energy expenditure did have some statistically significant relationships on the future body weights among children, however, these impacts depended on activity levels and the time of activity. In this manuscript, we developed an instrumental variable approach for addressing potential measurement errors associated with function-valued covariates in scalar on function regression models. The developed methods can be used for assessments of the impacts of data collected on biological markers obtained repeatedly over a dense time space on health outcomes. A limitation of our current approach is that the instrumental variable must be collected on the same time period as the unbiased measure for the true covariate. Thus, the developed methods are applicable for devices that collect data on multiple biological markers over the same time period.

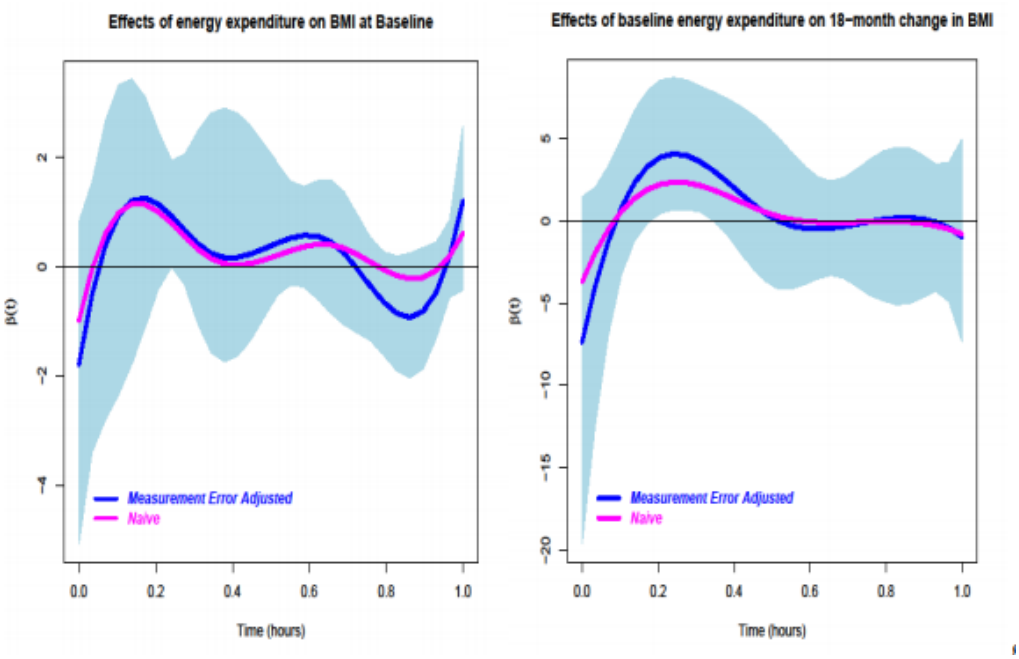


Figure 2: Plots of measurement error adjusted and naive estimates of $\beta(t)$ at baseline and also at 18 months. In (a), we estimate the effects of energy expenditure on BMI at baseline and in (b) we obtain plots of the effects of energy expenditure on 18-month change in BMI for the students included in our motivating example. The shaded regions are the 95% point-wise

Bootstrap confidence intervals, the blue line represents measurement error adjusted coefficients, while the pink line is the naive estimator that ignores potential measurement error.

References

1. Centers for Disease Control (CDC). Childhood obesity facts2017. https://www.cdc.gov/healthyschools/obesity/facts.htm.
2. Lanningham-Foster L, Foster RC, McCrady SK, et al. Changing the school environment to increase physical activity in children. Obesity. 2008;16(8):1849–1853.
3. Benden ME, Blake JJ, Wendel ML, Huber JC. The impact of stand-biased desks in classrooms on calorie expenditure in children. American Journal of Public Health. 2011;101(8):1433–1436.
4. Ramsay JO. Functional data analysis. Wiley Online Library; 2006.
5. Matthews CE, Hagströmer M, Pober DM, Bowles HR. Best practices for using physical activity monitors in population-based research. Medicine and Science in Sports and Exercise. 2012;44(1 Suppl 1):S68.
6. Mader JK, Feichtner F, Bock G, et al. Microdialysis. A versatile technology to perform metabolic monitoring in diabetes and critically ill patients. Diabetes Research and Clinical Practice. 2012;97(1):112–118.
7. Tudor-Locke C, Pangrazi RP, Corbin CB, et al. BMI-referenced standards for recommended pedometer-determined steps/day in children. Preventive Medicine. 2004;38(6):857–864.
8. Lee JA, Laurson KR. Validity of the SenseWear armband step count measure during controlled and free-living conditions. Journal of Exercise Science & Fitness. 2015;13(1):16–23.
9. Trinh A, Campbell M, Ukoumunne OC, Gerner B, Wake M. Physical activity and 3-year BMI change in overweight and obese children. Pediatrics. 2013;131(2):e470–e477.

How South Africa implemented a Smart Census

Nele Coghe
Hexagon Geospatial, Leuven, Belgium

Abstract

For many years the data collection for the census in South Africa was a manual process. Field workers used to receive paper maps to orientate themselves to their enumeration areas. This has been a tedious and complicated way of collecting data which required extra knowledge of map interpretation. With the improvement and democratization of technology, Statistics South Africa, the largest and arguably the most advanced national statistical office in Africa, benefits from the HxGN Smart Census solution. The HxGN Smart Census solution enables the use of imagery base maps in a web-based smart GIS application with predefined workflows that control and limit each user (including fieldworkers) to their allocated geographical areas and tasks. A mobile application, intelligent caching, data storage and backups make it possible for users, after only a limited amount of training, to have all the functionality required to do data capturing in the field without internet access.

Keywords

Census; GIS; Sustainable Development Goals; Statistics; Digital Transformation

1. Introduction

Technology is supposed to be a great equalizer. It is supposed to take the power out of the hands of the few and make it accessible to the many. It has made data available to millions, but in many cases the ability to collect, process, analyze, interpret, and present this data has remained with the few who have the domain and technology knowledge to understand it.

Remote Sensing and GIS are two fields that have been locked away. While the ability to use satellite imagery to analyze landcover and land use half a world away is fairly standardized and well-documented, access to the data has been restricted to only those few who can afford it. And processing the data is a technological barrier to entry: understanding the complex nature of satellite/aerial imagery capture and processing, GIS analysis, and geospatial analytics restricts the pool of potential users even further, including only those who have sufficient training and education. Added to this are the complexities of interpreting the data and understanding what is being communicated. Mapping and cartography are complex studies, and clear communication of the information is difficult.

Vast improvements have been made in these fields. The democratization of data through generous programs like Sentinel and Copernicus makes satellite imagery and radar data freely available to everyone. The Internet of Things and explosion of sensors (drones, webcams, video feeds, connected traffic sensors, etc.) have ensured that we have access to more data than ever before.

While the data itself is freely available, knowledge of the technology to convert the data to usable information is not. While we can get free radar coverage of every point on the earth every two weeks, processing that data is still complicated. We can unwrap radar data to find minute changes in the earth's surface (down to a millimeter), but the knowledge of those change detection algorithms and how to use them remain a black box – and there is no foreseeable end to that.

Even with the technology to perform change detection on satellite imagery, it still requires a human being to interact with the imagery and make a map. Frequently, this process can take weeks to months to complete, and by the time the imagery is processed, and a map is created, the information is out of date. Because making traditional maps takes so long and is so expensive, we try to make them do too much. Every map has to perform multiple purposes: land cover, land use, roadway mapping, and topography, to name a few. This compromises at least the intent of the map, if not the map's accuracy.

Information that can be derived from the data collected by all these sensors has great potential. What is needed is a way for domain experts to build sophisticated, reusable algorithms that can ingest streams of sensor data. We need a platform that allows data to be plugged in to the platform as soon as it is collected, and then have the system pull the data through processing steps so that analyses run automatically and generate updated maps. These maps would not only allow end users to see the current state of the land, but to see the entire time series so they can understand the patterns behind the change and begin to formulate predictions. Users need to see not only what was and what is, but also what can be.

This technology exists today. It is not a map; it is a Hexagon Smart M.App — a dynamic information service. By moving from the static map model — which collects data, analyzes it, and then produces a static printed, digital, or web-based map — to a dynamic information service, we can not only automate the process, but we can build job-specific and use-case-specific maps. Instead of multiple departments sharing a single multi-purpose map, each department can access its own view of the map containing information produced from the data specifically for them. Because this Smart M.Apps are lightweight and quickly produced, they are easy to prototype. Domain experts can build the map, incorporate feedback from users, and then make the map accessible to land use departments. As new data comes in, it is fed into the

system, and the map is updated, including all of the relevant analytics. It is time to stop using maps to communicate and to start using dynamic information systems that overcome the technological hurdles that keep people from using these powerful tools to analyze spatial data.

2. Methodology

A population census is the most important and costly statistical data collection exercise conducted by a national statistical office (NSO). It involves the total process of planning, collecting, compiling, evaluating, analyzing, and disseminating demographic, economic, and social data. This data is usually limited to a specified time and delimited area, whether that be an entire country or just a well- delimited part of a country. It is usually conducted every ten years, and the results provide a detailed, small-geographic-area snapshot of the demographic, socio-demographic, and housing status of a country. It also provides the basis for a wide range of sample surveys.

Statistics South Africa, the largest and arguably the most advanced national statistical office in Africa, now benefits from the HxGN Smart Census solution. HxGN Smart Census was developed on Hexagon Geospatial's M.App Enterprise platform. It combines traditional GIS functionality with a powerful workflow and workforce management tool to provide a total solution that covers all the phases of a census: pre-enumeration mapping, digital enumeration (including logistics, workforce training and management, integration with existing computer-assisted personal interviewing (CAPI) platforms, and dynamic progress reporting), as well as the dissemination of census results through dynamic Smart M.Apps. The traditional census has 4 distinct phases: Pre-Enumeration Planning, Enumeration, Post-Enumeration Processing, and Dissemination.

Pre-Enumeration Planning

In preparation for a population and housing census, the entire country is divided into small areas of land, each one small enough to be handled by one interviewer during the time of the census. This is referred to as pre-enumeration census mapping, or demarcation, and the resultant demarcated areas are called Enumeration Areas (EAs). In some countries, such as Sweden and Austria, door-to-door canvassing to collect census data has been replaced by a registration-based census. However, in some parts of the world, such as Africa, civil registration and housing registers are not complete or current. Therefore, a conventional census is an important source of information and will remain relevant, if not critical, for many years to come.

To conduct a door-to-door census, the census cartographer needs to provide the census enumeration team with a set of unique maps covering the entire country that accurately defines the boundaries within which each

interviewer (enumerator) has to work during the enumeration phase of the census. Furthermore, today's user community demands statistics to be provided within a spatial context. To facilitate this, desktop and server-based GIS solutions have become an important part of census data products and dissemination.

Censuses are by far the costliest statistical data collection project in a country. There is constant pressure on NSOs to improve efficiency while cutting the costs associated with a census. Traditional desktop GIS software, in combination with mobile GIS software, is the current standard in pre-enumeration census mapping. Since the costs associated with desktop and mobile GIS are determined by the numbers of users/licenses, it is usually significant – especially in countries where large numbers of temporary GIS operators and fieldworkers are deployed to do the work.

One license of HxGN Smart Census allows for unlimited users, resulting in significant cost savings. Although it does not have all the functionality of a high-end desktop and mobile GIS, HxGN Smart Census has all the GIS functionality required to do pre-enumeration census mapping and more. Raster and vector data are processed on the client side, rather than the server, which enables sophisticated GIS functionality, including vector data capturing, attributing, redlining, measuring, and querying, directly on the client. HxGN Smart Census therefore eliminates the need for numerous sophisticated and costly desktop and mobile GIS licenses.

HxGN Smart Census implies fewer desktop GIS licenses, but it does not imply replacement of the entire current GIS infrastructure that may exist at an NSO. Since it accesses any established spatial database server, such as Oracle Spatial or SQL Server, HxGN Smart Census can be used together with any desktop software, such as GeoMedia or ArcGIS. This is a huge advantage to NSOs where there is already an established GIS infrastructure and expertise on a particular platform. The status quo can either be retained with the addition of HxGN Smart Census to provide the increased software capacity required for the large temporary census workforce, or desktop GIS licenses that are no longer required can be replaced by HxGN Smart Census.

Although desktop and mobile GIS software have all the functionality required for census mapping, they have limitations when it comes to handling and managing project-specific workflows. This is a huge challenge which often requires use of a range of (usually unrelated) tools to establish and manage workflows. This leads to problems such as variances in interpretation of the methodology by different users and data integration issues that emerge as a direct result of using a set of disparate tools. Ultimately, it introduces unnecessary complexity that has a negative impact on data quality while increasing the overall risks associated with the project.

At its core, HxGN Smart Census has highly configurable rules and a powerful workflow engine. This is a key attribute of the software. This enables Subject Matter Experts (SMEs) to implement census mapping workflows and feature-level access control that is managed by a single, fully-integrated system where all the parties, each with associated posts and roles, access the same database(s) through a single web server. Each post is associated with one or more roles, and each role has a specific set of tasks. Furthermore, access to the system is through standard internet/intranet security protocols involving a username and password.

The access control system allows a supervisor to allocate a specific unit of work, referred to as a production unit (PU), to a particular user associated with a specific post. The assigned user can only work on the assigned PU and can only execute the tasks relevant to his/her role in that particular part of the overall workflow. This eliminates duplication of efforts, since it is impossible for two teams to work on the same PU at the same time. It is also impossible for a user to do anything other than the tasks related to his/her role with regards to the allocated PU at that particular step in the overall workflow.

Once the tasks for a PU are completed, the PU is submitted, and the supervisor receives a notification. This then triggers a set of quality control/quality assurance (QC/QA) steps, after which the work is either accepted or rejected and sent back for correction.

Streamlined Project Management

A census-mapping project is usually a huge undertaking, with a large workforce ranging from about 50 to more than 1000 persons, depending on the size of the country, the methodology, and the project timeframe. Project management is therefore a huge challenge. HxGN Smart Census enables the GIS manager to manage the entire census mapping project using a single tool. Since work allocation and task execution are performed on the same system, the GIS manager knows exactly who is doing what at any particular moment and, also, the overall status and progress of each phase of the project. And since HxGN Smart Census has a strong spatial component, work scheduling and progress tracking can be done using a combination of maps, tables, and lists.

This effectively manages, if not eliminates, most of the inefficiencies common to census mapping projects. These inefficiencies include: suboptimal scheduling of fieldwork that requires field teams to drive unnecessary distances between PUs, duplication of effort where the same PU is allocated to two teams or operators, and obtaining of status reports from the field and office and compiling of progress reports. HxGN Smart Census provides the ability to detect a problem swiftly – long before it turns into a major crisis.

Because numerous teams often work simultaneously on a census-mapping project, it is a huge training and management challenge to ensure consistency in the interpretation and execution of the methodology. HxGN Smart Census vastly improves the consistency of work, and this ultimately improves the overall quality of the data collected and processed by the respective parties in the workflow.

Every user/role executes a portion of the workflow with specific tasks configured in the software as required by the project. Each user can only execute the tasks associated with that particular part of the workflow – nothing else. The software literally guides the user through the steps. Only the software functionality required for execution of that particular workflow is available to the user. This ensures that all users perform the tasks in the same way, which leads to consistency in the execution of the overall methodology and, ultimately, increased data quality and uniformity.

Enumeration

The Enumeration phase, while quite short – usually two to three weeks – also requires the largest workforce and the most coordination. Over 200,000 people can work simultaneously collecting the enumeration data. To mobilize such a workforce requires a number of solutions.

Figure 1: Field interviewer collecting census data via a mobile device in South Africa

Since each user only has access to software functionality required to execute the tasks for the assigned portion of the workflow, extensive GIS skills and knowledge are not required to do the work. This is a huge advantage in Africa, where GIS skills among the population are relatively limited. Furthermore, workforce training can be much more focused and task-specific, resulting in significant savings in the training budget and quicker project

implementation. Simplified training also makes it much easier to replace staff during the project when required.

Many NSOs use third party software to provide the Computer-Aided Personal Interview (CAPI) platform. These third-party solutions utilize mobile devices and laptops to provide the interviewer with a list of questions and then records the results. The mobile component of Smart Census integrates with these third-party solutions giving the NSO the power to choose the provider that is right for their needs.

Census mapping software applications for fieldwork must provide the ability to work offline. In addition to openly connecting with third-party solutions for digital collection, HxGN Smart Census uses advanced caching to enable offline use in remote areas where internet access is limited or non-existent. If required, a field interviewer can work offline for days as in Figure 1, updating the captured data at intervals when connectivity can be restored.

With such a short window to collect all of the demographic data, the NSO needs to be able to react to changing needs as quickly as possible. With HxGN Smart Census, the data is updated on-the-fly, giving the planning center instant insight into where response may be necessary. Injuries, weather, technical difficulties, and other emergencies can all be mitigated by understanding the situation quickly and then adjusting the workforce as needed.

Processing

Because the entire project is digitized from start to finish, there is no need for a manual data processing step. This eliminates the bulk of the errors – from transcription or illegibility errors to data entry mistakes – from the process. It also eliminates the long lag time from capturing the data until it is processed and ready for analysis.

Post-Enumeration

One of the keys for census data is making it available to the stakeholders as soon as possible. They need to understand the makeup and needs of their citizens immediately, so they can begin planning and making policy. Census data is displayed in graphics, charts, and on a map simultaneously. This interactive Smart M.App as in Figure 2 lets stakeholders view both comprehensive statistics and detailed information about specific communities or areas. Results can be filtered based on multiple factors such as housing type or socio-economic status.

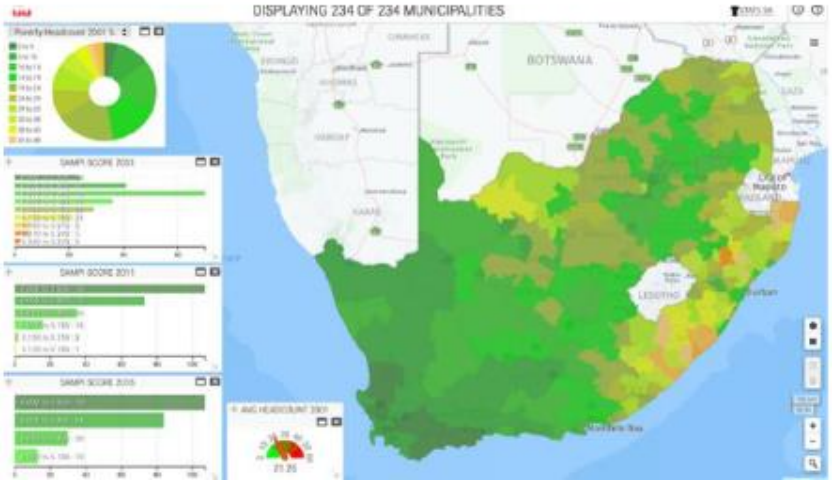


Figure 2: Example of an interactive Smart M.App

Once the information from the census is made available, the NSO may discover a need to gather more information about a specific pattern illuminated by the census data. Instead of having to wait 10 more years, the HxGN Smart Census infrastructure is already in place and can be used to plan, perform, and disseminate the data for post-censual surveys. These surveys can be done on any scale, whether national or limited to a specific area within the country.

3. Result

HxGN Smart Census provides an exciting solution to census mapping in South Africa. It provides an excellent platform for easy scheduling of work, project tracking, and reporting. It is easy to use and requires minimal training compared to conventional desktop GIS. The flexibility to configure the software to ensure that each user can only access tasks assigned to him/her eliminates duplication of effort and potential differences in methodology interpretation. This improves data quality and consistency.

Since software is installed only on the server, software/application updates are easy and non-disruptive. Each time a user logs in from a client, any updates or changes are automatically reflected. The offline functionality enables fieldwork to be done in remote areas. Advanced caching ensures that relatively small data packages are sent to and from the server, keeping data costs and connection time to a minimum — an important aspect from a fieldwork and budget perspective.

Flexibility of the software enables it to be used for planning and management of all phases of the census project. It can also ultimately be used as a tool to disseminate information to all stakeholders, including the public. HxGN Smart Census complements, rather than replaces, any existing GIS

infrastructure; however, it eliminates the need for mobile GIS software, and it reduces the need for high-end desktop GIS software.

4. Discussion and Conclusion

As we can see from the HxGN Smart Census case study, the challenge facing democratization of remote sensing and GIS workflows lies not in access to data, but in access to the processes that create the data and the information derived from the data. In this case, democratization of the data is not achieved by free access to satellite data, but rather by empowering NSOs to access technology that can make their processes more efficient, ultimately allowing them to produce more accurate data more quickly. By implementing a streamlined workflow system, the HxGN Smart Census solution empowers workforces of any size to be quickly trained and to focus on their particular duties instead of navigating through a complex and perhaps overly-powerful GIS system.

Under the hood, of course, there is a lot of specialized processing going on. But that is ameliorated by keeping it under the hood and limited to SMEs who can set up the processes and workflows, and then step back and allow the data to flow into the system. This allows SMEs to focus on QA/QC tasks instead of building map after map after map.

In addition, HxGN Smart Census ensures adherence to standards, simplifies the planning process, and automates report generation at the end of each cycle. Not only does this make it easier on the NSO, but it also makes this data more accurate and more accessible. This higher-quality data can be more quickly disseminated to stakeholders for their use in understanding the makeup of their country or region so they can better discern what is needed to shape smart change in their world.

How Resilient is Indian banking system? – IBS Perspective

Avijit Joarder, Ashis Nayak, Swapan-Kumar Pradhan[1]
Reserved Bank of India
Monetary and Economic Department, BIS, Basel, Switzerland

Abstract[2]

The paper focuses on specific aspect on resilience of the Indian banking system based on the insight of International Banking Statistics (IBS). We analyse long-term developments in overall business by the banking system, international business with banks' counterparties in India and foreign countries, and capital inflow to as well as capital outflow from Indian non-bank sector vis-a-vis banks abroad in various countries. We extensively use several financial statistics compiled by the Reserve Bank of India (RBI) and the Bank for International Settlements (BIS) to construct a number of indicators that capture changes in business trends. Finally, we use the indicators based only on IBS data to compare Indian context with selected Asian countries to establish that our proposed indicators could be utilised to monitor developments in Indian or similar banking system and identify potential vulnerabilities that could lead to stress or crisis.

Keywords

Bank Lending, Cross-border capital flows, Banking Crisis, Foreign Exchange Reserves

1. Introduction

In the aftermath of the Indian Balance of Payment (BoP) crisis during early 1990s, Indian economy began to liberalise with maximum impact of deregulatory policies was felt in its banking sector. In 1991, the Narasimham Committee appointed by the Government of India (GoI) suggested remedial measures relating to the structure, organization, functions and procedures of

[1] Avijit Joarder (ajoarder@rbi.org.in), Assistant Adviser, DICGC, RBI, Mumbai, India; Ashis Nayak (anayak@rbi.org.in), Research Officer, Department of Statistics and Information Management, RBI, Mumbai, India; Swapan-Kumar Pradhan (Swapan-Kumar.Pradhan@bis.org), Senior Statistical Analyst, Monetary and Economic Department, BIS, Basel, Switzerland

[2] The views expressed are those of the authors and do not necessarily those of the RBI or the BIS. The authors like to thank Stefan Avdjiev (BIS), Abhiman Das (Professor, IIM, Ahmedabad, India), Maximilian Jager (University of Mannheim), O. P. Mall & Anujit Mitra (Advisers, RBI), Madhusudan Mohanthy (BIS), Simone Saupe (Swiss National Bank) and Philip Wooldridge (BIS), for their helpful comments and suggestions ***on the detailed version of this paper***.

the financial system[3]. Banking reforms gradually transformed Indian banking landscape into level playing fields. In addition, the process enabled the banking system as stable and prosperous financial entities not only in India but also in foreign countries (Banga and Das [2012]). Foreign inflow and outflow of capital were not limited to banking sector also but also to non-bank sector as well. The non-bank sector remained active in placing their funds (deposits) with banks in foreign countries and foreign banks too built up confidence in lending (loans) to Indian non-bank sector.

While Indian banking system was recovering by mid-1990s, the Asian Financial Crisis (AFC) hit in the South East Asian countries. This led to loss of demand and confidence throughout the South-East Asian region. While the global impact of the Asian crisis was not possible to judge, it led to increased global attention to the health of financial institutions, particularly banks, and it was felt necessary to collect relevant data to help monitoring overall global situations. India and several other developing countries as well as important offshore financial centres were encouraged by the BIS to collect and compile international claims and liabilities of internationally active banks. Therefore, a concerted effort was made to improve the timeliness, frequency and coverage of the BIS international banking statistics. This led to increase in the global coverage of the BIS statistics.

The Global Financial Crisis (GFC, 2007-2009) underscored the need for granular data on funding and lending activities of both globally and regionally important banking systems. In a number of recent articles of post-Asian crisis developments, researchers at the BIS concluded that countries affected during the Asian crisis have gradually improved their financial conditions (Avdjiev et. al. [2018]). During past 20 years the cross-border claims of BIS reporting banks on the emerging Asian countries more than quadrupled, totalling $2 trillion in end 2017(Koch and Remolona [2018]). The latest BIS article in December 2018 Quarterly Review stated that over the past decade, the cross-border activity of banks from emerging market economies (EME) has been growing at a faster pace than that of banks from advanced economies, mainly driven by increasing EME-to-EME interlinkages (Eugenio, Koch and Pradhan [2018]). It has been reported that as of end-2017, Indian banks located abroad in the LBS reporting countries had lent 84% or $43.7 billion out of $52 billion, to non-resident counterparties and rest 16% from those located in India.

In view of the above background, we explore the RBI and the BIS statistics offering perspectives on following aspects:

- Longer-term development in business by scheduled commercial banks in India, including business by Indian-owned banks in foreign countries.

[3] M. Narasimham, the former Governor of RBI (from 2 May 1977 to 30 November 1977)

- Nature of international exposures of Indian banking system.
- Cross-border capital inflow and outflow of India through banks in foreign countries, with a special focus on foreign assets of Indian non-bank sector.
- Longer-term behaviour of certain indicators in Indian context compared to select Asian countries viz., Thailand, Indonesia, Malaysia, Philippines and Korea[4].

2. Methodology

In order to analyse long-term developments in Indian banking business, we rely on balance sheet of Scheduled Banks comprising both Indian and foreign owned banks. The RBI compiles such data and publishes on its website. The balance sheet data also provide breakdown of business by foreign affiliates of Indian banks. A subset of Scheduled Banks also reports international banking business, and we consider such statistics to analyse cross-border banking business by banks in India with counterparties in foreign countries. The IBS of India is in fact a part of global initiative by the BIS that among others compiles international banking statistics of banks operating in about four dozen countries around the world. The IBS statistics are useful to analyse exposure of international banks in foreign countries with their counterparties (banks and non-banks) in India. In addition, we use country-level data on foreign exchange reserves, GDP and the BIS international debt securities (IDS) to measure and understand banks' behaviours at different periods during the past 30 years.

It is widely acknowledged that the BIS international banking and financial statistics are the best source for analysis of cross-border capital flows. In particular, the locational banking statistics (LBS) measures gross cross-border claims and liabilities, including inter-office positions, of banking offices resident in a reporting country with their counterparts in other countries. The LBS also constitutes the key source of information on the instruments, currency, bank nationality, counterparty sector and counterparties' geographical composition of balance sheets positions. On the other hand, the consolidated banking statistics (CBS) measure worldwide-consolidated claims of banks headquartered in reporting countries, including claims of their foreign affiliates but excluding inter-office claims. In addition to LBS and CBS, the IDS statistics is a security-by-security data set built by the BIS using commercial data provider's information. The IDS are issued under international law or the debt securities that are issued outside the local market of the country. The statistics capture euro bonds and foreign bonds and

[4] We did not consider other jurisdictions such as Hong Kong SAR and Singapore as they are different in terms of financial activities (Maria [2009], Raymond and James [2004])

exclude negotiable loans. The IDS statistics are aggregated, among others, by sector of issuer, currency, and nationality and residence of the issuer.

Finally, we use Probit regression model to check and predict the warning signals of crises based on share of short-term international claims to foreign reserves, share of long-term international claims to GDP and GDP growths for six countries of our interest (India, Indonesia, Korea, Malaysia, Philippines and Thailand).

3. Findings

3.1: Longer-term development in business by Scheduled Commercial Banks (SCBs) in India

In order to understand long-term developments and emerging changes in Indian banking system since early 1990s, we examine aggregated balance-sheet positions of Indian SCBs comprising both domestic and foreign banks. While banks adjusted their business model over the three decades to cope with developments in domestic and international financial markets, the gross amounts of claims and liabilities have grown exponentially. As expected, deposits and advances are respectively the highest contributors to liabilities and assets. When measured as a ratio of country's GDP, total assets and total liabilities fell from 49% in early 1990s to 42% in 1996 but continued to grow thereafter to reach at 84% in 2018. It is noticeable that the growth in banking business did not stop during the AFC as well as during the GFC.

The changes in portfolio of assets and liabilities shows that the share of other assets has decreased significantly from 18.5% in end-March 1990 to merely 5.8% in end-March 2018. Similarly, the share of cash in hand and balances with the RBI also decreased by over 7% from 11.4% to 4.8% over the same period. These shifts in portfolio of assets during the period allowed banks to make more loans and advances, the share of which increased from 43.5% to 57.3%. On the liabilities side, other liabilities declined from 19.3% to only 4.6%. This shift resulted significant increase in reserves and surplus from below 1% to 7.1%, as well as increase in share of deposits from 70.6% to 77.3%. The primary reasons for these shifts are regulatory changes to resilience of Indian banking system. It is interesting to note that share of overall equity capital nearly remained the same at about 0.8% of total liabilities.

3.2: Nature of international exposures: international claims and liabilities of the SCBs

In comparison to total claims and liabilities of SCBs, the share of international claims and liabilities declined over years. While total assets and liabilities increased over years from ₹13 trillion in 2001 to over ₹140 trillion in 2017, the shares of international claims and liabilities in the total assets and total liabilities respectively declined over years, from 7.3% in 2001 to 3.9% in

2017 for claims and from 11.0% in 2001 to 8.7% for liabilities. The cross-border claims/liabilities in all currencies and local claims/liabilities in foreign currencies into sector non-banks and others (mainly banks) are asymmetrical. As of end-2017, the share of international claims on non-bank sectors was 70% whereas the same for international liabilities was 81%. The total claims and liabilities are the sum of world-wide consolidated balance sheet positions of Indian domestic banks, including operations of their foreign affiliates, plus unconsolidated total positions of foreign-owned banks for their operations in India. The domestic positions, i.e. positions vis-à-vis residents of Indian in Indian rupees, are thus assumed as the differences between total positions and international positions.

The country breakdown of international positions suggested that counterparties in the United States dominate share for both claims and

[4] We did not consider other jurisdictions such as Hong Kong SAR and Singapore as they are different in terms of financial activities (Maria [20091, Raymond and James [20041).

liabilities in recent years. In 2017, share of banks' foreign currency claims at home on residents stood at 28.3%, whereas share of cross-border claims on United States (US) stood at 36.7% of total international claims. Cross-border claims on United Kingdom (UK) declined over the years from 16.8% in 2001 to 6.1% in 2017. On liability side, the share of cross-border liabilities towards US was the highest (22.4% of total international liabilities, as of 2017) among other counterparties. Share of liabilities towards UK declined over years from 13.2% in 2001 to 10.5% in 2017. Share of liabilities towards United Arab Emirates slowed down before GFC and recovered slowly after GFC and stood at 19.2% in 2017.

3.3: Cross-border capital flow – counterparties in India versus banks in foreign countries

India has been the beneficiary of the capital flows as a promising investment destination and the GoI as well as the RBI are committed to continue to press the advantage by ushering in necessary reform measures (Mundra [2010]). In this context, we examine money inflow to India and money outflow from India through the banks in foreign countries.

Claims of Indian residents on foreign banks (i.e. outflow from India) increased from 6 billion US dollar (USD) in end-March 1990 to nearly 73 billion USD in end-2017. In early 1990s, the share of outflow from non-bank sector was above 60% of total but continued to fall to 13% in early 2006. The GFC triggered the rise with most funds directed mostly towards banks in the US, Offshore centres and UK, reaching the share close to 50% by mid-2011. The outflow from non-banks to banks located in UK, Euro area and offshore centres began to decline and reached at 12% by end-2017. On the other hand,

the share of outflow from banks was never below 34% and reached at 78% of the total by end-2017.

Liabilities of Indian residents to foreign banks increased from 13 billion USD in end-March 1990 to nearly 200 billion USD in end-2017. It means that lending banks in other countries are gradually getting more confident to put their money in India. In terms of sector breakdown, the share of inflow to non-banks has always been less than share of inflow to bank sector. Unlike bumpy outflow of funds, inflows to India from banks in foreign countries increased steadily over time, except for a brief period during end-2008 to end-2009. Nearly 40% of total inflow is routed through banks in offshore centres. It clearly stands out that non-bank sector attracted the largest share until end-2002 and thereafter inflow to banks steadily increased to current share at 53% of total inflow to India. The steady decline in inflow to non-banks from banks abroad is mainly attributed to alternative low cost sources of funds (e.g. issuance of euro-bonds).

Inflow of money from foreign banks is nearly 3 times more than money outflow. International banks in different countries play important role for capital inflow and outflow. As of end 2017, nearly 40% of total inflow (liabilities for resident Indians) are from international banks located in offshore centres, and another 40% from banks in US, UK, Switzerland, Japan and Australia. On the contrary, about one-third of assets (outflow from India) are placed with banks in offshore centres and about 40% with banks only in US and UK.

In terms of currency breakdown, USD denominated assets and liabilities remained the major currency of transactions with banks in foreign countries. Majority of these are accounted by deposits and loans. In case of outflow from India, Euro and British pound are other two preferred currencies for deposits abroad. In the case of inflow to India, Japanese yen and Euro are the other two preferred currencies for loans from banks in foreign countries.

3.4 Cross-border capital outflow from India non-bank sector to banks in foreign countries

There have been popular perceptions that Indians keep money with banks in Switzerland, but the BIS statistics reveal that money placed by non-bank sector of India with banks in Switzerland have reduced over time to negligible in terms of both size and share in such deposits. We conclude in the following analysis that the term "money in Swiss banks" from Indian non-bank sector is probably a generic term to mean foreign destinations. As we do not have any information on sources (of income) deposits, we only look at the destinations of total outflow from Indian non-bank sector to banks in foreign countries. The coverage of BIS locational banking statistics increased over time from 51%

in 1977 to 93% in 2016[5]. We thus assume that overage of total outflow (assets abroad) is at least above 90%.

At an aggregate level, nearly 100% of outflow from bank and non-bank sectors to banks in foreign countries was in deposits. The share of deposits with banks, however, sharply fell to about 78% during 2008-2009 before raising again to above 90% in subsequent years. As a result, we assume total outflow was in the form of deposits and those in debt securities and others are negligible. Another reason for such an assumption is that bilateral data on deposits with banks abroad in different reporting jurisdictions are not freely available on the BIS website. Further, in order to ensure data confidentiality, we used 4 quarters average in respective years for amount outstanding as well as for share of deposits. We noticed that traditionally non-bank sector of India has been placing their deposits with banks in UK, US and offshore centres rather than in Switzerland. The UK has been the favourite destination having the share above 20% in the past. After GFC, the banks in the US increased their share from 9% in 2008 to 15% in 2017. On the contrary, money with Swiss banks have started to fall continuously from its peak share of 35.3% in 1990 to 1.7% in 2017, and in fact started to fall almost in the same year of 2008 with increase with banks in the US. In terms of amount outstanding, such deposits was at 3,824 million USD in end March 2007 and reduced merely to 285 million USD in end 2017. The banks in offshore centres (Hong Kong, Singapore and 10 other centres) and those in developing countries (China, Russia, Malaysia, Chinese Taipei and 9 other countries) gradually picking up their share of deposits from non-bank sector in India. While the total outflow from non-bank sector significantly reduced since 2012, Switzerland is not among the top favourite destinations.

3.5: Comparison of longer-term behaviour - India versus selected five Asian countries

In 1997, India had several capital account restrictions that prevented inflow of short-term portfolio investments (often called "hot money") flowing into the country. Active capital inflows are a dual-edged sword for the recipient countries. On the positive side, capital flows support economic growth and provide welfare gains by financing productive investment opportunities and consumption smoothing. On the negative side, capital surges tend to bring inflationary pressures making the economy more vulnerable to external shocks. Prior to the GFC (2008), emerging markets in general appeared insulated from developments in the US and at the advent of GFC, emerging markets responded very strongly with policy measures to further insulate

[5] Global coverage estimated is published at https://www.bis.org/statistics/lbs_globalcoverage.pdf

themselves from deterioration in the US economy (Dooley and Hutchison [2012]). In the case of India, external debt did not change much from 2007-08 to 2008-09 (Viswanathan [2009]). Indian banking sector had negligible impact from the GFC due to no direct exposure to the sub-prime mortgage assets or to the failed institutions. The limited off-balance sheet activities or securitized assets resulted safe and healthy behaviour of the banking system (Subbarao [2010]).

Over the years banks and other financial institutions in our selected sample (India, Indonesia, Korea, Malaysia, Philippines and Thailand) are being gradually more integrated with the global financial system. The outstanding international debt securities of the financial institutions (comprising banks and other financial corporations) by country of issuers' nationality and by country issuers' residence prove that compared to other Asian countries in our sample, Indian financial institutions are increasingly issuing debt securities outside the home country is a reflection of global integration.

The forex exchange (FX) reserves of India rose from around 5.6 billon USD in 1990 to a very comfortable level of about 410 billion in end 2017, in a span of about 30 years. The portfolio is highly dominated by foreign currency assets (about 94% as of end 2017) and mostly denominated in USD, which hit the low at 23.4% in end 1990 during BoP crisis and now at 95% of total foreign currency assets. The significant increase of FX reserves hail to provide stability to Indian economy. In this context, we examine if the Indian forex reserves have been historically sufficient to meet short-term international claims on India by banks in foreign countries. Normally, low share of short-term international claims on a country compared to its FX reserves is the sign of stable financial system because at the time of a crisis foreign investors (including banks) moves money from volatile to stable market. As an evidence, the short-term international claims on India with residual maturity of less than one year was as high as 93% of India's FX reserves during the BoP crisis (1989-1990). Since the crisis, the share of short-term international claims fell sharply below the level of 10% by Q1 2003 but currently remain at around 20%. While the pace of international claims on India slowed down since 2013Q2, the level of FX reserves continued to rise, more than double of the size of international claims.

We look at comparative status of India compared to five other Asian countries that during the AFC were most affected (Thailand, Indonesia, Malaysia, Philippines and Korea). The international claims were raising and most of these claims were of short-term nature i.e. of less than 1 years of residual maturity. Such claims on India were the lowest compared to other five countries. After the crisis, international claims again started to increase. In the event of GFC during 2008, the international banks began increased lending again to Asian and other developing countries. Such type of lending peaked

in 2008 for Korea and then continued to fall but continued to increase for India, Indonesia, Malaysia, Philippines and Thailand.

The above mentioned Asian countries except India experienced a rapid expansion of total international claims on them from foreign banks. Ahead of the AFC, the short-term international claims, in particular, on these countries were booming. The most notable example is Thailand, whose share of short-term international claims reached to an unprecedented level of 71% of total international claim as of end-December 1993. Although it is not in the case of Thailand only, the pre-AFC expansions in total international claims including total short-term international claims on all remaining four countries were also very sizable. In the case of India, both total international claims and total short-term international claims followed similar upward trend in pre and post AFC. Among selected Asian countries, the share of short-term international claims on India was the lowest during the AFC. In terms of international claims as a share of GDP, India remains still at the lowest at below 7% and short-term international claims also are much lower (around 20%) compared to FX reserves.

3.5: Probit regression for prediction of banking crisis[6]

We finally aim at predicting banking crisis (warning signal) based on short-term international claims as a share of country's foreign reserves, long-term international claims as a share of country's GDP and growth in GDP. We consider consolidated international claims on a country by non-resident foreign banks, sourced from the BIS. In our analysis, for six countries (India, Indonesia, Korea, Malaysia, Philippines and Thailand) of our interest, we use the start of crisis years from the IMF Working Paper (WP/12/163)[7]. The data on foreign reserves and GDP at current price (USD equivalent) are sourced from World Economic Outlook (April 2018), all on annual basis.

The results for the probit models are estimated by using lagged variables to predict crisis over the period of 1983 to 2017. The results of the probit regression, are in line with the hypothesis of an overheating economy that grows too fast and therefore collapses during its growth path. Adding the GDP growth as a control variable does not influence sign or significance of the core variables of interest, and the probability of 0.0000 associated LR statistics = 1,065.21 rejects the null hypothesis that coefficients of all variables are simultaneously equal to zero. The GDP growth with one-year lag appears to be positive and significant showing that a crisis is most likely during an

[6] We thank Philip Wooldridge (BIS) for suggesting us to use Probit model and also thank Maximilian Jager, Centre for Doctoral Studies in Economics, University of Mannheim for his inside on probit/logit models.

[7] Systemic Banking Crises Database: An update (Table A1 on Banking Crises dates and Costs, 1970-2011), 1 June 2012 by Fabian Valencia and Luc Laeven.

economic upswing instead of a stagnation or a downswing. In other words, if we exclude the growth in GDP as a control variable the qualitative results of the both models are not influence, and the probability of 0.0000 associated LR statistics = 164.59 rejects the null hypothesis that coefficients of all variables are simultaneously equal to zero.

We conclude that short-term debt to foreign banks compared to foreign exchange reserve and long-term debt to foreign banks compared to GDP, could jointly serve as indicators for possible banking crisis at least years in advance. These are not only the indicators to monitor but are worthwhile to consider among other macro-economic and financial indicators.

4. Discussion and Conclusion

Starting from aftermath of Indian BoP crisis (in early 1990s), we covered two subsequent crises namely AFC and GFC to find out the impact of crises on domestic and international banking business of banks. The comparison of most affected Asian countries in the region is useful to understand symptoms before, during and after the crisis. In the detailed version of our paper, we demonstrated through visual analysis (graphs) and regression analysis (probit) that Thailand, Korea, Malaysia and Indonesia were severely affected by AFC, whereas Philippines was slightly less affected.

On the other hand, there was virtually no effect of AFC on Indian economy. In case of GFC, six countries in the region were not much affected. The RBI banking statistics together with the BIS statistics are very useful to monitor banking system and to get signals of overall developments in domestic and international markets, especially the inflow and outflow of money through banking channels. The analysis shows that not only foreign investors including international banks finds India as one of the safe destination for international investments but also banks and non-banks in India gradually increased their transactions with international banks in foreign countries. The signals from these statistics could thus be effectively utilised to monitor overall developments in financial markets and take corrective actions to avoid stress or crisis.

Further analysis could be considered to understand the signals arising from other macro-economic factors and also the change in lending behaviour of major creditors (e.g. bank nationalities) to India and other countries in the region. Similarly, it would be interesting to understand the influencing factors leading to change in behaviour of banks and non-banks of a country in the region with their counterparts in foreign countries.

References

1. Avdjiev, Stefan, Bat-el Berger & Hyun Song Shin (2018): Gauging pro-cyclicality and financial vulnerability in Asia through the BIS banking and financial statistics, BIS WP No. 735, July 2018.
2. Cerutti, Eugenio, Catherine Koch and Swapan-Kumar Pradhan (2018): The growing footprint of EME banks in the international banking system, BIS Quarterly Review, December 2018, pp 27-37.
3. Jung Changoon and Clark Cal (2010): "The Impact of the Asian Financial Crisis on Budget Politics in South Korea", Asian Affairs, Vol.37, No.1 (Jan-Mar 2010), pp 27-45.Koch, Catherine and Eli M Remolona (2018): "Common lenders in emerging Asia: their changing roles in three crises", BIS Quarterly Review, March 2018, pp17-28.
4. Mundra S S, (2016): "Financial stability in a weak global environment", 7th SEACEN High Level Seminar, Mumbai.
5. Rashmi Banga and Abhijit Das (2012):"Twenty Years of India's Liberalization–Experiences & Lessons" –UN Conference on Trade & Development(UNCTAD),Centre for WTO Studies, UN.
6. Subbarao, Duvvuri (2009): "Impact of the Global Financial Crisis on India Collateral Damage and Response", Speech delivered at the symposium organised by the institute of International Monetary Affairs, Tokyo on February 18, 2009.
7. Viswanathan, K. G (2010): "The Global Financial Crisis and its Impact on India", Journal of International Business and Law, Volume 9, Issue 1, Article 2.

What can data science do for economic statistics?

Louisa Nolan, Jeremy Rowe, Steven Hopkins, Sonia Williams
Data Science Campus, Office for National Statistics, UK

Abstract

The Data Science Campus of the UK's Office for National Statistics was set up to explore how data science could change the evidence base for the UK. One of the key areas for exploration is economics. More than two years after the Campus was set up, this presentation looks at what has been achieved.

How have we enhanced or supplemented traditional economic statistics? What are the challenges for incorporating unstructured data, collected by third parties into official statistics? How can we best share what we have learned, and what are the challenges for implementing data science prototypes into production?

Here, we present examples of economic data science from the Campus, including from our Faster economic indicators project, and use these to illustrate how we have addressed the challenges described.

Keywords

data science; economics; big data; official statistics

1. Introduction

The appetite for faster, more granular and more comprehensive information has never been higher. Policymakers and analysts demand faster, better insights into the state of economies and societies in order to make well-informed, timely decisions on national and international matters.

With the growing availability of big data and large administrative datasets, and the tools, technology and skills to understand, process and analyse these, National Statistics Institutes (NSIs) are being challenged to produce outputs that meet the growing demand for richer, more timely data.

In this paper, we use three economics projects from the Campus to illustrate how we have been meeting this challenge:

- initial work from our Faster indicators of UK economic activity programme (1), which uses three datasets: her Majesty's Revenue and Customs (HMRC) UK Value Added Tax (VAT) returns; ship tracking data from automated identification systems (AIS) for UK waters; and road traffic sensor data for England
- understanding the characteristics of high growth companies using non-traditional data sources (2)

- Optimus - a tool to turn free text into hierarchical datasets (3) to understand the movement of goods in the UK.

These projects use a range of administrative data, geospatial data, data from the Internet of Things, and text, and tools and techniques including machine learning, natural language processing and distributed computing to give fresh insights into the UK economy and the business behaviours which drive it. In Section 2, we briefly summarise our approach, and in Section 3, present an overview of the results. In Section 4, we discuss these results, and conclude with a summary of our findings.

2. Methodology

2.1 Faster indicators of UK economic activity

The faster indicator project set out to fast indicators of the UK economy using novel data sources based on administrative data and data from the Internet of Things. It has three objectives:

- to identify close-to-real-time data which represent useful economic concepts
- to create early warning indicators of potentially large economic changes
- to provide new insights into economic activity, particularly around UK ports

It is important to note that we are not attempting to forecast or predict gross domestic product (GDP) or other headline economic statistics here, and the indicators should not be used in this way. Rather, by exploring big, closer-to-real-time datasets of activity likely to have an impact on the economy, we provide an early picture of a range of activities that supplement official economic statistics and may aid economic and monetary policymakers and analysts in interpreting the economic situation in a timely way.

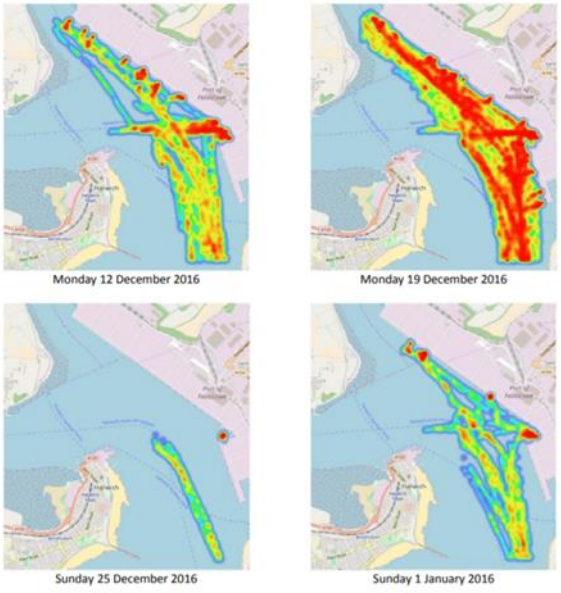
From UK VAT data, we have created, with industry breakdowns where possible:

- several monthly and quarterly diffusion indices from turnover and expenditure VAT returns
- novel indicators tracking changes in VAT reporting (counts of repayments, re-inputs and replacements)
- a proxy for firm births, based on counts of new VAT reporters.

We created shipping indicators from automatic identification system (AIS) data, which tracks ship location every few seconds whilst the ship is moving and every couple of minutes whilst it is in port, using data from the UK Maritime and Coastguard Agency, and, via the UN Global Platform, from ORBCOMM. We have used the data to construct monthly indicators of the time spent in port by ships, and the frequency of visits to ports, for the 10

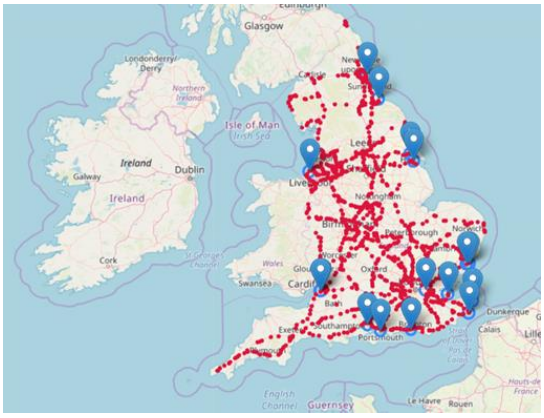
largest ports in the UK. These indicators are likely to be important in supplementing our understanding of international trade activity. Figure 1 shows a heatmap visualisation of the number of ships in Felixstowe port for four days in December 2016.

Figure 1: heatmap showing the number of ships in Felixstowe port for four days in December 2016. There were few ships in port on Christmas day.



Finally, we have used road traffic sensor data for England, published by Highways England to construct monthly indicators of average traffic counts and average traffic speeds for the whole of England and 13 main English ports. Figure 2 shows the location of road traffic sensors around England. The road traffic indicators were produced for all-England, and for 13 English port areas, and split into five length categories, which allows, for example, heavy good vehicles (HGVs) to be analysed separately from cars and motorbikes.

Figure 2: location of road traffic sensors in England (red dots) and of the ports analysed (blue pins)



2.2 Understanding the characteristics of high-growth companies using novel data sources

The goal of this project was to use novel data sources to explore the characteristics of firms with high growth. We used four data sources: the UK's statistical business register (the inter-departmental business register, IDBR); a high-growth flag constructed by the Department for Business, Energy and Industry Strategy (BEIS) from HMRC VAT data; a dataset from GlassAI, a start-up, who shared a random sample of data from 30,000 UK company websites, including company descriptions, sectors, mentions, news articles, job adverts and bios; and geolocations of UK retail clusters from the Ordnance Survey.

The IDBR, high-growth flag and GlassAI data were linked, giving a total sample of 5,500 companies, of which 8.6% were high-growth.

Supervised learning classification, using a gradient boosted classifier (GBC) was used to identify the features of high growth firms, firstly from the IDBR data alone, and then from the IDBR data linked to the GlassAI data.

Spatial analysis was carried out to investigate whether high growth is related to geographical location in retail clusters, where retail clusters may be seen as a broad proxy for density of economic activity. And finally, topic analysis was carried out on the GlassAI textual data.

2.3 Optimus - a tool to turn free text into hierarchical datasets

Many datasets contain variables that consist of short free-text descriptions of items or products. Optimus is a tool developed with the Department for Environment, Food and Rural Affairs (DEFRA) to understand shipping manifests of ferry journeys. The manifests are short, messy text descriptions of cargo on lorries boarding ferries. The huge variation in detail, scale of description and how items are recorded (such as incorrect spellings or

syntactic differences for identical products) make it difficult to automatically clean the data to a structured state that is ready for aggregation and analysis.

Optimus is a natural language processing (NLP) pipeline that retrieves vector representations of item descriptions and allows tiered grouping of both syntactically (words that look the same) and semantically (words which are contextually similar) similar descriptions. This produces a structured dataset where each item can be classified across multiple hierarchical tiers. Data can then be aggregated to an appropriate level or linked to existing taxonomies.

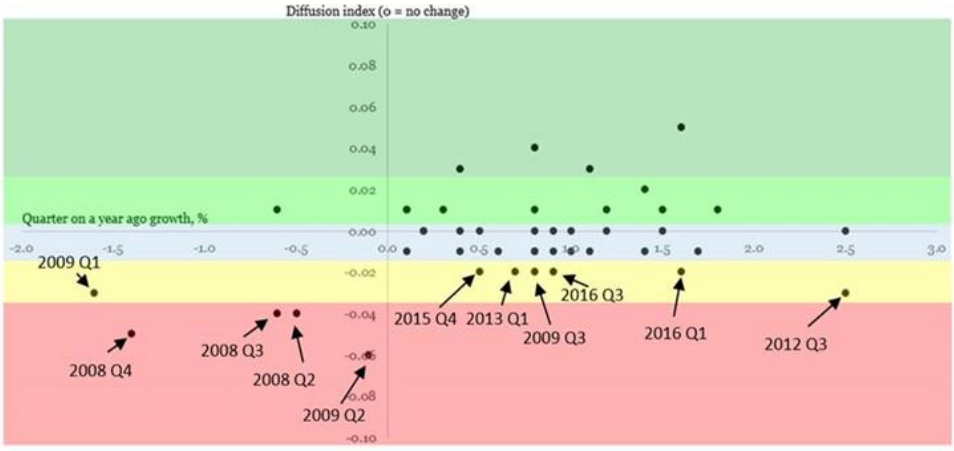
3. Results

All the faster indicators show promise as early-warning indicators of large changes in the economy, although none, on their own, should be used as proxies for gross domestic product (GDP) or other official economic statistics. Figure 3 shows a comparison of the quarterly VAT turnover index with GDP from 2008 to 2018. The index is strongly negative during the recession, but there is a much more scattered relationship with GDP during periods of greater stability. This demonstrates how novel uses of data can give a new window on the economy (we may identify economic turning points more quickly than waiting for official figures) but that care should be taken with interpretation (the diffusion index is not a proxy for GDP).

The shipping indicators and road traffic indicators for HGVs are, as one might expect, more closely correlated with trade in goods than with GDP, but again, there is enough scatter in the relationship to make them unreliable as a direct proxy, not least because we do not, at present, know what – if anything – is being transported the ships or lorries. However, the geographical granularity of these indicators is useful. Changes in activity in and around ports is likely to have an economic impact, and if we can link this to the types of products shipped at particular ports, we have the potential not only for an early warning, but also perhaps some information about the types of industries that might be affected.

ONS is now publishing the new faster indicators on a monthly basis (4). They are published as research outputs, to reflect the fact that they are still in development, whilst allowing users to explore them and offer feedback.

Figure 3: GDP quarter-on-quarter growth rates (current prices) and the VAT turnover diffusion index, both seasonally adjusted. The index captures the last recession quite well (most periods of the recession lie in the bottom left quadrant, red band), but is less good at tracking small changes in GDP during periods of greater stability (scatter in the top right quadrant, blue, green).



The pilot project identifying the characteristics of high-growth firms had mixed results. Supervised learning classification did not perform well and was no better when the GlassAI data were added to the IDBR data. This may in part be because, after linking the 3 datasets, we were left with a relatively small sample of around 5,500 firms. There is a need to further develop robust linking methodologies, so that novel data sources can be linked with sufficient quality to business registers. It may also be that a wider range of data are required to fully understand what drives high growth.

However, there were some interesting insights from the topic modelling and spatial analysis. Figure 4 shows a summary of the topic analysis. High-growth firms in the sample were more likely to talk about management, services, teams and (perhaps unsurprisingly!) awards, and less likely to talk about tax, law and manufacturing. A much larger sample is required to understand whether these are real features of high-growth firms, or whether they reflect different sectors, and some sectors are more likely to be high-growth than others. The spatial analysis showed that high-growth firms are more likely to be located in retail clusters, and retail clusters are likely to be a proxy for urban density.

A sample of the output for Optimus is shown in Figure 6. It can be seen, for example, that 'horse feed' and 'hrose feed', which are syntactically similar, sit close together, as do 'whiskey' and 'vodka', which are semantically similar. The green lines can be followed up through the levels, to give the required group level.

A web application was also developed for Optimus, which allows users to explore the clusters and labels, and amend these if required. This enables the data experts to carry out fast, intuitive quality assessment on the outputs, and is an important component of assurance on this new dataset, constructed from complex natural language processing algorithms. This is a good example of how humans and artificial intelligence (AI) can complement each other.

Figure 4: summary of main words used by high growth companies for different free text collected from their websites. Text in green is more likely to be mentioned by high growth firms, whilst text in red is less likely to be mentioned for the different free text entries.

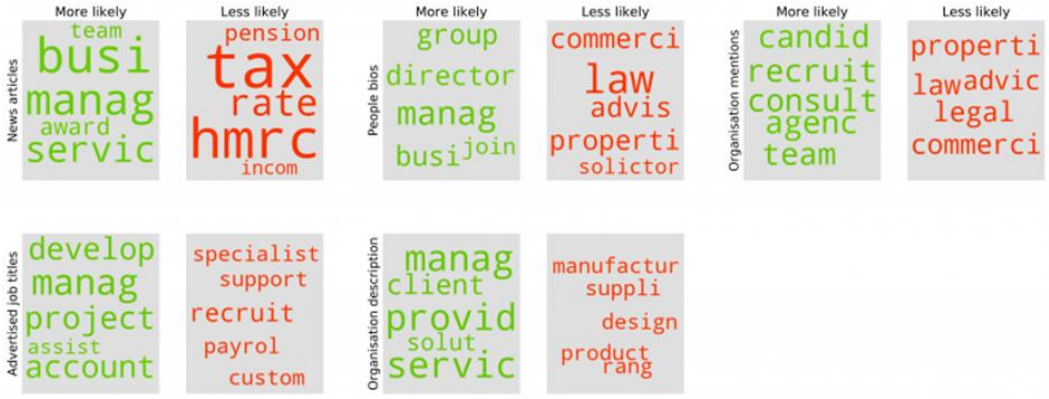


Figure 5: dendrogram showing the result of the first iteration clustering for a dataset of product descriptions using Optimus

4. Discussion and Conclusion

4.1 Discussion

In the faster indicators programme, we have supplemented official statistics using novel data sources:

- identifying close-to-real-time big data and administrative datasets, which represent useful economic concepts
- creating a set of indicators that allow early identification of large economic changes
- providing insight into economic activity, at a level of timeliness and granularity not currently possible with official economic statistics.

Although our indicators are not – and are not intended to be – a proxy for GDP or other official statistics, they act as an early warning system for the UK economy, addressing the need for faster economic information for decision-makers. We currently publish them as monthly research outputs, and, although these are not 'official statistics' they are regular, timely indicators, which provide us with a new angle on economic activity, using data in a novel way. We plan to further develop the shipping and road traffic indicators to provide new indicators for the movement of goods into and around the UK.

The pilot project exploring the characteristics of high-growth firms has produced some tantalising insights into how 'non-traditional' data can be used to supplement our understanding of firm growth. We were able to demonstrate how administrative, geospatial and textual data can be linked, at a firm level, and used machine learning to give a richer insight into firm behaviour than is possible with standard analyses of aggregate statistics. Aspects of this work have now been adopted by the UK Department for Business, Energy and Industrial Strategy (BEIS), to inform their policy development.

The ability to take messy free text and translate it into syntactic and semantic hierarchies has many potential applications, not only in economics. In our project, we were able to summarise the goods being shipped by lorry across ferry routes around the British Isles. It is unlikely that these will become official statistics, as there is a wide variation in the quality of the data recorded on the manifests, and there is no information on the volume and value of goods being transported. However, in the absence of an existing survey or access to more complete administrative data, the output of this project delivered a rapid and valuable insight on the local movement of goods, where previously very little information was available.

4.2 Conclusions

In this paper, we showcase some of the work of the UK's Data Science Campus. The projects discussed demonstrate how the combination of novel data sources and the tools and techniques of data science can enhance the evidence base for economics.

We learn that:

- Big Data is not a silver bullet, but we can create useful evidence, even where outputs are not – or not yet – of the same quality as official statistics
- building a new knowledge base for quality, not just of novel data sources, but also of novel techniques, is challenging, but the principles remain the same: understand user needs so that these are adequately met, and explain quality clearly, so that users are confident in knowing how and when to use new outputs; we may also need new approaches

to quality assurance, which give the subject matter experts confidence in results from complex algorithms, such as the web application developed for Optimus
- collaboration is key for data science teams, and that includes collaboration with subject matter exerts and data experts, as well as the ultimate users
- AI is also not a silver bullet, and human-in-the-loop approaches allow humans to focus on oversight and quality assurance for difficult cases, with AI processes relieving humans from the burden of repetitive processing
- scale-up, from development and prototyping to production, requires a different set of skills from development, and possibly a different team, again, collaboration is vital in ensuring a smooth handover.

All this work is more fully described in the references (1,2,3).

References

1. Nolan L. et al. (2019), Faster indicators of UK economic activity, https://datasciencecampus.ons.gov.uk/faster-indicators-of-uk-economic-activity/
2. Pugh D., Williams S. & Johnson C. (2019), Understanding the characteristics of high growth companies using non-traditional data sources, https://datasciencecampus.ons.gov.uk/projects/understanding-the-characteristics-of-high-growth-companies-using-non-traditional-data-sources/
3. Hopkins S., Clews G. & Eidukas A. (2019), Optimus – A natural language processing pipeline for turning free-text lists into hierarchical datasets, https://datasciencecampus.ons.gov.uk/projects/optimus-a-natural-language-processing-pipeline-for-turning-free-text-lists-into-hierarchical-datasets/
4. Research Output: Economic activity, faster indicators, UK, https://www.ons.gov.uk/economy/economicoutputandproductivity/output/articles/economicactivityfasterindicatorsuk/previousReleases

Statistical performance index - Assessing country-level statistical capacity on a global scale

Mustafa Dinc, Juderica Dias
World Bank

Abstract

Data and statistics are a fundamental requirement for evidence-based decisions. They are central to tracking results, and critical to holding decision-makers accountable. Governments, international agencies, civil society organizations, and the private sector cannot deliver on their respective mandates without data and statistics. Demands on national systems are rising. National statistical systems need to become more agile and maintain and deepen their overall capacities to meet the increasing demand for data and statistical services. This, in turn, requires continuous monitoring and evaluations of these systems. Statistical Performance Index (SPI) is one such tool that can help monitor and evaluate the performance of national statistical systems.

Keywords

Statistical Capacity; Assessment tools; Statistical performance index; national statistical system

1. Introduction

It has now been widely recognized that relevant, timely and quality statistics are essential for evidence-based policy formulation, decision making and for monitoring and evaluation of development progress. It has also been acknowledged that the quality of statistics is as good as the agencies that produce them. Availability of and access to timely and good quality data are critical to holding decision-makers accountable. Governments, international agencies, civil society organizations, and the private sector cannot deliver on their respective mandates without data and statistics.

The data is produced and disseminated at the country level by the national statistical systems to inform a range of issues – social, economic, environmental, and political. These statistics must be relevant, timely, and credible.

National statistical systems need to become more agile and maintain and deepen their overall capacities to meet the increasing demand for data and statistical services. This, in turn, requires continuous monitoring and evaluations of these systems. Statistical Performance Index (SPI) is one such tool that can help monitor and evaluate the performances of national

statistical systems. This paper provides a succinct discussion about the Statistical Performance Index (SPI).

2. Statistical capacity and assessment tools

Statistical capacity is the ability of National Statistical Systems to meet user needs for relevant and good quality statistics on a timely manner. A well-functioning statistical system should be able to collect, analyze, and disseminate high-quality data about its population and economy. Just producing data, however, is not enough to demonstrate that the statistical system has the capacity. The indicators, data series and other outputs must meet the needs of users and must be provided in a form that supports their widespread use. The data outputs need to be of good quality, published within a time frame that means the data are still relevant and provided in a format that users can access and use. It is also important that statistical systems are open and transparent about their methods and procedures and provide access to adequate metadata – detailed descriptions of the methods and procedures used to produce the data. A capable statistical system should also be able to bring about transformation that is generated and sustained over time from within.

As an object of study, 'capacity' is too broad a concept with many intangibles and elements that are difficult to measure and often quite subjective. This fact is also true for the capacity of national statistical systems that involve different data producing agencies and departments within a given country. Therefore, it is very difficult to measure statistical capacity directly.

Even though it is quite difficult to measure, and it could be costly, it is necessary to understand and assess the capacity of national statistical system to identify weaknesses and strengths and to provide guidance for necessary interventions.

In the absence of direct measures, certain tools are customized to assess capacity to for different purposes and audiences. In general, there are four main drivers to assess capacity:

- To inform the national planning and strategic process that could help national authorities develop strategic plans for the improvement of national statistical system. It could provide guidance for improving the quality and scope of products available to users, identifying priorities for improvement, and bringing coherence for donor support.
- To inform program/project design and implementation process focusing on the strengths and weaknesses of the national statistical systems, their current human resource capacity, infrastructure and governance arrangements.
- To assess compliance with codes of practice endorsed by relevant international and/or regional organizations to assure and maintain the

standards and Codes of Practice required as part of membership to these organizations.

- To assess the statistical performance of countries on a global scale as Global Public Goods in terms of the scope of data products; compliance with international standards; periodicity, timeliness and the accessibility of results.

Over the years, a number of tools and approaches have been developed utilizing certain type of questionnaires to be completed by staff from national statistical system or by experts recruited for this purpose. The assessment process takes place in the country with involvement of external experts and relevant staff from the national statistical system. Such assessments are country specific and could provide a deeper and better understanding of the national statistical system, but this process is costly, time consuming and often imposes additional burden on already weak capacity of statistical systems. It could also result in different interpretation of questions and hence different answers that, in turn, could make international comparison difficult.

The SPI is one of these tools that provides a globally consistent and comparable assessment of country statistical systems by focusing on a smaller set of indicators and using publicly available information.

Irrespective of the specific purpose, interventions derived from SPI assessments have the same long-term objective – a national statistical system that should be able to sustainably collect, analyze, and disseminate high-quality data about its population and economy to inform evidence-based policy making and monitoring and evaluating development programs.

Resource-intensive country assessments lack cross-country comparability and difficult to summarize. Moreover, given the degree of subjectivity associated with in-depth assessments, different assessors could arrive at a different overall summary.

Given the difficulties of synthesizing detailed assessments to determine progress on a global scale, there is an understandable desire to form a single composite index drawing from publicly available information. The SPI in conjunction with Country Statistical Profiles provide the means to compare countries and to track performance over time.

3. The Statistical Performance Index (SPI)

The SPI framework is designed to capture different aspects of national statistical capacity by employing most relevant and representative variables that are publicly available. The SPI can be used to gauge statistical performance of individual countries over time or cross-country comparisons of performance at a point in time.

The SPI aims to provide an objective, justifiable/verifiable assessment of the statistical performance of countries over time by using publicly available

information from international agencies and country websites that were produced by national statistical systems. The SPI framework helps countries and development partners identify the strengths and weaknesses of national statistical systems and areas of potential improvements. It could also provide actionable guidance for national statistical systems in areas that may require further and deeper assessment.

Key features of the SPI are:

- Uses only publicly accessible data
- Transparent methodology
- Easily replicable
- Provides a long-time series to track progress in performance
- Captures outcomes and supporting elements
- Reflects the SDGs.
- Facilitates at-a-glance comparisons on a global scale

SPI Methodology

Due to their complex and multi-dimensional nature socio-economic phenomena cannot be measured by a single descriptive indicator. Instead, generally a composite index method is utilized to measure and understand such phenomena. In constructing a measure that is policy relevant it is helpful to follow a series of basic steps.

The first step asks the question: what phenomenon is being measured? A clear conception helps orient the process by which the measure is assembled and will prove valuable in communicating its underlying meaning.

The second step asks: for what purpose or purposes is the index being sought? Knowing how the index will be used can greatly affect subsequent choices in its construction, and its eventual suitability. In particular, it will help define the unit of analysis both for data gathering and reporting purposes.

The third step identifies a list of essential characteristics, or desiderata, that the methodology should exhibit. This list of "pre-axioms" helps orient the construction process and define what success means.

A fourth step identifies the conceptual space in which measurement is to take place. If there are multiple conceptual dimensions, consideration must also be given to the relative importance of each.

The fifth step selects the form of the variables to be used and the aggregation method to be employed – how the variables are to be combined into an overall measure.

The sixth step identifies a set of axioms that the resulting index should satisfy to have the greatest practical utility. Axioms are not sterile mathematical requirements, but rather contain the salient nuggets of policy required of the index: which aspects of the data should be ignored, which should be reflected, and helpful consistency requirements over subsets of

data. Together, these six steps comprise the core theoretical elements of our proposed measurement technology.

Characteristics of the Statistical Performance Index (SPI)

Constructing a measure of statistical capacity entails many distinct choices that can appear to be arbitrary and unrelated to one another if no context is provided. A set of desired characteristics, or criteria, can provide the guiding principles that help organize these choices to obtain a relevant and useful measurement tool. SPI is designed to satisfy seven criteria. The SPI should be:

1. Simple. It must be understandable and easy to describe
2. Coherent. It must conform to a common-sense notion of what is being measured
3. Motivated. It must fit the purpose for which it is being developed
4. Rigorous. It must be technically solid
5. Implementable. It must be operationally viable
6. Replicable. It must be easily replicable
7. Incentive Compatible. It must respect country incentives

The SPI also satisfies three axioms. The symmetry axiom requires that the index value is unaffected when variable levels are switched. The dominance axiom requires that the index value rises whenever one variable rises from 0 to 1 and the rest of the variables do not fall in value. The subgroup decomposability axiom allows the index to be divided into salient sub-indices and linked back to the original index for policy analysis.

SPI Dimensions

The production process for statistical outputs has certain similarities to the traditional production model from economics and begins with a technology that is used in generating the statistical products, and the level of this technology is clearly a relevant component of statistical capacity. The resulting statistical outputs might be divided into two general categories. First are the intermediate products, which have direct use for specialists but require additional processing to create products suitable for general use. For example, a census can be helpful for policy analysts but must be processed to obtain useful statistics. Second are the final products, which are available in a form that can be understood by the public. The key macro statistics of a country would naturally be viewed as final products. Even after the products have been created, their existence does not imply that potential users will actually have access to them. Statistical products may be available to only a few users, or available to all. The final dimension then covers the extent to which statistical products are disseminated.

This simple framework helps to identify four coherent dimensions for a measure of statistical capacity, namely: (i) Methodology, Standards and

Classifications (MSC), which provides information on the technology being used by the NSS; (ii) Census and Surveys (CS), which describes the intermediate products of the NSS; (iii) Availability of Key Indicators (AKI), which focuses on key final products needed for policy; and (iv) Dissemination Practices and Openness (DPO), which evaluates the extent to which products are publicly disseminated. It is easy to see that each of these dimensions is centrally related to the statistical capacity of an NSS.

SPI Dimension 1: Methodology, Standards and Classifications (MSC):

Internationally accepted and recommended methodology, classifications and standards provide the basis for national statistical offices (NSOs) on data integration, facilitating data exchange and providing the foundation for the preparation of relevant statistical indicators. This dimension aims to assess whether national statistical systems have the necessary capacity to adopt and comply with international statistical standards.

To keep the dimension simple and comparable across countries, the selection is based on the categories under IMF SDDS standards that basically cover the following sectors: real sector (national accounts, production index, labor market, price indices etc.), fiscal sector (central government operations), financial sector and external sector (balance of payments, external debt). Standards in the field of social statistics were also examined but not selected due to lack of verifiable and actionable assessment criteria for all countries. A variable on Civil Registration and Vital Statistical (CRVS) and GSBPM variable are included under this dimension.

These 12 indicators cover a major portion of relevant standards and methodologies. It is assumed that if national statistical systems have necessary capacity (human, physical and financial), they will be able to adopt and employ these standards and receive higher scores and rankings.

SPI Dimension 2: Censuses and Surveys (CS):

Data collection is the key responsibility of national statistical systems where information on a nation's population, economy, health and other aspects are recorded in an accurate and timely manner. Through censuses and surveys, sometimes together with administrative systems, the NSSs collect data and generate aggregate indicators based on the results. This dimension aims to check the availability and frequency of key censuses and inter-census surveys. The use of administrative data in producing official statistics, particularly in advanced statistical systems, has become an integral part of the data production process. However, due to verification and comparability issues administrative systems are not included in the ranking.

This dimension covers 8 indicators on population and housing census, agricultural census, business census, income and expenditure surveys and other surveys on agriculture, health, labor force and establishments.

SPI Dimension 3: Availability of Key Indicators (AKI):

Transforming source data into statistical outputs (indicators) and releasing them on a timely basis shows that the statistical systems are utilizing their capacity in data production. Reporting relevant data to specialized international agencies on time and getting them published in their respective databases demonstrates that statistical systems meet required quality standards and timeliness. Therefore, this dimension evaluates national statistical systems by reviewing the availability of country data for the most recent year in international databases. By looking at the data availability in international databases, it also makes the assessment cost-effective.

These selected indicators cover key socio-economic and SDG indicators that have well established standards and methodology in the area of poverty, health, education, and economic development. It is assumed that if national statistical systems have capacity in the first two dimensions they should be able to produce these selected indicators.

SPI Dimension 4: Dissemination Practices and Openness (DPO):

Data users are seen as an integral part of the national statistical system, which is crucial to the improvement of collection, processing and dissemination of quality data. Therefore, dissemination practices of statistical systems reflect an important part of the overall statistical capacity. This dimension is built on the principle that quality statistics should be delivered to the public in a timely, easily accessible manner for free. It includes 10 indicators under two sub-sections: Dissemination capacity of NSO and Openness of Data.

These four dimensions are closely linked and capture the production cycle of national statistical systems in collecting, producing and disseminating quality statistics. By following internationally recommended standards and classifications, statistical systems will have the basic foundations for data collection that will make produced data comparable with other countries. Then, ideally with a combination of administrative sources and timely censuses and surveys, statistical systems will collect, process and analyze relevant data and generate necessary indicators as data products covering different aspects of households and establishments. Finally, statistical systems will disseminate these final data products through their official websites, regular publications and by submitting them to relevant international organizations. Through this cycle statistical systems produce statistics that reflect the socio-economic conditions of the nation and inform decision making.

SPI Variables and Aggregation

Once the dimensions have been specified, attention turns to identifying the variables and selecting an appropriate aggregation method. The criteria suggest that the variables in an index should be coherent with the concept

being measured; they should be publicly available. The aggregation method should be selected with rigor in mind, including the axioms or properties that the method satisfies. At the same time, it should aim for simplicity to maximize general understanding and impact.

In the domain of statistical capacity, a variable is typically derived from a simple "yes-no" question concerning a normative guideline that a country's NSS should meet. Each of these "yes-no" questions generate a dichotomous variable having a 0-1 representation, where 1 means that the underlying test or target has been successfully achieved, while 0 indicates it has not.

The indicator selection process is guided by conventions of international agencies, expert opinions on statistical performance and the principles of SDGs. However, given the cost and time constraints and the accuracy concerns of assessment, trade-offs have to be made to build an actionable, cost-effective and internationally comparable index.

One such trade-off is the equal weighting of each dimension and individual indicator, even though some of them may be more important than others or countries may assign higher priority to some than others. The equal weighting selection may be, to some extent subjective, partly failing to address the relative importance of the dimensions and indicators. This could be checked by simulations that will show the sensitivity of SPI scores and rankings in relation to alternative weights.

When variables are dichotomous (or can be dichotomized), a measurement approach called a "counting method" is applicable and, indeed, has become standard for many types of measurement exercises. This method is used here to aggregate the scores of four dimensions into an overall SPI score and to create the composite index. Hence:

Total SPI Score =(MSC+CS+DPO+AKI)/4

Each of the four dimensions has a scale of 1-100 and are aggregated into a total score which also ranges from 1 to 100.

4. Conclusions

This SPI could be the first step before more resource-intensive country-specific assessments to inform multi-year improvement plans. The SPI framework is also flexible enough to allow for future revisions as the global data landscape evolves. For example, it is possible to incorporate new indicators such as whether an NSO uses cloud computing to store their data or implements household panel surveys in the relevant dimensions without creating major changes to the total scores. The SPI may also be relevant to the construction of other indexes in related areas, such as tracking the global SDGs or child development. Since the SPI will be produced every year, it will provide time series data for monitoring the progress over time.

Measuring recruitment costs of migrant workers through household surveys: Results of a pilot test from Lao PDR Labour Force Survey 2017

Tite Habiyakare[1], Kuangjie Zhong[2]

[1]International Labour Organization (ILO), Regional Office for Asia and the Pacific, Bangkok, Thailand

[2]Ministry of Human Resources and Social Security (MOHRSS), Beijing, China

Abstract

Lao PDR implemented its 2nd national labour force survey (LFS) in September 2017. The survey included a module testing the measurement of international labour migration, return migrants, and absentees, as well as recruitment costs of migrant workers. The module was meant to contribute to methodological work for the 2030 Sustainable Development Goals (SDGs) indicator 10.7.1 on "Recruitment cost borne by employee as a proportion of monthly income earned in country of destination". The indicator is currently considered as Tier II in the SDG Global Indicator Framework, with a new measurement methodology not yet broadly implemented by countries. The pilot process in Lao PDR LFS 2017 was done before the methodological work on this indicator was completed. Results of this pilot process are presented in this paper. The main data used in this analysis are those on return migrant workers: there were 52,600 return migrant workers in Lao PDR in 2017, representing only 0.7 per cent of the total population, and their average recruitment costs were estimated at USD141, i.e. one third of the monthly salary during the last job abroad.

Keywords

International migration; return migrant; employment; earnings; country of origin

1. Introduction

Migration has seen an increased role in the SDGs unlike the previous Millennium Development Goals (MDGs), including a dedicated Target 10.7 on safe migration (UN, 2015). For the monitoring of this Target two indicators have been adopted, and one of these, i.e. SDG indicator 10.7.1, is on the costs that migrant workers have to pay to get a job abroad. Migration is prominent into the SDGs in three ways: (i) with a migration-specific target, i.e. SDG Target 10.7 (Facilitate orderly, safe, regular and responsible migration and mobility of people, including through the implementation of planned and well-managed migration policies); (ii) as part of at least some other 5 Targets (such as Target 5.2- trafficking of women and girls, Target 8.7- forced labour and human

trafficking, Target 8.8- migrant workers' rights, Target 10.c- remittances, and Target 16.2- trafficking of children); and (iii) as an overarching disaggregation variable (as stated in Target 17.18). In 1990 the UN General Assembly adopted the International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families in its resolution 45/158, with a legal definition of a migrant worker (UN, 1990). However, it is only recently that the international community adopted a statistical definition of international labour migration at the 20th International Conference of Labour Statisticians (ICLS) of October 2018 (ILO, 2018). The Guidelines on statistics for SDG indicator 10.7.1 (ILO & World Bank, 2018) were also endorsed by the Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs) in November 2018. Its methodology is therefore still new. Lao PDR LFS 2017 was among the first country-level pilot tests to contribute to developing these Guidelines.

2. Methodology

This section covers the main concepts used in this paper in line with current international standards, as well as the estimation methodology for SDG indicator 10.7.1 from Lao PDR LFS 2017 data. The survey, implemented by Lao Statistics Bureau (LSB) from mid-July to end August 2017, was a onetime stand-alone national household survey covering a representative sample of 10,520 households.

Main concepts used in this paper

International migrant: the UN recommendations on statistics of international migration define international migrants as "the set of persons who have ever changed their country of usual residence, that is to say, persons who have spent at least a year of their lives in a country other than the one in which they live at the time the data are gathered" (UN, 1998). In practice such information corresponds to the total number of usual residents born abroad (foreign-born population), or usual residents who are not citizens (foreign population), as in the recent Principles and Recommendations for Population and Housing Censuses, Revision 3 (UN, 2017).

Migrant worker: the 1990 UN Migrant worker's convention defines a migrant worker as "a person who is to be engaged, is engaged or has been engaged in a remunerated activity in a State of which he or she is not a national" (Art.2.1). As per international migration, the reference population for international labour migration covers all persons who are usual residents of the measurement country. However, it also includes "persons who are not usual residents in the country but who are, nevertheless, in the labour force or potential labour force or any other forms of work in that country" (ILO, 2018), with the exception of refugees and asylum seekers.

The 20th ICLS Guidelines concerning statistics of international labour migration define therefore migrant workers as international migrants and non-resident foreign persons who are in the country's labour force. However, the concept excludes: (i) foreign military and diplomatic personnel, (ii) international travellers on tourism whose main purpose is not to work, and (iii) non-resident staff of call centres and those providing services from a foreign location.

Return migrant worker: the term comprises "all current residents of the country who were previously international migrant workers in another country" (ILO, 2018), irrespective of their citizenship, birth place, current labour force status, or whether they were residents in the foreign country of work. This paper identifies a return migrant worker as any usual resident who lived in another country in the past, or who travelled abroad at any time in the past, even if for a short period, for the purpose of working or looking for work. In practice return migrant workers are proposed as the main target population when running recruitment costs surveys in a migration sending country, or country of origin, while for the country of destination the proposed target population is that of usual resident migrant workers.

Recruitment costs and components: in the ILO General principles and operational guidelines for fair recruitment and definition of recruitment fees and related costs, the concept of recruitment fees or related costs refers to "any fees or costs incurred in the recruitment process in order for workers to secure employment or placement, regardless of the manner, timing or location" (ILO, 2019), as long as those costs are borne (directly or indirectly) by the migrant worker.

The current Guidelines for statistics for SDG indicator 10.7.1 presents details of some 14 recruitment costs items that should be included in the calculation of the indicator, i.e.: (1) Recruiter/job broker charges; (2) Visa costs; (3) Inland transportation expenses; (4) International transportation; (5) Passport fees; (6) Medical fees; (7) Insurance fee; (8) Security clearance fee; (9) Pre-departure briefing; (10) Language training; (11) Skills assessment fee; (12) Contract approval fee; (13) Welfare fund fee; and (14) Interest payment on debt incurred to cover recruitment costs.

For the pilot test in Lao PDR and for this paper, recruitment costs were grouped into three main items:

(a) travel costs to and back from the destination country,
(b) recruitment agencies or brokers' fees and related costs, including costs paid to friends and relatives, and
(c) other costs including preparations costs for work abroad, passport, visa, insurance and any medical costs.

Monthly income: the concept refers to the actual income earned as a wage/salary, as defined in the Resolution concerning an integrated system of

wages statistics adopted by the 12th ICLS (October 1973). "The concept of earnings, as applied in wages statistics, relates to remuneration in cash and in kind paid to employees, as a rule at regular intervals, for time worked or work done together with remuneration for time not worked, such as for annual vacation, other paid leave or holidays" (ILO, 1973). Earnings exclude employers' contributions to social security and pension schemes, as well as severance and termination pay.

Estimation methodology of SDG indicator 10.7.1 from Lao PDR LFS 2017

The Guidelines for statistics for SDG indicator 10.7.1 "recommend that the statistics/estimates on costs and earnings used to calculate 10.7.1 should refer to the first job obtained in the last country of destination within recent years (for example, in the 3 years prior to the survey year)" (ILO & World Bank, 2018), and earnings should be collected for the first month of that job. However, this pilot test was implemented before the Guidelines were finalised, and estimates presented in this paper refer to the typical monthly earnings during the last job abroad, as in the LFS questionnaire (LSB, 2018).

Recruitment costs indicator (RCI): In this paper the RCI is defined only for the subset M of those return migrant workers with non-zero recruitment costs and non-zero earnings abroad, so that the indicator can be produced and analysed at individual level, as the equivalent number of months of salary to recover the recruitment cost. Statistics on those migrant workers with no recruitment costs or with no earnings should be published separately in addition to the RCI. The RCI indicator is a proportion of costs in earnings at individual level. It can be expressed as a function of the costs and earnings of the return migrant worker k in the subset of M migrant workers;
i.e.:

$$RCI = f(\frac{Ck}{Ek})$$

Where

Ck = is the recruitment costs paid by individual k, among the subset of M migrant workers who declared both costs and earnings (non-zero costs and non-zero earnings);

Ek = is the monthly earnings of the same individual k, among the subset of M migrant workers.

At aggregate levels the measure can be equated to using a proportion of totals (total costs and total earnings); i.e.:

$$RCI = \frac{\sum_{k=1}^{M} Ck}{\sum_{k=1}^{M} Ek}$$

Caution on the results

The sampling design for the Lao PDR LFS 2017 was aimed at estimating reliable employment and unemployment statistics for the country. Sample allocation such as by urban and rural areas and by provinces was based on this requirement. However, migrants do not come equally from all provinces, and this consideration was not used during the survey sampling design. To provide better estimates on migrant workers, basic information on international migration should be used in the sampling design.

The actual sample size after completion of the survey includes some 52,166 individual cases, with 1,612 cases as return migrants (3.1 per cent of the sample), and only 284 cases are return migrant workers from paid employment (0.5 per cent of the overall sample of individual cases). Some disaggregation of data from this pilot may therefore not be statistically significant due to small sample size, and we have limited the disaggregation to fewer categories when presenting results.

3. Results

In this paper we present selected results of the pilot test, starting with a summary on recruitment costs, earnings, and the RCI. We then look at the structure (distribution) of recruitment costs and earnings, and finally present data for the main corridor (Thailand), as well as for the skill levels of migrant workers (low- versus high-skilled migrant workers).

Summary results

A summary of main results on return migrant workers, recruitment costs, monthly earnings, and the RCI, is presented below.

Table 1: Return migrant workers and key recruitment costs statistics (recruitment costs and earnings in USD*)

Statistic	Sample cases	Estimate (person)	Estimate (mean, or percent)	Min	Max	Standard error (of mean)	Totals (for costs & earnings)
Population (person)	**52,162**	**6,915,559**	**...**	**...**	**...**	**...**	**...**
Male	25,736	3,408,996	...	...	...	...	...
Female	26,426	3,506,563	...	...	...	...	...
Return MW (person)	**280**	**52,639**	**...**	**...**	**...**	**...**	**...**
Male	172	31,037	...	...	...	...	...
Female	108	21,602	...	...	...	...	...
Costs (mean)**	**254**	**48,429**	**141.31**	**2.41**	**1,060.56**	**.64**	**6,843,640**
Male	153	27,919	141.35	2.65	1,060.56	.91	3,946,239
Female	101	20,510	141.27	2.41	964.15	.88	2,897,401
Earnings (mean)**	**254**	**48,429**	**430.88**	**4.82**	**3,266.04**	**1.90**	**20,866,748**
Male	153	27,919	435.49	22.30	2,548.96	2.39	12,158,304
Female	101	20,510	424.60	4.82	3,266.04	3.08	8,708,444
RCI (months, mean)	**254**	**48,429**	**0.33**	**.01**	**2.67**	**...**	**...**

Male	153	27,919	0.32	.01	2.50	...	...
Female	101	20,510	0.33	.01	2.67	...	...
MW with no costs (%)	**11**	**2,574**	**4.9**	**...**	**...**	**...**	**...**
Male	8	2,000	6.4	...	...	...	...
Female	3	574	2.7	...	...	...	...
MW with no earnings (%)	**20**	**2,864**	**5.4**	**...**	**...**	**...**	**...**
Male	15	2,174	7.0	...	...	...	...
Female	5	690	3.2	...	...	...	...

Source: Authors calculations based on LFS data from the 2017 labour force survey of Lao PDR.

Notes: (*) = September 2017 UN exchange rate: 1 USD = 8,297.500 Laotian Kip (LAK).

(**) = Subset of migrant workers with non-zero costs & non-zero earnings; four outlier cases (all males) who earned more than USD5,000 per month were removed from the analysis (by identifying unusual cases).

(...) = Denotes Not Applicable.

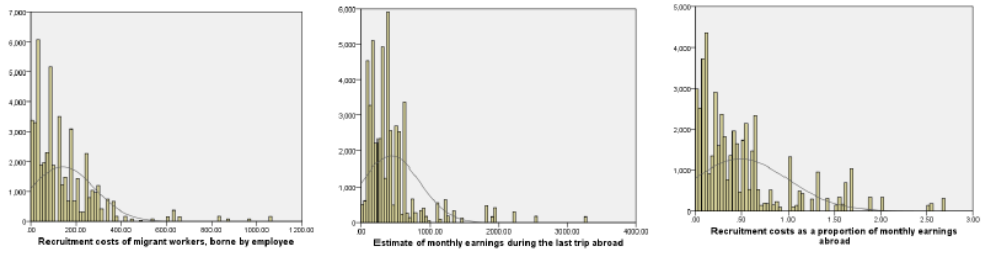
(MW) = Denotes migrant worker.

The average recruitment costs is about USD141; however, the median is at USD96, i.e. about USD45 below the average, and the mode is at USD36, signalling low recruitment costs for many Laotian migrants, as the majority goes in the neighbouring Thailand, and use mostly informal channels to obtain a job there. Only 6.4 percent went through formal channels, i.e. through a job transfer or registration (see also Harkins B et al., 2017). However, the sample was too small to capture reliable data on costs incurred for formal channels of migration which, in the case of Lao PDR is likely to be higher. Return migrant workers received an average monthly wage of USD431 (slightly higher for males than for females, i.e. USD435 against USD425). On average they paid about 33 per cent of monthly earning of their last job abroad in recruitment costs. The total of all recruitment costs was estimated at USD 6.8 million (i.e. 0.04 per cent of the 2017 country's GDP).

Distribution of recruitment costs and earnings

The figure 1 below present the distribution of costs and earnings, as well as the recruitment cost indicator (in 3 panels).

Figure 1: Distributions of costs (Panel A), earnings (Panel B), and RCI (Panel C)



Panel A Panel B Panel C

Both three distributions are skewed as any costs or earnings. Recruitment costs and earnings are concentrated at the lower side (the median of earnings is about USD100 lower than the average, even after removing the outliers). However, 15.8 per cent of return migrant workers paid at least one month of wage of their last job abroad as recruitment costs (1.00 and above in Panel C). About 4.9 per cent did not pay any recruitment costs (see Table 1).

Recruitment costs per corridors and skill levels

Table 2 below presents statistics on recruitment costs per corridors and low- versus high-skilled workers, for those migrant workers with non-zero values for both recruitment costs and earnings. Only Lao PDR-Thailand corridor is presented as it is the main corridor, with 94.4 per cent of return migrant workers (statistics on the other corridors may not be reliable due to small sample size).

Table 2: Recruitment costs of return migrant workers with non-zero values on both costs and earnings (in USD*) by sex, geographic location, last country of destination (corridors), and skill levels (**)

Selected variables	Frequency, return MW (person, %)	Recruitment costs (mean, USD)			Proportion in monthly earnings abroad		
		Total	Male	Female	Total	Male	Female
Total	**52,639**	**141.31**	**141.35**	**141.27**	**0.33**	**0.32**	**0.33**
Geographic location							
Urban	20.3	92.00	109.91	71.26	0.22	0.23	0.20
Rural	79.7	151.62	147.38	157.61	0.35	0.35	0.36
Country of destination (corridor)							
Thailand	94.4	138.04	136.85	139.59	0.33	0.32	0.33
Others	5.6	247.82	231.00	450.78	0.36	0.33	0.67
Skill levels							

High-skilled	0.7	482.07	-	482.07	1.00	-	1.00
Low-skilled	99.3	141.08	141.35	140.71	0.33	0.32	0.33

Source: Authors calculations based on LFS data from the 2017 labour force survey of Lao PDR.

Notes: (*) = September 2017 UN exchange rate: 1 USD = 8,297.500 LAK.

(**) = Estimated by educational levels in this paper as in ISCO-08 (ILO, 2012); the ideal should be by occupations abroad.

(-) = Denotes zero value.

(MW) = Denotes migrant worker.

One notes that return migrant workers in Lao PDR were mostly low-skilled (99.3 per cent) and were mostly living in rural areas (79.7 per cent). Therefore, statistics presented in Table 2 are not reliable for high-skilled and urban workers due to small sample size. However, one can note that recruitment costs seem to be higher for rural than for urban migrant worker, and are likely to be higher for high skilled than for low-skilled return migrant workers (in the Lao PDR context).

4. Discussion and Conclusion

Despite possible issues with the small sample size, Lao PDR LFS 2017 provides an insight and data on the labour migration process in and from Lao PDR, and on recruitment costs: return migrants were estimated at 208,500 persons, and only 25.2 per cent of them (52,600 persons) were return migrant workers as currently defined in the SDG indicator 10.7.1 Guidelines. The main destination country of Laotian migrant workers is Thailand: about 94.4 per cent of return migrant workers in 2017 were coming back from Thailand.

The recruitment costs of return migrant workers as a proportion of their monthly earnings is estimated at 33 per cent, with no significant differences between women (33 per cent) and men (32 per cent). Laotian return migrant workers were mostly found in rural areas and were predominantly low-skilled workers (99.3 per cent). However, some statistics of this study such as those on high-skilled migrant workers are to be considered with caution due to small sample size. Surveys on SDG indicator 10.7.1 will need to make sure the sampling design takes into consideration the existing data on both migrants in urban and rural areas, in high and low skill levels, as well as those with formal versus informal migration channels.

References

1. Harkins B., Lindgren D. & Suravoranon T. (2017). Risks and rewards: Outcomes of labour migration in South-East Asia. A joint report of the International Labour Organization (ILO) & the International Organization for Migration (IOM), Bangkok.
2. International Labour Organization (2019). General principles and operational guidelines for fair recruitment & Definition of recruitment fees and related costs. International Labour Office - Fundamental Principles and Rights at Work Branch & Labour Migration Branch, Geneva.
3. ___ (2018). Guidelines concerning statistics of international labour migration. 20th International Conference of Labour Statistician (ICLS), Geneva.
4. ___ & World Bank (2018). Statistics for SDG indicator 10.7.1: Guidelines for their Collection. ILO website (https://www.ilo.org/global/topics/fair-recruitment/lang--en/index.htm).
5. ___ (2012). International Standard Classification of Occupation: ISCO-08. International Labour Office, Geneva.
6. ___ (1973). Resolution concerning an integrated system of wages statistics. 12th International Conference of Labour Statistician (ICLS), Geneva.
7. Lao Statistics Bureau (2018). Lao PDR Labour force survey 2017: Survey finding report. Vientiane.
8. United Nations (2017). Principles and recommendations for population and housing censuses, Revision 3. New York.
9. ___ (2015). Integrating migration into the 2030 Agenda for Sustainable Development. Population Facts, No 2015/5, Population Division, United Nations Department of Economic and Social Affairs (UNDESA), New York.
10. ___ (1998). Recommendations on Statistics of International Migration, Revision 1. Statistical Papers, Series M, No 58, Rev.1, New York.
11. ___ (1990). International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families. Adopted by General Assembly resolution 45/158, New York.

Modelling volatility with outlier detection in asymmetric GARCH (p, q) models on JSE index

D.L Sepato[1], N.D Moroke[2], J.T Tsoku[2]
[1]Nelson Mandela University
[2]North West University

Abstract

Financial data often contain observations caused by unexpected events, called interventions, and such extreme returns are often found to disturb volatility less than a standard time series model would forecast. The main purpose of this study is to assess the performance of GARCH type family models with outlier(s) from a set of data. An iterative procedure is given for the outlier's detection and correction method to check the presence of any type of the four common outliers in financial data. The study explored through the guidance of the ACFs and PACFs the ARMA enhanced GARCH models such as the ARMA (0, 2)-GARCH (1, 1), ARMA (0, 2)-EGARCH (1, 1) and ARMA (0, 2)-GJR-GARCH (1, 1) models to assess the volatility in stock returns data. The study used daily time series data from the year 03 January 2011 until 21 April 2016 was sourced from the JSE database. ARMA (0, 2)-EGARCH (1, 1) model was confirmed to be adequate and was appropriate after the outliers were removed. The model was recommended for further analyses and were later used for producing forecasts of JSE stock returns.

Keywords

ARCH; GARCH-variants; Additive outlier; Level shift outlier; Temporary change outlier; Innovation outlier; JSE index

1. Introduction

Modelling volatility of financial time series is a key area of investigation in econometrics. Though there are many customary volatility Generalised Autoregressive Conditional Heteroscedastic (GARCH) models being used, to capture the stylised facts of financial time series, a drawback of the model is that it cannot capture the asymmetric features found in financial returns, thus, to bridge the gap various asymmetric GARCH models have been proposed (Raziq, Iqbal and Talpur, 2017).

There is a dearth of literature on studies that detect outliers in time series prior to modelling and producing forecasts. The main objective of this study was to assess of the performance the asymmetric GARCH-type models in outlier free data. Therefore, outliers may also introduce bias in the estimated parameters of GARCH models.

This study, despite the different types of outliers focused on four basic and commonly used types of outliers in time series namely, the Additive Outlier (AO), Innovational Outlier (IO), Temporary Change (TC) and Level Shift (LS) outlier discussed in Fox (1972) proved that the power of the test performed better than that of the iid assumption. Studies by (Andersen et al., (2007); Boudt et al. (2013); Laurent et al. (2016), and Verhoeven & McAleer (2004) caution that the consequence of neglecting jumps (outliers) in GARCH models usually overestimate the volatility during several days, if not weeks, after the occurrence of these jumps. Though these models have enjoyed the success in studying the evolution of financial time series modelling, they are unable to explain the high frequency and size of extreme jumps commonly occurring in practice which may wrongly suggest conditional heteroscedasticity (Carnero et al., 2007).

2. Methodology

Data used in this study consist of the daily closing prices of the primary South African indices the JSE index obtained from the Johannesburg Stock Exchange (JSE) ranging from January 3, 2011, to April 29, 2016. The percentage log returns r_t are calculated as:$r_t = 100 * log$ ($_X{}^X t-{}^t 1$), where *Xt* is the daily closing price of JSE index stock market at time *t,* and *Xt-1* denotes lagged stock price on day *t-1,* this is used to forecast ARCH/GARCH *(p, q)* models.

2.1. General outlier detection in ARCH and GARCH models

An outlier detection procedure used by Fox (1972) follows the following steps to detect outliers in a time series data:

1. Model a GARCH (1, 1) using the original data under the assumption that there are no outliers
   $y_t = x_t'\xi + \varepsilon_t, \varepsilon_t|\Gamma_{t-1} \sim N(0, h_t)$ (3)
   Γ_{t-1} is the division up to time *t.* In practice, x_t consists of a constant term
2. The outlier effect $\varpi(\tau)$ and the residual variance, σ_v are computed from the residuals.
3. Using (2), the test statistics $\tau_i(\tau)$is calculated for all possible $\tau = 1,\dots n$, where *i* denotes outlier type. When $V_i(L) = {}^1\!/_{\pi(L)}$ this indicates an IO, $V_i(L) = 1$ implies an AO, $V_i(L) =$
   ${}^1\!/1 - (L)$which implies a LS outlier. Then $e_t = \pi(L)z_t \Rightarrow e_t = \omega_i x_t + a_t$, where the IO
   $\omega_i = \omega_I$ and $x_t = I^{(T)}$. AO $\omega_i = \omega_A$ and$x_t = \pi(L)I^{(T)}$, LS $\omega_i = \omega_L$ and $x_t = \pi(L)(1 - L)^{-1}I^{(T)}$. This will test the null hypothesis for an outlier
   H_0:$\omega_A = \omega_I = \omega_L = 0$ against H_A: $\omega_A \neq 0$, H_I: $\omega_I \neq 0$,H_L: $\omega_L \neq 0$ (4)
   This tests the likelihood ratio test statistics for testing H_0 vs. H_A, H_I, and H_L respectively, $\lambda_{iT} = \hat{\omega_i}/\sigma_i$ for *i* = I, A, and L where σ_i is the standard

deviation of the estimate. Under the null hypothesis of no outliers, these statistics are asymptotically distributed as N (0, 1) (Jesús Sánchez & Pena, 2003).

4. The maximum of the absolute value of these test statistics, $\tau_{\max=Max} t=1,\ldots,n|\tilde{\imath}_{\tau}(\tau)|$ is computed. If the value of the test statistic exceeds the pre-specified critical value, C (significant) then an outlier is detected. Thus, the point t where τ_{max} occurs is the point detected as having the outlier. The procedure is to compute the effect of outliers on residuals following (Tsay, 1988) where the actual parameters π and $\sigma^2$are known in the modelling procedure estimated by any consistent estimator by employing: AO: $\lambda_A = max_{\{T:1\leq T\leq n\}}|\lambda_{A,T}|$, IO: $\lambda_I = max\{T{:}1\leq T\leq n\}|\lambda I,T|$, LS: $\lambda L = max\{T{:}1\leq T\leq n\}|\lambda L,T|$, (4,5,6) where these are used as testing criterions for outlier detection. Comparing the test statistics with critical value C the existence of outliers can be detected, where the time points at which the above maxima occur are timings of the corresponding outliers.
5. The final step is to determine whether the observations are outliers and by removing each outlier from the series by deducting the value of the effect of ω, then apply the GARCH modelling procedure to obtain the most adequate model and use it for forecasting future values of the series.

2.2. Information criterion for model selection between the candidate models

The model with the highest subsequent probability is the one that minimises BIC, a desirable model is one that minimises the AIC or the BIC (Ngailo, 2011).

2.3. Model diagnostics

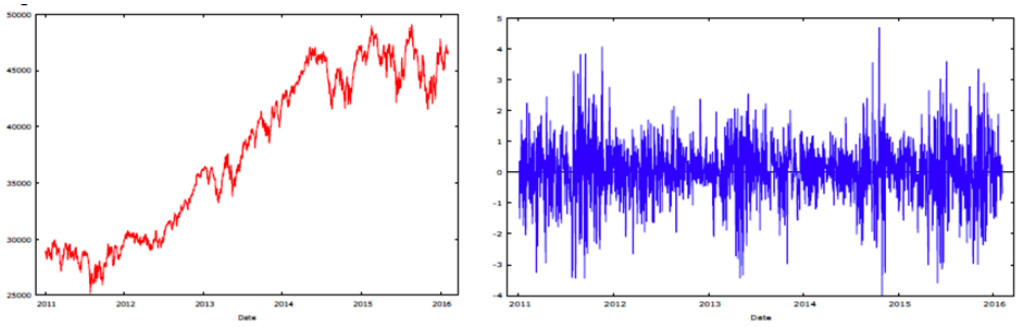
This section discusses the model diagnostic tests for the selected model such as runs test for independence, normality tests, Test for autocorrelations and heteroscedasticity in that respect.

2.4. Forecasting performance evaluation

This study uses three evaluation measures to evaluate the forecast accuracy of JSE top 40 index for the proposed model ARMA-GARCH model applied, namely the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) statistics.

3. Result

This section provides and discuss the results that quantify, summarise and check the distributional properties of the financial time series.



The JSE top 40 index also shows a nonlinear upward stochastic trend. Thus, the log transformation has produced a stationary process. Although the plot depicts a constant mean across sub samples of the data, which is consistent with other studies of financial returns, where the mean is often found to be stationary, there is suspicion that the series is somehow volatile. This is due to abnormalities roughly above and below the mean.

Table 1: KPSS test of the JSE top 40 index and returns

	KPSS Test statistic (Prob.)*	Kurtosis	Jarque-Bera
JSE top 40 index	4.215939 (0.0000) ***	-1.5613544	136.8483(< .0001)
Returns	0.06508 (0.2206) *	1.29215392	94.8555 (<0.0001)

*Notes: ***, **, * denotes significance at 1%, 5% and 10% levels respectively.*

The KPSS test was also used to confirm stationarity for the JSE top 40 index. The test statistic is 4.215939 and a corresponding probability value of 0.000 is less than the critical values at all levels. Therefore, JSE top 40 index is non-stationary. The results of KPSS test show that the returns are stationary since LM-stat is greater than 0.10 at all asymptotic critical values. Thus, it can be concluded that the time series has no unit root.

As a result, the distribution of the series is platykurtic due to negative excess kurtosis. Inferences concerning non-normality is maintained by the Jarque-Bera test statistics for the JSE top 40 index and returns which show that the null hypothesis is rejected at 5% level of significance. Thus it can be concluded that the JSE top 40 index have a non-normal distribution which is a common occurrence in stock markets. As suggested by Sigauke et al. (2014), lack of non-normality of the distribution is due to volatility clustering.

The ARMA-GARCH type modelling results - "outlier free"

This section focused on outlier detection processes to analyse the outlier-adjusted returns $\{y_t^*\}$ and to also examine time-varying volatility. The study also investigates the effect of outliers through test statistics, and performance of the criteria and outlier detection procedure Stage I: Locate outliers

Table 2: outlier detection summary

Outlier Detection Summary	
Maximum number searched	1328
Number found	99
Significance used	0.01

Table 2 shows the maximum number of outliers found is 99 at 1% level of significance under the tests of hypothesis. There are two types of outliers detected namely the AO and LS at 0.01 level of significance as shown in Table 3 (e.g. $at\ t$ = 991 $or\ t = T$ = 18/12/2014,τ_{AO} = 4.5309) . Thereforeω = 99, with a threshold critical value of $cval$ = 3.5. Following the Chen and Liu (1993) procedure, outliers are detected through inner and outer loops indicated in Table 4 (Appendix). Iterations around the function locate outliers until no additional outliers are found or the maximum number of iterations is reached. After each iteration, the effect of the outliers on the residuals of the fitted model is removed and the t-statistics are obtained again for the modified residuals. No model selection or refit of the model is conducted within this loop. At the end of each iteration, the detected outliers are removed from the original data and a new check for the presence of outliers is carried out. Figure 1 shows the data for JSE top 40 index the plot shows the measure of outlier effects, $\hat{\omega}_{TP} = TC$ (t =1… 1000). All $\hat{\omega}_{TC,t}$ lie in the interval [-4, 4]. As it was observed in Table 4 there are five TC outliers of size ω = 3.5. Consequently, this indicates that an ARMA-GARCH model will be able to isolate time point at which TC occurs. Figure 1 shows the original data (grey line), the adjusted series (blue line), the location of the detected outliers (red points) and their estimated effects (red line) for the return series. Therefore, due to the nature of the outliers detected the effect of outliers is not permanent as it affects a single observation at a particular time.

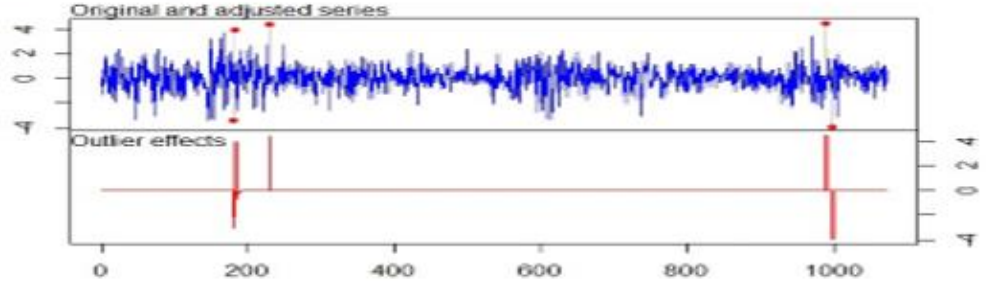


Figure 1: Effects of AO and TC

Stage II: Remove outliers

If any of the outliers turn to be non-significant then they are removed from the set of potential outliers. An outlier free data was used for further analysis. Thus the new ARMA (p, q)-GARCH (p, q) type models were modelled with 1312 observations after removal of 18 significant outliers from a set of data.

Stage III: iterate stages I and II for the adjusted series (model fit)

After the outliers were removed the ACF and PACF were plotted. According to the results, AR (2), MA (2) are seen to be significant in this section. The, parameters for ARMA (0, 2) were estimated. ARMA (0, 2)-GARCH (1, 1) and ARMA (0, 2)-EGARCH (1, 1) and ARMA (0, 2)-GJR-GARCH (1,1) are fitted.

Table 5: GARCH type models with outlier free model estimation summary

Parameter	ARMA(2,2)-EGARCH (1, 1)	ARMA(2,2)-GJR (1, 1)	ARCH (2)	ARMA(0,2)-GARCH (1, 1)
μ	0.0126 (0.5861) *	0.01964 (0.4292) *	0.0391 (0.1432) *	0.05823 (0.0158) **
θ_1	-0.005998 (0.8384) *	-0.00561 (0.8458) *	0.7543 (<.0001) ***	0.1014 (0.0004) ***
θ_2	0.0942 (0.0007) ***	-0.0998 (0.0005) ***	0.1691 (<.0001) ***	0.025665 (0.0001) ***
ω	0.000901 (0.8133) *	0.02182 (0.0226) **	0.2037 (<.0001) ***	0.088371 (0.0000) ***
α	0.0541 (<0.0001) ***	0.036161 (0.0009) ***		0.891771 (0.0000) ***
β	0.9811 (<0.0001) ***	0.99933 (0.000) ***		
γ	-2.6465 (0.0001) ***	0.90835 (0.000) ***		

Notes: *, **, * denotes significance at 1%, 5% and 10% levels respectively. Values in the parentheses are the probabilities of the tests statistics of the estimated parameters.**

Table 7: GARCH Outlier Uncontaminated Model selection summary

Criterion	ARCH (2)	ARMA(0, 2)-GARCH (1, 1)	ARMA(0, 2)-EGARCH	ARMA(0, 2)-GJR-GARCH(1, 1)
MSE	1.17866	1.16091	1.16272	1.16283
SBC	3868.0396	3776.48725	3704.3122	3710.516
AIC	3847.3224	3745.4114	3668.05701	3674.2715
MAE	0.80501	0.79761	0.80025	0.79998
MAPE	106.78218	118.74248	110.0026	112.058
Log-Likelihood	-1919.661	-1886.7057	-1827.029	-1830.136
Rank	(4)	(3)	(1)	(2)

Table 7 shows the results that suggest ARMA (0, 2)-EGARCH (1, 1) as an adequate model based on the information criterion and the model performance evaluations. Therefore, this model was used to make forecasts for outlier free series. The returns of ARMA (0, 2)-EGARCH (1, 1) are formulated as:

$r_t = 0.0126 + \varepsilon_t$

$ln\ \sigma_t^2 = -0.005998 + 0.0942\varepsilon_{t-1} + 0.000901{\varepsilon_t}^2{}_{-2} + 0.0541\hbar_{t-1} + 0.9811\hbar^2{}_{t-1} - 2.6465\gamma$

$\lambda = 1.0352$

Outlier free ARMA (0, 2) - GARCH (1, 1) model forecasts are presented in Figure 2.

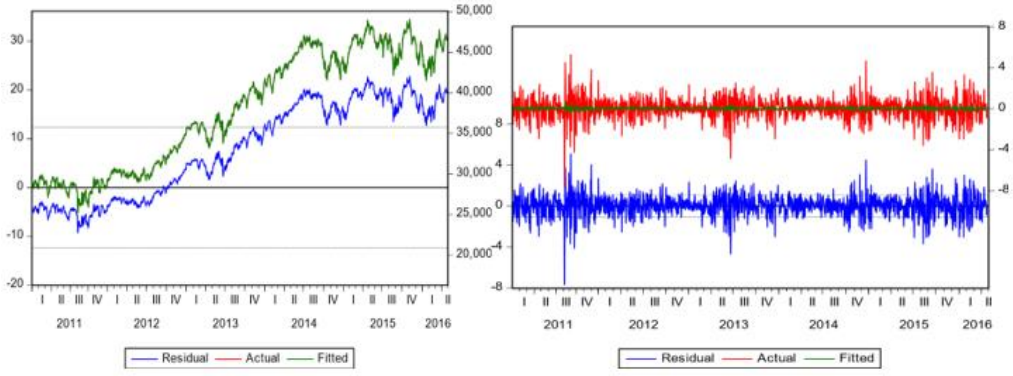


Figure 2: Forecast plot for ARMA (0, 2)-EGARCH (1, 1)

4. Discussion and Conclusion

The RSME/MSE is lower thus according to Brooks (2008) the forecast with the smallest RMSE and MAE provides the most accurate forecasts. MAPE shows that the original series is the best since the MAPE value is closest to 100. Therefore, SBC and AIC will be used to select the best model. This is a confirmation that outlier free ARMA (2, 2) - EGARCH (1, 1) model is good for the data since it has a small forecasting error. In conclusion, the outlier free data provides good forecasts to predict volatility of JSE top 40 index.

References

1. SIGAUKEA, C., MAKHWITING, R.M. & LESAOANA, M. 2014. Modelling conditional heteroskedasticity in JSE stock returns using the Generalised Pareto Distribution. African Review of Economics and Finance. 6(1):41-55.
2. FOX,A.J. 1972. Outliers in time series. Journal of the Rovail Statistical Society. Series B (Methodological):350-363.
3. CHEN, C. & LIU, L.-M. 1993. Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88(421):284-297.
4. TSAY, R.S. 1988. Outliers. level shifts. and variance changes in time series. Journal of Forecasting.7(1):1-20
§. JESUS SANCHEZ, M. AND PENA, D.. 2003. The identification of multiple outliers in ARIMA models. Communications in Statistics-Theory and Methods. 32(6). pp.1265-1287.
6. LAURENT, S., LECOURT, C. & PALM, F.C.. 2016. Testing for jumps in conditionally Gaussia ARMA-GARCH models. a robust approach. Computational Statistics & Data Analysis. 100. pp.383-400.

7. BOUDT, K., DANIELSSON, J. & LAURENT, S. 2013. Robust forecasting of dynamic conditional correlation GARCH models. International Journal of Forecasting. 29(2):244-257.
8. ANDERSEN, T.G., BOLLERSLEV, T. & DIEBOLD, F.X. 2007. Roughing it up: Including jump components in the measurement. modelling. and forecasting of return volatility. The review of economics and statistics. 89(4):701-720.
9. CARNERO, M., PENA, D. & RUIZ, E. 2007. Effects of outliers on the identification and estimation of GARCH models. Journal of time series analysis. 28(4):471-497.
10. VERHOEVEN, P. 2004. Modelling Outliers and Extreme Observations for ARMA-GARCH Processes. (In. Econometric Society World Congress 2000 Contributed Papers organised by: Econometric Society.
11. RAZIQ, A., IQBAL, F. AND TALPUR, G.H.. 2017. EFFECTS OF ADDITIVE OUTLIERS ON ASYMMETRIC GARCH MODELS. Pakistan Journal of Statistics. 33(1).
12. NGAILO E. Modelling and forecasting using time series GARCH models: An application of Tanzania inflation rate data. Unpublished Master's Thesis). University of Dar es Salaam. 2011.

Appendix

Table 4: Outlier detection: inner and outer loops and Table 3: showing outlier details

	OUTER LOOP				INNER LOOP		
TYPE	TIME ID	COEFHAT	TSTAT	TYPE	TIME ID	COEFHAT	TSTAT
AO	07-Sep-11	3.870239	4.418528	AO	10-Mar-11	-3.316543	-3.786392
AO	22-Sep-11	-3.73737	-4.266836	AO	04-Aug-11	-3.37602	-3.854295
AO	27-Sep-11	3.904602	4.45776	AO	18-Aug-11	-3.185753	-3.637073
AO	30-Nov-11	4.159271	4.748508	AO	29-Aug-11	3.079603	3.515885
AO	18-Dec-14	4.32413	4.936721	AO	07-Sep-11	3.870239	4.418528
AO	05-Jan-15	-3.699429	-4.22352	AO	22-Sep-11	-3.73737	-4.266836
TC	20-Jun-13	-2.43836	-4.006183	AO	27-Sep-11	3.904602	4.45776
AO	11-Aug-11	3.463976	4.014625	AO	30-Nov-11	4.159271	4.748508
TC	02-aug-211	-2.406181	-4.021456	AO	11-Jun-13	-3.235001	-3.693297
TC	05-Aug-11	-2.433364	-4.066888	AO	21-Nov-14	3.243048	3.702485
TC	04-Aug-11	-2.549548	-4.261066	AO	05-Jan-15	-3.699429	-4.22352
TC	05-Aug-11	2.682333	4.483131	TC	02-aug-211	-2.170064	-3.565378
TC	09-Aug-11	2.733591	4.568801	TC	05-Aug-11	-2.143677	-3.522024
AO	18-Aug-11	-3.651	-4.28634	AO	11-Aug-11	3.351046	3.825783
				TC	20-Jun-13	-2.43836	-4.006183
				AO	18-Dec-14	4.32413	4.936721
				AO	14-Jan-15	-3.275653	-3.843651
				TC	04-Sep-11	-2.153391	-3.636322

Obs	Time ID	Type	Estimate	Chi-Square	Approx Prob>ChiSq
991	18-DEC-2014	Additive	4.53805	24.69	<.0001*
230	30-NOV-2011	Additive	4.16483	20.98	<.0001*
1000	05-JAN-2015	Additive	-3.87796	18.26	<.0001*
147	04-AUG-2011	Additive	-3.84758	17.98	<.0001*
1151	12-AUG-2015	Additive	-3.67482	16.46	<.0001*
1176	16-SEP-2015	Additive	3.61362	15.95	<.0001*
170	07-SEP-2011	Additive	3.49695	14.95	0.0001*
184	27-SEP-2011	Additive	3.47061	14.75	0.0001*
181	22-SEP-2011	Additive	-3.46914	14.76	0.0001*
616	20-JUN-2013	Additive	-3.46561	14.76	0.0001*
156	18-AUG-2011	Additive	-3.43613	14.55	0.0001*
49	10-MAR-2011	Additive	-3.41307	14.37	0.0002*
973	21-NOV-2014	Additive	3.38122	14.14	0.0002*
163	29-AUG-2011	Additive	3.29155	13.40	0.0003
383	13-JUL-2012	Shift	0.06359	7.21	0.0072*

On moments of folded and truncated multivariate extended skew-normal distributions

Christian E. Galarza[1]*, Larissa Avila Matos[2], Victor Lachos Davila[3]
[1] Department of Statistics, Campinas State University, Campinas, Brazil.
[2] Department of Statistics, Campinas State University, Campinas, Brazil.
[3] Department of Statistics, University of Connecticut, Storrs, CT, USA.

Abstract
Following Kan & Robotti (2017), this paper develops recurrence relations for integrals that involve the density of multivariate extended skew-normal distributions, which includes the well-known skew-normal distribution introduced by Azzalini & Dalla-Valle (1996) and the popular multivariate normal distribution. These recursions offer fast computation of arbitrary order product moments of truncated multivariate extended skew-normal and folded multivariate extended skew-normal distributions with the product moments of the multivariate truncated skew-normal, folded skew-normal, truncated multivariate normal and folded normal distributions as a by product. Finally, from the application point of view, these moments open the way to propose analytical expressions on the E-step of the Expectation-Maximization (EM) algorithm for complex data, such as, asymmetric longitudinal data with censored and/or missing observations. These new methods are provided to practitioners in the R MomTrunc package

Keywords
Product moments, Truncated distributions, Censored models.

1. Introduction

In many applications, researches often generate a large number of datasets with values restricted to fixed intervals. For example, variables such as pH, grades, viral load in HIV studies and humidity in environmental studies, have upper and lower bounds due to detection limits, and the support of their densities is restricted to some given intervals. Thus, the necessity of studying the truncated distributions along with their properties arises naturally. In this context, there has been a growing interest in evaluating the moments of truncated distributions. Also, these variable are often skewed, departing from the traditional assumption of using symmetric distributions. From Tallis (1961) to Arismendi (2013), several works have pursued to compute formulae for the rst two moments as well as higher order moments of truncated univariate/multivariate distributions as the truncated normal (TN), truncated t-Student (TT), truncated skew-normal (SN) (Azzalini & Dalla-Valle, 1996) distributions among others. Main applications involve environmental studies,

macroeconomics and actuarial data. For instance, Arismendi (2013) provided explicit expressions for computing arbitrary order product moments up to order 4 of the TMN distribution by using the moment generating function (MGF). However, the calculation of this approach relies on differentiation of the MGF and can be somewhat time consuming.

Instead of differentiating the MGF of the TN distribution, Kan & Robotti (2017) recently presented recurrence relations for integrals that involve directly the density of the multivariate normal (MN) distribution for computing arbitrary order product moments of the TN distribution. Although some proposals to calculate the moments of the univariate truncated skew-normal distribution (Flecher et al., 2010) and truncated univariate skew-normal/independent distribution (Flecher et al., 2010) has recently been published, so far, to the best of our knowledge, there is no attempt on studying neither moments nor product moments of the folded multivariate extended skew-normal (FESN) and truncated multivariate extended skew-normal (TESN) distributions. Moreover, this approach allows to compute as a by-product the moments of folded and truncated distributions, of the N (Kan & Robotti, 2017), SN (Azzalini & Dalla-Valle, 1996), and its respective univariate versions. The proposed algorithm and methods are implemented in the new R MomTrunc package.

Over the last decade or so, censored modelling approaches have been used in various ways to accommodate increasingly complicated applications. Many of these extensions involve using N and its symmetrical extensions, however statistical models based in distributions to accommodate censored, missing and skewness, simultaneously, have remained relatively unexplored in the statistical literature from the likelihood-based perspective. The results of this paper allow, for instance, to derive analytical expressions on the E-step of the EM algorithm for multivariate SN responses with censored and/or missing observation.

The rest of this paper is organized as follows. In Section 2 we briefy discuss some preliminary results related to the multivariate ESN, TESN and FESN distributions and some of its key properties. Section 3 presents a recurrence formula of an integral for the essential evaluation of moments of the TESN distributions. Explicit expressions for the rst two moments of the TESN distribution are also presented. In Section 4, we finally present interesting results for the FESN distribution as well as explicit expressions for the univariate case. Some concluding remarks are presented in Section 4. Proofs and two interesting applications have been omitted due to the lack of space.

2. Methodology

We denote a random variable by an upper-case letter and its realization by the correspondent lower case and use boldface letters for vectors and matrices. Let Ip represent a p×p identity matrix and a ones matrix respectively, A> be the transpose of A, and |X| = (|X1|. . . . , |Xp|) > mean the absolute value of each component of the vector X. For multiple integrals, we use the shorthand notation

$$\int_{\mathbf{a}}^{\mathbf{b}} f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{a_1}^{b_1} \cdots \int_{a_p}^{b_p} f(x_1, \ldots, x_p)\mathrm{d}x_1 \ldots \mathrm{d}x_p.$$

where $\mathbf{a} = (a_1, \ldots, a_p)^\top$ and $\mathbf{b} = (b_1, \ldots, b_p)^\top$.

General results to calculate the probability of a bounded random vector are summarized in the followings theorems:

Theorem 1. *Let* $\mathbf{X}$ *be a p-variate random vector with joint probability density function (pdf)* $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ *and cumulative density function (cdf)* $F_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$. *Let* $\mathbb{A}$ *be a Borel set in* $\mathbb{R}^p$ *of the form*

$$\mathbb{A} = \{(x_1, \ldots, x_p) \in \mathbb{R}^p : \ a_1 \leq x_1 \leq b_1, \ldots, a_p \leq x_p \leq b_p\} = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}. \quad (1)$$

Then

$$P(\mathbf{X} \in \mathbb{A}) = \sum_{\mathbf{s} \in S(\mathbf{a},\mathbf{b})} (-1)^{n_s} F_{\mathbf{X}}(\mathbf{s};\boldsymbol{\theta}),$$

where $S(\mathbf{a},\mathbf{b}) = \{\mathbf{s} : \mathbf{s} = (s_1, \ldots, s_p),\ with\ s_i = \{a_i, b_i\},\ i = 1, \ldots, p\}$ *and* $n_s = \sum_{i=1}^{p} \mathbb{1}(s_i = a_i)$.

Theorem 2. *Let* $\mathbf{X}$ *be a p-variate random vector with joint pdf* $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ *and joint cdf* $F_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$. *If* $\mathbf{Y} = |\mathbf{X}|$, *then the joint pdf and cdf of* $\mathbf{Y}$ *that follows a folded distribution are given, respectively, by*

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{s} \in S(p)} f_{\mathbf{X}}(\boldsymbol{\Lambda}_s \mathbf{y};\boldsymbol{\theta}), \qquad and \qquad F_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{s} \in S(p)} \pi_s F_{\mathbf{X}}(\boldsymbol{\Lambda}_s \mathbf{y};\boldsymbol{\theta}), \quad for\ \mathbf{y} \geq \mathbf{0},$$

where $S(p) = \{\mathbf{s} : \mathbf{s} = (s_1, \ldots, s_p),\ with\ s_i = \pm 1, i = 1, \ldots, p\}$, $\boldsymbol{\Lambda}_s = Diag(\mathbf{s})$ *and* $\pi_s = \prod_{i=1}^{p} s_i$.

It is important to stress that Theorem 2 generalizes the results found in Chakraborty & Chatterjee (2013) for the FMN case to all distributions of belonging the multivariate location-scale family.

a. The extended multivariate skew-normal distribution (ESN)

We say that a p × 1 random vector Y follows a ESN distribution with p × 1 location vector μ, p × p positive definite dispersion matrix Σ, a p×1 skewness parameter vector λ ∈ R p , and shift parameter τ ∈ R, denoted by Y ∼ ESNp(μ, Σ,λ, τ), if its pdf is given by

$$ESN_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) = \xi^{-1}\phi_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad (2)$$

with $\xi = \Phi_1(\tau/(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2})$. Note that when $\tau = 0$, we retrieve a skew-normal distribution that except by a straightforward difference in the parametrization, it corresponds to that introduced by Azzalini & Dalla-Valle (1996) that is, $ESN_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},0) = SN_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$. It is also interesting to note that $ESN_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) \longrightarrow \phi_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$, as $\tau \to +\infty$. The following propositions are crucial to our methodology.

Proposition 1. *Let* $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau)$ *and* $\mathbf{Y}$ *is partitioned as* $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$ *of dimensions* p_1 *and* p_2 $(p_1 + p_2 = p)$, *respectively. Let*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top, \quad \boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top \quad and \quad \boldsymbol{\varphi} = (\boldsymbol{\varphi}_1^\top, \boldsymbol{\varphi}_2^\top)^\top$$

be the corresponding partitions of $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ *and* $\boldsymbol{\varphi} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda}$. *Then,*

$$\mathbf{Y}_1 \sim ESN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, c_{12}\boldsymbol{\Sigma}_{11}^{1/2}\tilde{\boldsymbol{\varphi}}_1, c_{12}\tau), \quad \mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim ESN_{p_2}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}, \boldsymbol{\Sigma}_{22.1}^{1/2}\boldsymbol{\varphi}_2, \tau_{2.1})$$

where $c_{12} = (1 + \boldsymbol{\varphi}_2^\top \boldsymbol{\Sigma}_{22.1}\boldsymbol{\varphi}_2)^{-1/2}$, $\tilde{\boldsymbol{\varphi}}_1 = \boldsymbol{\varphi}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\varphi}_2$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$ *and* $\tau_{2.1} = \tau + \tilde{\boldsymbol{\varphi}}_1^\top(\mathbf{y}_1 - \boldsymbol{\mu}_1)$.

Proposition 2. *If* $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$, *then for any* $\mathbf{y} \in \mathbb{R}^p$

$$F_{\mathbf{Y}}(\mathbf{y}) = P(\mathbf{Y} \leq \mathbf{y}) = \frac{\Phi_{p+1}\big((\mathbf{z}^\top, \tilde{\tau})^\top; \mathbf{0}, \boldsymbol{\Omega}\big)}{\Phi(\tilde{\tau})}, \tag{3}$$

where $\mathbf{z} = \mathbf{y} - \boldsymbol{\mu}$ *and* $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma} & -\boldsymbol{\Sigma}^{1/2}\boldsymbol{\psi} \\ -\boldsymbol{\psi}^\top\boldsymbol{\Sigma}^{1/2} & 1 \end{pmatrix}$, *with* $\boldsymbol{\psi} = \boldsymbol{\lambda}/(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}$ *and* $\tilde{\tau} = \tau/(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}$.

Hereinafter, for $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$, we will denote to its *cdf* as $F_{\mathbf{Y}}(\mathbf{y}) \equiv \tilde{\Phi}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ for simplicity.

Truncated extended multivariate skew-normal

The doubly truncated extended multivariate skew-normal (TESN) distribution is obtained by conditioning on $\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}$, where $\mathbf{X} \sim \mathrm{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$. Let $\mathbf{Z}$ be the resulting truncated ESN random variable vector, denoted by $\mathbf{Z} \sim \mathrm{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau, [\mathbf{a}, \mathbf{b}])$. It follows that $\mathbb{E}[\mathbf{Z}^{\boldsymbol{\kappa}}]$ can be expressed as

$$\mathbb{E}[\mathbf{Z}^{\boldsymbol{\kappa}}] = \frac{\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{z}^{\boldsymbol{\kappa}} ESN_p(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)\mathrm{d}\mathbf{z}}{\int_{\mathbf{a}}^{\mathbf{b}} ESN_p(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)\mathrm{d}\mathbf{z}}, \ \mathbf{a} \leq \mathbf{z} \leq \mathbf{b}. \tag{4}$$

For the special case $\tau = 0$, we refer to this distribution as a multivariate TSN, i.e., $\mathrm{TSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, [\mathbf{a}, \mathbf{b}])$.

3. Results

3.1 On moments of the doubly truncated multivariate ESN distribution

For two vectors $\mathbf{x} = (x_1, \ldots, x_p)^\top$ and $\boldsymbol{\kappa} = (k_1, \ldots, k_p)^\top$, let $\mathbf{x}^{\boldsymbol{\kappa}}$ stand for $(x_1^{\kappa_1}, x_2^{\kappa_2}, \ldots, x_p^{\kappa_p})$, and let $\mathbf{a}_{(i)}$ be a vector $\mathbf{a}$ with its ith element being removed. For a matrix $\boldsymbol{\Delta}$, we let $\boldsymbol{\Delta}_{i(j)}$ stand for the ith row of $\boldsymbol{\Delta}$ with its jth element being removed. Similarly, $\boldsymbol{\Delta}_{(i,j)}$ stands for $\boldsymbol{\Delta}$ with its ith row and jth columns being removed. Besides, let $\mathbf{e}_i$ denote a $p \times 1$ vector with its ith element equaling one and zero otherwise.

Let

$$\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \int_{\mathbf{a}}^{\mathbf{b}} \xi^{-1}\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\tau + \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}))\mathrm{d}\mathbf{x},$$

where $\xi = \Phi(\tilde{\tau})$, with $\tilde{\tau} = \tau/(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}$.

We are interested in evaluating the integral

$$\mathcal{F}_{\boldsymbol{\kappa}}^p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \int_{\mathbf{a}}^{\mathbf{b}} \mathbf{x}^{\boldsymbol{\kappa}}\xi^{-1}\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\tau + \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}))\mathrm{d}\mathbf{x}. \tag{5}$$

The boundary condition is obviously $\mathcal{F}_{\mathbf{0}}^p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$. When $\boldsymbol{\lambda} = \mathbf{0}$ and $\tau = 0$, we recover the multivariate normal case, and then $\mathcal{F}_{\boldsymbol{\kappa}}^p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{0}, 0) \equiv F_{\boldsymbol{\kappa}}^p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \mathbf{x}^{\boldsymbol{\kappa}}\phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\mathbf{x}$, with boundary condition $\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{0}, 0) \equiv L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \phi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{d}\mathbf{x}$. Note that we use calligraphic style for the integrals of interest $\mathcal{F}_{\boldsymbol{\kappa}}^p$ and $\mathcal{L}_p$ when we work with the skewed version.

Theorem 3. *For $p \geq 1$ and $i = 1, \ldots, p$,*

$$\mathcal{F}^p_{\boldsymbol{\kappa}+\mathbf{e}_i}(\mathbf{a},\mathbf{b};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) = \mu_i \mathcal{F}^p_{\boldsymbol{\kappa}}(\mathbf{a},\mathbf{b};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) + \delta_i F^p_{\boldsymbol{\kappa}}(\mathbf{a},\mathbf{b};\boldsymbol{\mu}-\boldsymbol{\mu}_b,\boldsymbol{\Gamma}) + \mathbf{e}_i^\top \boldsymbol{\Sigma}\mathbf{d}_{\boldsymbol{\kappa}}, \tag{6}$$

where $\eta = \phi_1(\tau;0,1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})/\xi$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_p)^\top = \eta\boldsymbol{\Sigma}\boldsymbol{\varphi}$, $\boldsymbol{\varphi} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda}$, $\boldsymbol{\mu}_b = \tau\boldsymbol{\Gamma}\boldsymbol{\varphi}$, $\boldsymbol{\Gamma} = [\gamma_{ij}] = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{\Psi} = \mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, and $\mathbf{d}_{\boldsymbol{\kappa}}$ is an p-vector with jth element

$$\begin{aligned} d_{\boldsymbol{\kappa},j} \;=\; & k_j \mathcal{F}^p_{\boldsymbol{\kappa}-\mathbf{e}_j}(\mathbf{a},\mathbf{b},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) + a_j^{k_j}\, ESN_1(a_j;\mu_j,\sigma_j^2,c_j\sigma_j\tilde{\varphi}_j,c_j\tau)\mathcal{F}^{p-1}_{\boldsymbol{\kappa}_{(j)}}(\mathbf{a}_{(j)},\mathbf{b}_{(j)};\tilde{\boldsymbol{\mu}}_j^{\mathbf{a}},\tilde{\boldsymbol{\Sigma}}_j,\tilde{\boldsymbol{\Sigma}}_j^{1/2}\boldsymbol{\varphi}_{(j)},\tilde{\tau}_j^{\mathbf{a}}) \\ & - b_j^{k_j}\, ESN_1(b_j;\mu_j,\sigma_j^2,c_j\sigma_j\tilde{\varphi}_j,c_j\tau)\mathcal{F}^{p-1}_{\boldsymbol{\kappa}_{(j)}}(\mathbf{a}_{(j)},\mathbf{b}_{(j)};\tilde{\boldsymbol{\mu}}_j^{\mathbf{b}},\tilde{\boldsymbol{\Sigma}}_j,\tilde{\boldsymbol{\Sigma}}_j^{1/2}\boldsymbol{\varphi}_{(j)},\tilde{\tau}_j^{\mathbf{b}}), \end{aligned} \tag{7}$$

where

$$\tilde{\boldsymbol{\mu}}_j^{\mathbf{a}} = \boldsymbol{\mu}_{(j)} + \boldsymbol{\Sigma}_{(j),j}\frac{a_j-\mu_j}{\sigma_j^2}, \quad \tilde{\boldsymbol{\mu}}_j^{\mathbf{b}} = \boldsymbol{\mu}_{(j)} + \boldsymbol{\Sigma}_{(j),j}\frac{b_j-\mu_j}{\sigma_j^2}, \quad \tilde{\varphi}_j = \varphi_j + \frac{1}{\sigma_j^2}\boldsymbol{\Sigma}_{j(j)}\boldsymbol{\varphi}_{(j)}, \quad \tilde{\boldsymbol{\Sigma}}_j = \boldsymbol{\Sigma}_{(j),(j)} - \frac{1}{\sigma_j^2}\boldsymbol{\Sigma}_{(j),j}\boldsymbol{\Sigma}_{j,(j)},$$
$$c_j = \frac{1}{(1+\boldsymbol{\varphi}_{(j)}^\top\tilde{\boldsymbol{\Sigma}}_j\boldsymbol{\varphi}_{(j)})^{1/2}}, \qquad \tilde{\tau}_j^{\mathbf{a}} = \tau + \tilde{\varphi}_j(a_j-\mu_j), \qquad \text{and} \qquad \tilde{\tau}_j^{\mathbf{b}} = \tau + \tilde{\varphi}_j(b_j-\mu_j).$$

When $k_j = 0$, the first term in (7) vanishes. When $a_j = -\infty$, the second term vanishes, and when $b_j = \infty$, the third term vanishes.

3.2 Mean and covariance matrix of multivariate TESN distributions

Let consider $\mathbf{Y} \sim \mathrm{TESN}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau,[\mathbf{a},\mathbf{b}])$. In light of Theorem 3, we have that

$$\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} + \frac{1}{\mathcal{L}}[L\boldsymbol{\delta} + \boldsymbol{\Sigma}(\mathbf{q}_a - \mathbf{q}_b)], \tag{8}$$

where $\mathcal{L} \equiv \mathcal{L}_p(\mathbf{a},\mathbf{b};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau)$, $L \equiv L_p(\mathbf{a},\mathbf{b};\boldsymbol{\mu}-\boldsymbol{\mu}_b,\boldsymbol{\Gamma})$ and the j-th element of $\mathbf{q}_a$ and $\mathbf{q}_b$ are

$$q_{a,j} = ESN_1(a_j;\mu_j,\sigma_j^2,c_j\sigma_j\tilde{\varphi}_j,c_j\tau)\mathcal{L}_{p-1}(\mathbf{a}_{(j)},\mathbf{b}_{(j)};\tilde{\boldsymbol{\mu}}_j^{\mathbf{a}},\tilde{\boldsymbol{\Sigma}}_j,\tilde{\boldsymbol{\Sigma}}_j^{1/2}\boldsymbol{\varphi}_{(j)},\tilde{\tau}_j^{\mathbf{a}}), \tag{9}$$
$$q_{b,j} = ESN_1(b_j;\mu_j,\sigma_j^2,c_j\sigma_j\tilde{\varphi}_j,c_j\tau)\mathcal{L}_{p-1}(\mathbf{a}_{(j)},\mathbf{b}_{(j)};\tilde{\boldsymbol{\mu}}_j^{\mathbf{b}},\tilde{\boldsymbol{\Sigma}}_j,\tilde{\boldsymbol{\Sigma}}_j^{1/2}\boldsymbol{\varphi}_{(j)},\tilde{\tau}_j^{\mathbf{b}}). \tag{10}$$

Denoting $\mathbf{D} = [\mathbf{d}_{\mathbf{e}_1},\ldots,\mathbf{d}_{\mathbf{e}_p}]$, we can write

$$\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \boldsymbol{\mu}\mathbb{E}[\mathbf{Y}]^\top + \frac{1}{\mathcal{L}}[L\boldsymbol{\delta}\mathbb{E}[\mathbf{W}]^\top + \boldsymbol{\Sigma}\mathbf{D}], \tag{11}$$
$$\mathrm{cov}[\mathbf{Y}] = [\boldsymbol{\mu} - \mathbb{E}[\mathbf{Y}]]\mathbb{E}[\mathbf{Y}]^\top + \frac{1}{\mathcal{L}}[L\boldsymbol{\delta}\mathbb{E}[\mathbf{W}]^\top + \boldsymbol{\Sigma}\mathbf{D}], \tag{12}$$

where $\mathbf{W} \sim \mathrm{TN}_p(\boldsymbol{\mu}-\boldsymbol{\mu}_b,\boldsymbol{\Gamma},[\mathbf{a},\mathbf{b}])$, that is a p-variate truncated normal distribution on $[\mathbf{a},\mathbf{b}]$.

Sample codes for this approach using our `MomTrunc` R package and a comparison of the processing time vs. Monte Carlo (MC) simulations based on 500 samples, can be found in the supplementary material.

3.3 On moments of folded multivariate ESN distributions

Let $\mathbf{X} \sim ESN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau)$, we now turn our attention to discuss the computation of any arbitrary order moment of $|\mathbf{X}|$. First, we established the following corollary from Theorem 2.

Corollary 1. *If $\mathbf{X} \sim ESN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau)$ then $\mathbf{Z}_s = \boldsymbol{\Lambda}_s\mathbf{X} \sim ESN_p(\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s,\boldsymbol{\lambda}_s,\tau)$ and consequently the joint* pdf *and the $\boldsymbol{\kappa}$th raw moment of $\mathbf{Y} = |\mathbf{X}|$ are, respectively, given by*

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{s}\in S(p)} ESN_p(\mathbf{y}_p;\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s,\boldsymbol{\lambda}_s,\tau) \qquad \text{and} \qquad \mathbb{E}[\mathbf{Y}^{\boldsymbol{\kappa}}] = \sum_{\mathbf{s}\in S(p)} \mathcal{I}^p_{\boldsymbol{\kappa}}(\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s,\boldsymbol{\lambda}_s,\tau),$$

where $\mathbf{y}_s = \boldsymbol{\Lambda}_s\mathbf{y}$, $\boldsymbol{\mu}_s = \boldsymbol{\Lambda}_s\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_s = \boldsymbol{\Lambda}_s\boldsymbol{\Sigma}\boldsymbol{\Lambda}_s$, $\boldsymbol{\lambda}_s = \boldsymbol{\Lambda}_s\boldsymbol{\lambda}$ and $\mathcal{I}^p_{\boldsymbol{\kappa}}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau) = \int_0^\infty \mathbf{y}^{\boldsymbol{\kappa}} ESN_p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\tau)\mathrm{d}\mathbf{y}$.

Note that $\mathcal{I}_{\boldsymbol{\kappa}}^{p}$ is a special case of $\mathcal{F}_{\boldsymbol{\kappa}}^{p}$ that occurs when $a_i = 0$ and $b_i = +\infty$, $i = 1, \ldots, p$. In this scenario we have $\mathcal{I}_{\boldsymbol{\kappa}}^{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \mathcal{F}_{\boldsymbol{\kappa}}^{p}(\mathbf{0}, +\boldsymbol{\infty}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$. When $\boldsymbol{\lambda} = \mathbf{0}$ and $\tau = 0$, that is, the normal case we write $\mathcal{I}_{\boldsymbol{\kappa}}^{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{0}, 0) = I_{\boldsymbol{\kappa}}^{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From Corollary above, we have that the mean and variance-covariance matrix can be calculated as a sum of 2^p terms as well. In Subsection 3.3.2, we propose another scheme that will let us circumvent this cumbersome calculation.

3.3.1 Univariate case

Supposing that $X \sim \mathrm{ESN}(\mu, \sigma^2, \lambda, \tau)$, a direct consequence of Corollary 1 is that

$$\mathbb{E}[|X|^k] = \mathcal{I}_k^1(\mu, \sigma^2, \lambda, \tau) + \mathcal{I}_k^1(-\mu, \sigma^2, -\lambda, \tau).$$

Thus, using the recurrence relation on $\mathcal{I}_k$ and the notation in theorem 3 (using lowercase Greek symbols instead), next we present explicit expressions for $\mathbb{E}[|X|^k]$ with $k = 1, 2$, these are given by

$$\mathbb{E}[|X|] = \mu(1 - 2\tilde{\Phi}_1(0; \mu, \sigma^2, \lambda, \tau)) + 2\sigma^2 ESN_1(0; \mu, \sigma^2, \lambda, \tau) + \lambda\eta\sigma(1 - 2\Phi_1(0; \mu - \mu_b, \gamma^2)), \quad (13)$$

$$\mathbb{E}[|X|^2] = \mu^2 + \sigma^2 + \lambda\eta\sigma(2\mu - \mu_b). \quad (14)$$

We have readily obtained the first four raw moments as well as for others univariate folded distributions that are special cases of the ESN distribution. These have been intentionally omitted in this extended abstract.

3.3.2 Explicit expressions for mean and covariance matrix of multivariate folded ESN distribution

Let $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\lambda}^*, \tau^*)$. To obtain the mean and covariance matrix of $|\mathbf{Y}|$ boils down to compute $\mathbb{E}[|Y_i|]$, $\mathbb{E}[|Y_i^2|]$ and $\mathbb{E}[|Y_i Y_j|]$. In subsection 3.3.1, we have already computed $\mathbb{E}[|Y_i|]$ and $\mathbb{E}[|Y_i^2|]$. It remains to obtain $\mathbb{E}[|Y_i Y_j|]$ for $i \neq j$, which can be obtained as

$$\begin{aligned}\mathbb{E}[|Y_i Y_j|] =& \mathcal{I}_{1,1}^2(\mu_i, \mu_j, \sigma_i^2, \sigma_{ij}, \sigma_j^2, \lambda_i, \lambda_j, \tau) + \mathcal{I}_{1,1}^2(\mu_i, -\mu_j, \sigma_i^2, -\sigma_{ij}, \sigma_j^2, \lambda_i, -\lambda_j, \tau) \\ &+ \mathcal{I}_{1,1}^2(-\mu_i, \mu_j, \sigma_i^2, -\sigma_{ij}, \sigma_j^2, -\lambda_i, \lambda_j, \tau) + \mathcal{I}_{1,1}^2(-\mu_i, -\mu_j, \sigma_i^2, \sigma_{ij}, \sigma_j^2, -\lambda_i, -\lambda_j, \tau), \quad (15)\end{aligned}$$

as pointed in Corollary 1, with $(Y_i, Y_j) \sim \mathrm{ESN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ denoting an arbitrary bivariate partition. Using (15) and the recurrence relation on $\mathcal{I}_{\boldsymbol{\kappa}+\mathbf{e}_i}^2$, we can obtain $\mathcal{I}_{1,1}^2$ setting $\boldsymbol{\kappa} = (1, 0)^\top$ and $\mathbf{e}_2 = (0, 1)^\top$. This leads to an explicit expression for $\mathbb{E}[|Y_i Y_j|]$ given by

$$\begin{aligned}\mathbb{E}[|Y_i Y_j|] =& (\mu_i\mu_j + \sigma_{ij})(1 - 2(\tilde{\Phi}^{(i)} + \tilde{\Phi}^{(j)})) + (\delta_i\mu_j + \delta_j(\mu_i - \mu_{bi}))(1 - 2(\Phi^{(i)} + \Phi^{(j)})) \\ &+ 2\mu_j\left[\sigma_i^2\tilde{\phi}^{(i)}(1 - 2\tilde{\Phi}^{(i)}) + \sigma_{ij}\tilde{\phi}^{(j)}(1 - 2\tilde{\Phi}^{(j)})\right] + 2\sigma_j^2\tilde{\phi}^{(j)}\mathbb{E}[|Y_{i.j}|] \quad (16) \\ &+ 2\delta_j\left[\gamma_i^2\phi(\mu_i; \mu_{bi}, \gamma_i^2)(1 - 2\Phi(0; m_{j.i}, \gamma_{j.i}^2)) + \gamma_{ij}\phi(\mu_j; \mu_{bj}, \gamma_j^2)(1 - 2\Phi(0; m_{i.j}, \gamma_{i.j}^2))\right],\end{aligned}$$

with $Y_{i.j} \sim \mathrm{ESN}_i(\mu_{i.j}, \sigma_{i.j}^2, \sigma_{i.j}\varphi_i, \tau_{i.j})$ and where $\mathbb{E}[|Y_{i.j}|]$ can be readily computed using (13). All other parameters details have been omitted but can be found in found in the main work.

This approach is 10× faster than using corollary 1 and 50× faster than Monte Carlo methods.

4. Discussion and Conclusion

In this paper, we have developed recurrence relations for integrals that involve the density of MESN distributions. These recursions allow fast computation of arbitrary order product moments of TMESN and FMESN distributions as well as the first two moments of the TMESN and FMESN as a by product, generalizing results obtained by Kan & Robotti (2017). We also presented some general results aiming to compute these moments for other multivariate skewed distribution belonging to the location-scale family. Finally, the proposed method has been coded and implemented in the R MomTrunc package, which is available for the users on CRAN repository

References

1. Arismendi, J. C. (2013). Multivariate truncated moments. Journal of Multivariate Analysis, 117, 41-75.
2. Azzalini, A. & Dalla-Valle, A. (1996). The multivariate skew-normal distribution. Biometrika, 83(4), 715-726.
3. Chakraborty, A. K. & Chatterjee, M. (2013). On multivariate folded normal distribution. Sankhya B, 75(1), 1-15.
4. Flecher, C., Allard, D. & Naveau, P. (2010). Truncated skew-normal distributions: moments, estimation by weighted moments and application to climatic data. Metron, 68, 331-345.
5. Kan, R. & Robotti, C. (2017). On moments of folded and truncated multivariate normal distributions. Journal of Computational and Graphical Statistics, 25(1), 930-934.
6. Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution.
7. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 23(1), 223-229.

Least trimmed squares estimators for functional principal component analysis

Holger Cevallos-Valdiviezo[1], Stefan Van Aelst[2], Matias Salibian-Barrera[3]
[1]Escuela Superior Politécnica del Litoral (ESPOL), Facultad de Ciencias Naturales y Matemáticas (FCNM), Km 30.5 Vía Perimetral, Guayaquil, Ecuador
[2]KU Leuven, Department of Mathematics, Section of Statistics, Celestijnenlaan 200B B-3001, Leuven, Belgium
[3]The University of British Columbia, Department of Statistics, 3182 Earth Sciences Building (ESB), Vancouver, BC V6T 1Z4, Canada

Abstract
Classical functional principal component analysis can yield erroneous approximations in presence of outliers. To reduce the infuence of atypical data we propose two methods based on trimming: a multivariate least trimmed squares (LTS) estimator and its coordinatewise variant. The multivariate LTS minimizes the multivariate scale corresponding to hsubsets of curves while the coordinatewise version uses univariate LTS scale estimators. Consider a general setup in which observations are realizations of a random element on a separable Hilbert space H. For a fixed dimension q, we aim to robustly estimate the q dimensional linear space in H that gives the best approximation to the functional data. Our estimators use smoothing to first represent irregularly spaced curves in a high-dimensional space and then calculate the LTS solution on these multivariate data. The solution of the multivariate data is subsequently mapped back onto H. Poorly fitted observations can therefore be flagged as outliers. Simulations and real data applications show that our estimators yield competitive results when compared to existing methods when a minority of observations is contaminated. When a majority of the curves is contaminated at some positions along its trajectory coordinatewise methods like Coordinatewise LTS are preferred over multivariate LTS and other multivariate methods since they break down in this case.

Keywords
Functional Data Analysis; Robust Methods

1. Introduction

For a fixed dimension q, functional principal component analysis (FPCA) aims to estimate the q dimensional linear space that gives the best approximation to the functional data. For instance, the classical approach for FPCA has the property of providing optimal approximations in the L2 sense. However, this approach is very sensitive to abnormal functional data. To reduce the influence of outliers we propose two methods based on trimming: a multivariate least

trimmed squares (MVLTS) estimator and a coordinatewise least trimmed squares (CooLTS) estimator. The MVLTS estimator was introduced in Maronna (2005) for multivariate data. We extend this estimator to the functional case. Moreover, we also introduce the CooLTS estimator for both multivariate data and functional data. For both methods we propose an algorithm based on estimating equations to find a local minimum of their objective function.

For multivariate data we use the following notation. Consider n observations $x_i \in \mathbb{R}^p, i = 1, \dots, n.$ with corresponding sample mean $\bar{x}$ and sample covariance matrix $\mathfrak{S}$. Let $B_q \in \mathbb{R}^{p\times q}$ be an orthogonal matrix, i.e. $B_q^T B_q = I_q$ with rows $b_j^T, j = 1, \dots, p$. Let $A_q \in \mathbb{R}^{n\times q}$ be a matrix with rows $a_i^T, i = 1, \dots, n,$ and $m \in \mathbb{R}^p$. The corresponding approximations of the observations are given by $\hat{x}_i(B_q, A_q, m) \equiv \hat{x}_i = m + B_q a_i$, or elementwise $\hat{x}_{ij} = m_j + a_i^T b_j$. The associated multivariate residuals are given by $r_i = x_i - \hat{x}_i \in \mathbb{R}^p$ with components $r_{ij} = x_{ij} - \hat{x}_{ij}$. Its Euclidean norm is denoted by $d_i(B_q, A_q, m) \equiv d_i = \|r_i\|_{\mathbb{R}_p}$.

2. Methodology

a. Multivariate least trimmed squares estimator in $\mathbb{R}^p$

It is easy to see that the classical PCA solution is found by minimizing a scale estimate $\hat{\sigma}^2(d(B_q, A_q, m))$ of the Euclidean distances of the residuals $d(B_q, A_q, m) = (d_1, \dots, d_n)$, given by

$$\hat{\sigma}^2\left(d(B_q, A_q, m)\right) = \frac{1}{n}\sum_{i=1}^{n} d_i^2(B_q, A_q, m) \qquad (1)$$

This classical scale estimator based on a quadratic loss function is clearly not robust against outliers. Maronna (2005) robustified the classical approach by replacing $\hat{\sigma}^2$ by a least trimmed squares (LTS) scale, defined by

$$\hat{\sigma}_{LTS}^2\left(d(B_q, A_q, m)\right) = \frac{1}{h}\sum_{i=1}^{h} d_{i:n}^2(B_q, A_q, m) = \frac{1}{h}\sum_{i=1}^{n} w_i d_i^2(B_q, A_q, m) \qquad (2)$$

where $d_{(1:n)}(B_q, A_q, m) \le \cdots \le d_{(n:n)}(B_q, A_q, m)$ is the ordered sequence of Euclidean distances and $h = n - [n\alpha]$ for some $0 \le \alpha \le 1$. Note that the weights w_i are:

$$w_i = \begin{cases} 1, & for\ d_{(1:n)} \le \cdots \le d_{(h:n)} \\ 0, & otherwise \end{cases} \qquad (3)$$

The multivariate LTS-stimulator (MVLTS) is now defined as the solution $(\hat{B}_{MVLTS}, \hat{A}_{MVLTS}, \hat{m}_{MVLTS})$ of the minimization problem

$$\min_{B_q, A_q, m} \hat{\sigma}_{LTS}^2\left(d(B_q, A_q, m)\right) \qquad (4)$$

Where $B_q \in \mathbb{R}^{p\times q}$ is an orthogonal matrix.

b. Coordinatewise least trimmed squares estomatpr in $\mathbb{R}^p$

Boente & Salibian-Barrera (2015) noted that the classical PCA problem can be rewritten as

$$\min_{B_q, A_q, m} \sum_{j=1}^{p} s_j^2 \left(B_q, A_q, m\right) \tag{5}$$

A robust alternative can thus be obtained by replacing the nonrobust sample variance by a robust estiator of scale. We use the univariate LTS scale-estimator, which is defined as

$$\hat{\sigma}_{LTS,j}^2 = \frac{1}{h}\sum_{i=1}^{h}(r_{ij}^2)_{i:n} = \frac{1}{h}\sum_{i=1}^{n} w_{ij}\left(x_{ij} - m_j - a_i^T b_j\right)^2 \tag{6}$$

For the jth variable, and where the weights w_{ij} are given by

$$w_{ij} = \begin{cases} 1, & if\ r_{ij}^2 \leq (r_{ij}^2)_{h:n} \\ 0, & if\ r_{ij}^2 > (r_{ij}^2)_{h:n} \end{cases} \tag{7}$$

The corresponding coordinatewise LTS-estimator (CooLTS) is now defined as the solution $(\hat{B}_{CooLTS}, \hat{A}_{CooLTS}, \hat{m}_{CooLTS})$ of the minimization problem

$$\min_{B_q, A_q, m} \sum_{j=1}^{p} \hat{\sigma}_{LTS,j}^2 \left(B_q, A_q, m\right) \tag{8}$$

with B_q an orthogonal matrix.

c. Algorithm

Explicit first-order conditions can be obtained by differentiating (4) or (8) with respect to a_i, b_j and m_j. After setting them to zero and rearranging terms we obtain

$$\sum_{j=1}^{p} w_{ij}\left(x_{ij} - m_j\right) b_j = \left(\sum_{j=1}^{p} w_{ij} b_j b_j^T\right) a_i \,, 1 \leq i \leq n \tag{9}$$

$$\sum_{j=1}^{n} w_{ij}\left(x_{ij} - m_j\right) a_j = \left(\sum_{j=1}^{p} w_{ij} a_i a_i^T\right) b_i \,, 1 \leq j \leq p \tag{10}$$

$$\sum_{j=1}^{p} w_{ij}\left(x_{ij} - a_i^T b_j\right) = \sum_{i=1}^{n} w_{ij} m_j \tag{11}$$

Note that for CooLTS the weights w_{ij} are given by (7) while for MVLTS the weights $w_{ij} = w_i$ are given by (3). These equations suggest an iterative re-weighted least squares procedure to find a local minimum of (4) or (8). As

starting values random orthogonal matrices are generated and as initial location estimate we use the spatial median of the data. The best local minimum that is reached is then the approximation for the global optimum (see also Cevallos-Valdiviezo & Van Aelst (2019);
Boente & Salibian-Barrera (2015)).

d. Functional data

In most applications, curves are only partially observed at different design points $t_{ij}, 1 \leq j \leq m_i, 1 \leq i \leq n$, i.e. $x_{ij} = X_i(t_{ij})$. To extend MVLTS and CooLTS to the functional case we use smoothed robust principal components by the Sieves method introduced in Bali et al. (2011). The Sieves smoothing method uses B—splines as a smoothing tool. Hence, we first project the functional data on a finite dimensional space by using appropriate basis functions, then we estimate the principal components by MVLTS or CooLTS in the finite dimensional space and finally we transform the solution back to the original functional space.

3. Result

To assess robustness of MVLTS and CooLTS with functional data we carried out the experiments in Boente & Salibian-Barrera (2015) which introduce complicated patterns of contamination. We compare our methods to classical PCA (LS) and other robust PCA techniques such as the coordinatewise S-estimator (CooS) of Boente & Salibian-Barrera (2015), the Multivariate S-cstimator (MVS) of Maronna (2005), and the sieve projection-pursuit approach (PP) of Bali et al. (2011). We also included in the comparisons the best q—dimensional linear space (True) according to the data generating process as a benchmark for all methods. For the S-estimates we consider the Tukey's bisquare function for p with constants $c = 1.54764$, $b = 0.50$ and $c = 3$, $b = 0.2426$. For the LTS estimates we consider $\alpha = 0.5$.

We generate functional data from a model with finite-rank process (Model 1) and from a model with an infinite-rank process (Model 2). For Model 1 we estimated one-dimensional approximations while for Model 2 we estimated four-dimensional approximations since this choice explains 95% of the total variance. In all cases $n = 70$ functional observations were generated where each curve was observed at $m = 100$ equidistant instants in the interval $[0,1]$. A fraction ϵ of the curves has been contaminated. Figure 1 shows an example of data generated from Model 1 with $\epsilon = 0.30$ (left) and from Model 2 with $\epsilon = 0.90$ (right). Regular curves are shown in blue while contaminated curves are shown in red color. A total of 500 replications was generated for each setting. A cubic B -spline basis of dimension $p = 50$ was used to project the functional data. To assess the performance of estimators we examine their

mean prediction errors for outlying data and clean data separately ($\overline{Out}$ and $\overline{Clean}$). An FPCA method that is robust will give a high $\overline{Out}$ value and a low $\overline{Clean}$ value.

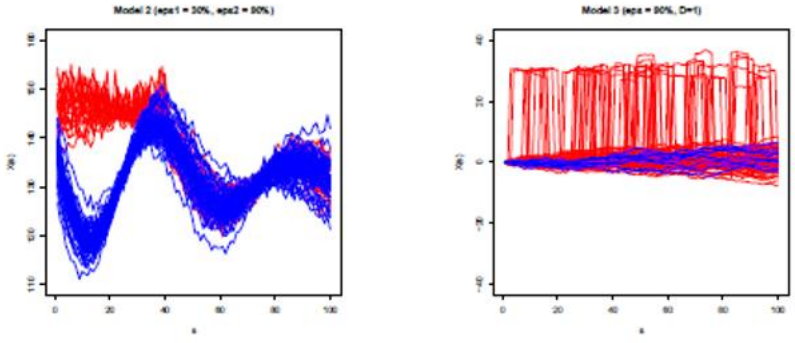


Figure 1: Examples of functional data from Model 1 with $\epsilon = 0.30$ (left) and from Model 2 with $\epsilon = 0.90$ (right), Regular curves are shown in blue color while contaminated curves are shown in red color

Table 1 shows that without contamination classical FPCA (LS) does a good job while the robust methods perform slightly worse. However, for contaminated data classical PCA does not perform well. When the fraction of contamination becomes larger {Model 1, $\epsilon = 0.30$) the CooLTS, MVLYTS and MVS ($c = 1.5$) methods outperform the other procedures. For Model 2 with $\epsilon = 90\%$, we clearly see that the multivariate methods (MVS, MVLVS) break down because too many curves are contaminated. On the other hand. the coordinatewise approaches CooLTS and CooS remain robust, because the fraction of contamination in each coordinate still remains below 50% (see Figure 1).

	Model 1 $\epsilon = 0$	Model 1 $\epsilon = 0.10$		Model 1 $\epsilon = 0.30$		Model 2 $\epsilon = 0.90$	
	Clean	Out	Clean	Out	Clean	Out	Clean
True	1.36	100.59	1.36	100.51	1.36	44.11	0.31
LS	1.34	19.53	4.51	8.48	5.87	36.28	4.31
CooLTS	1.49	100.45	1.52	100.22	1.45	45.53	1.45
CooS(c=1.5)	1.40	97.21	2.30	83.26	4.81	44.81	0.52
CooS(c=3)	1.35	99.23	1.54	16.23	5.55	43.87	0.39
MVLTS	1.38	99.85	1.38	99.40	1.35	42.37	5.84
MVS(c=1.5)	1.34	100.71	1.33	100.62	1.31	37.41	3.95
MVS(c=3)	1.34	100.63	1.33	22.59	5.08	36.55	4.05
PP	1.43	90.70	1.59	55.22	2.94	44.69	0.38

Table 1: Mean prediction errors over 500 replications

4. Discussion and Conclusion

Simulation studies showed that CooLTS and MVLTS yield competitive results when compared to existing methods when a minority of observations is contaminated. When a majority of the curves is contaminated at some positions along its trajectory coordinatewise methods like Coordinatewise LTS are preferred over multivariate LTS and other multivariate methods since they break down in this case.

References

1. Bali, J. L., Boente, G., Tyler, D. E. & Wang, J. L. (2011). Robust functional principal components: A projection-pursuit approach. Annals of Statistics, 39, 2852-2882.
2. Boente, G. & Salibian-Barrera, M. (2015). S-estimators for functional principal component analysis. Journal of the American Statistical Association, 110(511), 1100-1111.
3. Cevallos-Valdiviezo, H. & Van Aelst, S. (2019). Fast computation of robust subspace estimators. Computational Statistics & Data Analysis, 134, 171-185.
4. Maronna, R. A. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. Technometrics, 47, 264-273.

Nowcasting modelling of volatile and non-volatile food prices using crowdsourcing data (case study of some food commodities prices on Lombok island in 2015)

Tiffany Rizkika[1], Wida Siddhikara P[2], Budi Yuniarto[3], Ricky Yordani[3], Setia Pramana[3]

[1]BPS Solok Selatan
[2]BPS Statistics Indonesia
[3]Politeknik Statistika STIS

Abstract

Monitoring of food prices in real-time or called nowcasting is important to maintain price transparency so that the government easy to detect inflation early and the economy becomes more stable and also can reduce the potential of economic turmoil. This research investigate the usage of crowdsourcing data to nowcast the volatile and non-volatile food prices in Mataram, Indonesia. This research uses two approaches to nowcast the prices, i.e., historical data-based to nowcast volatile food prices, and present data-based to nowcast non-volatile food prices. Historical data-based uses two methods, namely statistical modelling using Distributed Lag Model and machine learning using Neural Network RPROP, also using two types of periods daily and weekly. The result of this approaches shows that the suggested model is Distributed Lag Model, and the recommended data period is weekly. While nowcast based on present data uses two methods, they are time series-based (Nowcast Model) and statistical filtering-based which is followed by cubic smoothing spline modelling (IQR-Spline Model, KDE-Spline Model). We found that the IQR-Spline model is better than KDE-Spline for the commodities of long beans, and mackerel. The KDE-Spline model is better than IQR-Spline for the commodities of instant dry noodles, peanuts, and vegetable tomatoes. In time series-based models, the nowcast model with parameter modification has better results than the nowcast model with parameters as in the previous studies.

Keywords

Nowcasting; Food Price; Volatile; Non-Volatile

1. Introduction

Transparency of food prices is needed by government to making a right decision about food price policy. The right decision of this essential parts of economy, can reduce the potential of economic turmoil. To monitor the price of goods and services at consumer level, BPS-Statistics of Indonesia conducts Consumer Price Survey collecting consumer price data from markets. The

survey is done every week for volatile commodities, and every two weeks for non-volatile products. The publication of this data is published every month.

In making decisions related to fluctuating food prices, data with a collection time very close to the time of decision making is needed. The closer the time span, the better decisions can be made. Currently the consumer prices are published every month. Nowadays when quick decision making on prices with rapid changes, more frequent and almost real time data and prediction so called now casting is needed. The term "nowcasting" is originally used in meteorology to forecast weather in the present and in the next few hours. Several approaches can be implemented to monitor the food prices in almost real-time to maintain price transparency therefore government will be easy detecting inflation early to reach the stability of economy.

Pramana, et al. (2016) shows a proof of concept of using crowdsourcing data as one resource for food price nowcasting. They also discuss that in general there is a similar movement of official price data patterns between several commodities that collected by BPS-Statistics Indonesia, with data collected through crowdsourcing techniques carried out by Pulse Lab Jakarta. They used simple data cleaning and smoothing for nowcasting and only few commodities.

One of big challenges in big data is its veracity and followed by the analysis. Hence, further research on nowcasting technique required, including filtering and modelling, to obtain accurate prediction is needed. Therefore, in this study, several nowcasting approaches are investigated to obtain the best nowcast for food price data.

2. Methodology

a. Data Sources. This study uses the following secondary sources:
    1) Crowdsourcing data obtained from collaboration of Pulse Lab Jakarta, United Nation's Food and Organization (FAO), and Premise Data Corporation in 2015.
    2) Market data obtained from google to identify places/markets
    3) Official Consumer Price Data, from BPS Statistics Indonesia consumer price survey.
b. Food Commodities:
    1) Volatile: chicken beef, egg, onion, chili, low quality rice, and premium quality rice.
    2) Non-volatile: mackerel, long bean, instant dry noodles, peanuts, and vegetable tomatoes.
c. This study is taking place at Lombok Island, during March until July 2015.

3. Nowcasting

We use the different nowcasting approaches based on the volatility of the food prices:

a. Volatile food price nowcasting

Historical data-based uses two methods, namely statistical modelling using Distributed Lag Model (DLM) and machine learning using Neural Network RPROP (NN RPROP), also using two types of periods daily and weekly. Before modelling, pre-processing techniques are applied to prepared data.

The step of data pre-processing for historical data-based approaches is data cleaning, data transformation, smoothing, and data imputation. Data cleaning includes filtering to filter data according to time and place of study, removing incomplete record, extreme prices, and outliers. Data transformation includes standardizing the unit price and calculating daily and weekly price. Smoothing is used to minimize the fluctuation pattern, using smoothing spline. Data imputation is needed to complete the unavailable data in a certain day, also to get daily price from weekly price using temporal disaggregation. After pre-processing, data is divided into training and testing data. Training data is used to making models, while testing data is used to calculate MAPE to get the best model.

Modelling using DLM involves data in the current and past time of the independent variable X. According to Baltagi (2011), DLM is a dynamic model that have a form as follows:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-1} + u_t \quad (1)$$

$$t = 1,2,\ldots,T \quad (2)$$

where, Y_t is the t-th observation on the dependent variable X, and X_ (t-s) is an independent variable of observation. α is an intercept, and β_0, β_1, ... β_s are coefficients at the present time and at lag time, and u_t is a stationary error. The lag is the time required for X independent variables to influence non-independent variables Y (Supranto, 1995). One type of distributed lag model is the infinite lag model, where the length of the lag is known or determined. The amount of lag will affect the number of observations used as samples. The more number of lags, the lower the number of sample data.

Modelling using Neural Network solves the problem by learning from training examples (Michael,2015). The algorithm that applied is Resilient Backpropagation implements two stages of learning, namely the forward propagation stage to get the output error and the backward propagation stage to change the values of weights. Changing the weight and network bias in NN RPROP, according to the gradient behaviour in each training

epoch, so that the number of epochs needed to reach the desired target is far less.

b. Non-volatile food price nowcasting

Nowcast based on present data use two methods, i.e., time series-based (Nowcast Model) and statistical filtering-based which is followed by cubic smoothing spline modelling (IQR-Spline Model, KDE-Spline Model). There are several data pre-processing techniques applied in this study, namely data cleaning, data integration, data reduction, and data transformation. Because of preprocessing techniques are not mutually exclusive, so deep the process does not have to work separately, it can be done simultaneously (Han, Kamber, & Pei, 2012).

1. Data Cleaning. Application of data cleaning in statistical filtering-based (IQR.Spline Model, KDE.Spline Model) is to eliminate missing value. Whereas in time series-based (Nowcast Model), apply imputation to overcome the missing value using formula that include in Nowcast Model (Kim, Cha, & Lee, 2017). The statistical filtering-based will eliminate extreme data (outliers) before continuing with cubic smoothing spline modelling. Two outlier detections are used in statistical filtering-based, namely parametric (Interquartile Range-IQR) and non-parametric (Kernel Density Estimation-KDE) detections.
2. Data integration is implemented to detect fraudulent in crowdsourcing data. The addition of market identity data can illustrate how data is distributed according to commodities, markets, days and time.
3. Data Reduction. The data is adjusted to the scope of research based on commodities, location, and time.
4. Data Transformation is carried out to standardize units according to their commodities.

Here is the cubic smoothing spline formula which is used for statistical filtering-based modelling (Model 1 and Model 2):

$$PLS(f) = \underbrace{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}_{(a)} + \underbrace{\lambda \int [f''(x)]^2\, dx}_{(b)}$$

In that function, part (a) is a function of the distance between the data and the estimate or the number of squares, section (b) is a measure of the smoothness of the curve in data roughness penalty. Lambda (λ) is a smoothing parameter as a balance controller between the compatibility of the data and the smoothness of the curve which has a value range of 0 <λ <1. The greater the lambda value, the greater the smoothness of weight and the smaller the variance produced. The value of f (x) used in the PLS function is obtained from the P_i (x) polynomial. The polynomial used is the third order polynomial (k =

3), natural cubic spline. Natural cubic spline is a continuous segmentation function and remains continuous on all derivatives so that it makes it very smooth.

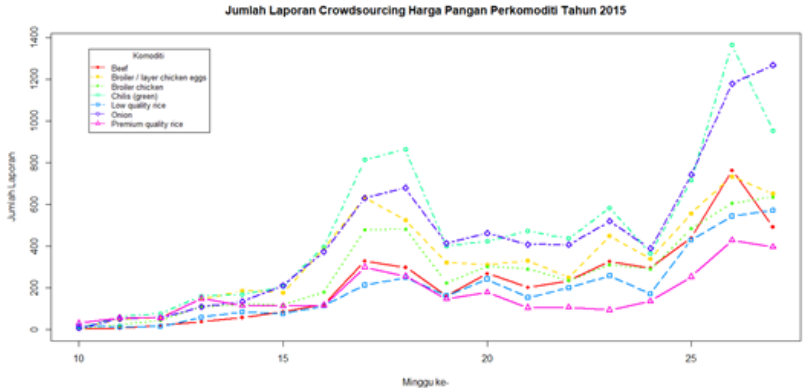
Model 1 uses a parametric aprroach to detect outlier (IQR), so the resulting model is called IQR.Spline. While **Model 2** uses a non-parametric approach to detect outlier (KDE), so the resulting model is called KDE.Spline. The time series-based will use previous period data in modelling, using the Nowcast Model proposed by Kim, Cha, and Lee (2017).

4. Result

a) Historical Data-Based (Volatile food price nowcasting)

1) Visualization Crowdsourcing Data

The picture below shows the number of crowdsourcing data report in weekly periods every commodity. It shows that the number of report is unstable that cause the frequency of data wasn't same everytime. So before modelling, data will be preprocessed.



2) Modelling with Training Data

Commodity	DLM Lag 1		Neural Network (2)		Neural Network (3)		Neural Network (3,2)	
	Daily	Weekly	Daily	Weekly	Daily	Weekly	Daily	Weekly
Chicken	0,08332466	0,059171286	0,0827	0,0477	0,0831	0,0592	0,0834	0,0579
Beef	0,001968903	0,001844379	0,0020	0,0018	0,0020	0,0018	0,0020	0,0018
Egg	0,019875244	0,02287047	0,0171	0,0159	0,0171	0,0221	0,0174	0,0230
Chili	0,077096627	0,058217853	0,0578	0,1428	0,0757	0,1214	0,0651	0,1203
Onion	0,108276312	0,099677708	0,1024	0,0675	0,0970	0,0712	0,0937	0,0308
Low quality rice	0,046325108	0,038070219	0,0266	0,0172	0,0260	0,0081	0,0256	0,0222
Premium quality rice	0,060732601	0,058284652	0,0424	0,0233	0,0334	0,0165	0,0174	0,0147

Table above shows the MAPE value of modelling using Distributed Lag Model with Lag 1, NN RPROP with (2) Neuron, (3) Neuron, and (2,3) Neuron. Also with the different period of data, daily and weekly. The minimum MAPE for DLM model (cell with yellow highlight) is dominated by the model with weekly period data, where 6 commodities have the model with lower MAPE value. While the minimum MAPE value for NN RPROP model (cell with blue highlight) also dominated with by the model with weekly period data, but in different number of neuron.

3) Obtain the Best Model with Testing Data

After getting the best period of data in each modelling technique, testing data is used to get the best technique between DLM and RPROP. The result is shows below.

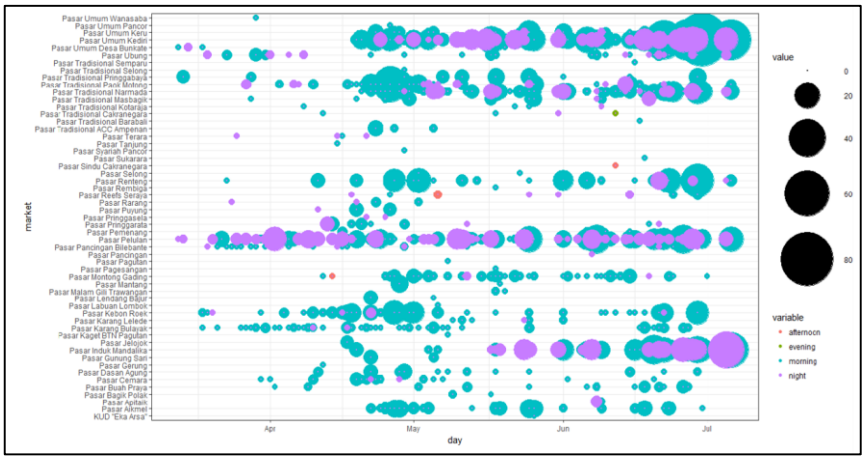
Tabel 3.4 The Best Model Result

Commodity	Model	Period
Chicken	$Y_t = -16146.441 + 1.03X_t + 0.55X_{t-1}$	Weekly
Beef	$Y_t = 99547.486 + 0.012X_t + 0.067X_{t-1}$	Weekly
Egg	1 Hidden layer dengan 2 neuron	Weekly
Chili	$Y_t = 9253.229 + 0.186 + 0.643X_{t-1}$	Weekly
Onion	$Y_t = 26177.519 + 0.773X_t - 0.835X_{t-1}$	Weekly
Low quality rice	2 Hidden layer with 3 and 2 neuron	Weekly
High quality rice	$Y_t = 7554.825 - 0.807X_t + 0.968X_{t-1}$	Weekly

b) Present Data-Based (Non-volatile food price nowcasting)

1) Visualization Data

The following is an illustration of the distribution of crowdsourcing data on mackerel commodities. This visualization is obtained after integrating data with market identity data. It is seen that the more days the officers have a tendency to collect data in the morning and tend to gather in certain places.



2) Statistical Filtering-based Model

Here is MAPE value of IQR.Spline dan KDE.Spline Model :

Commodity	MAPE IQR Smooth Spline	MAPE KDE Smooth Spline
(1)	(2)	(3)
Long Bean	16,95	17,83
Mackerel	18,01	25,04
Instant Dry Noodles	25,62	0,81
Peanuts	118,42	62,33
Vegetable Tomato	15,33	15,03

Based on table above, IQR.Spline Model is better for commodities Long Bean and Mackerel, while KDE.Spline is better for commodities Instant Dry Noodles, Peanuts, and Vegetable Tomato.

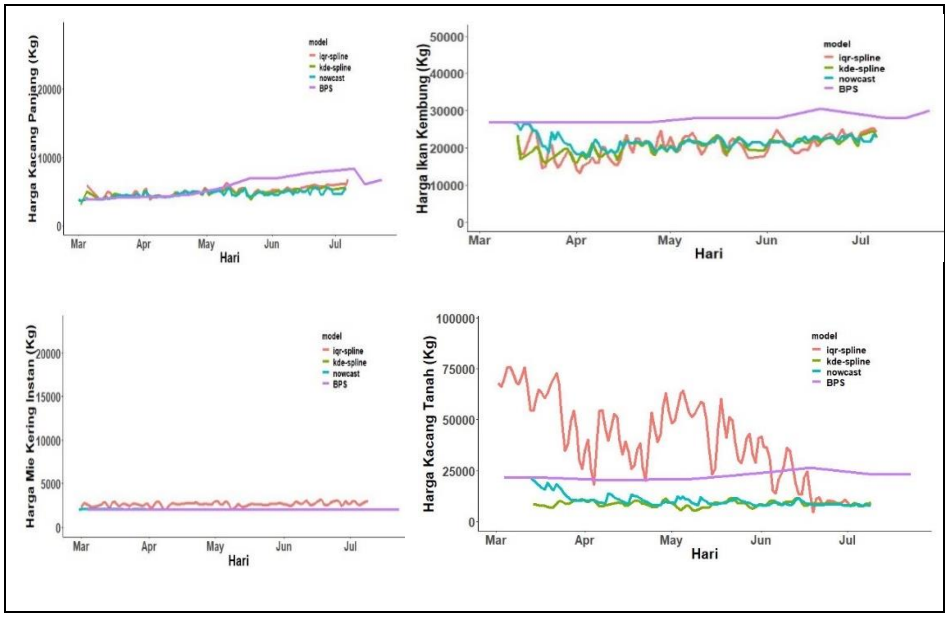
3) Time series-based Model

Here is MAPE value of time series-based model:

Komoditas	MAPE Value					
	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Long Bean	16,06	16,06	25,73	18,95	16,40	27,29
Mackerel	34,69	34,69	34,69	23,05	22,14	21,75
Instant Dry Noodles	20,43	20,43	27,60	1,18	0,97	10,34
Peanuts	106,46	106,46	145,41	62,01	49,03	71,92
Vegetable Tomato	25,93	25,93	26,78	26,19	13,02	27,04

Based on the table above, model 7 is the best of Nowcast Model.

4) Comparison Model for Non-volatile food price
Here is comparison graph of HK-BPS price, IQR.Spline Model, KDE.Spline Model, Nowcast Model (Model 7).



5. Conclusion

1. The study has shown that the best model to nowcast volatile food price (weekly) is Distributed Lag Model.
2. For non-volatile commodities, the IQR-Spline model perform best for long bean and mackerel commodities whereas the KDE-Spline model best for the commodity of instant dry noodles, peanuts, and vegetable tomato, long beans, and mackerel.

References

1. BPS. (2012). Diagram Timbang Indeks Harga Konsumen Hasil Survei Biaya Hidup 2012. Jakarta.
2. Baltagi, B. H. (2011). Distributed Lags and Dynamic Models. In *Econometrics* (S. 131-147).
3. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0
4. Kim, J., Cha, M., & Lee, J. G. (2017). Nowcasting commodity prices using social media. PeerJ Computer Science, 3, e126. https://doi.org/10.7717/peerj-cs.126
5. Michael, A. N. (2015). *Neural Networks and Deep Learning.* Determination Press.
6. Pramana, S., Yuniarto, B., Kurniawan, R., Lee, J., Amin, I., Putu, N. L., … Riyadi, Y. (2016). Big Data for Government Policy : Potential Implementations of BigData for Official Statistics in Indonesia.
7. Premise. (2015). Food security in Indonesia.
8. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/. 2014

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 978-967-2000-70-9

#ISIWSC2019