62nd ISI World Statistics Congress

ISI 2019 Kuala Lumpur

18 - 23 August 2019

PROCEEDING

CONTRIBUTED PAPER SESSION

VOLUME 7

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 7)

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 7, 2019. 403 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 7

Data boosting on short bivariate time series data by sieve bootstrap

Ma. Salvacion B. Pantino[1], Erniel B. Barrios[2], Joseph Ryan G. Lansangan[2]

[1] Mathematics Program, College of Science, University of the Philippines Cebu, Philippines
[2] School of Statistics, University of the Philippines Diliman, Quezon City, Philippines

Abstract

A model is postulated given a short bivariate time series data with high-dimensional inputs. The correlated response vectors are functions of the contemporaneous effects of the input series. The model is then estimated using a hybrid of methods embedded into the backfitting algorithm. It is noted from the simulation studies that the estimation procedure produces parameter estimates with lower relative bias and better predictive ability compared to $VAR(1)$. The estimation method is also robust to misspecification errors.

Keywords

Short bivariate time series data; high-dimensional inputs; backfitting algorithm

1. Introduction

Data are being collected at various lengths and frequencies depending on their availability and cost. Optimizing the use of the database of various industries have been the trend towards answering questions about productivity, prediction, and diagnoses of arising problems that could be resolved by crafting solutions based on what was observed. Considering the contemporaneous effects of a set of variables has been a concern towards determining their dynamic effects over time. More so, insightful explanations could be derived if a wide array of covariates would be considered to explain two or more interrelated variables of interest.

Shen and Huang (2008) proposed the sparse principal component analysis via regularized singular value decomposition (sPCA-rSVD) which solves a low-rank matrix approximation problem by imposing regularization penalties to produce sparse principal component (PC) loadings. Witten et al. (2009) developed a method using penalized matrix decomposition (PMD) which decomposes the wide array of covariates using sparse vectors.

The bootstrap method of resampling proposed by Efron (1979) has become a powerful nonparametric method for estimating the distribution of a statistical quantity. Bühlmann (1997) developed the sieve bootstrap as a method of generating a bootstrap sample by resampling from the residuals of

a time series model as it produces consistent estimators of the variance of time series statistics.

Nonparametric multivariate analysis simultaneously tracks the effects of the array of explanatory variables on multiple variables of interest (response variables) while letting the data speak for itself. Opsomer and Ruppert (1997) came up with a bivariate additive model and fitted it by local polynomial regression and showed that it has the same rate of convergence as its univariate counterpart.

Friedman and Stuetzle (1981) suggested the use of an additive model which assumes that the conditional expectation function of the response variable can be written as the sum of smooth terms of covariates. Buja et al. (1989) estimated an additive nonparametric regression model using linear smoothers by the backfitting algorithm and provided proof on the convergence of the method. The backfitting algorithm sequentially estimates parameters of interest until convergence in an iterative manner. Hastie and Tibshirani (1986) proposed the generalized additive model (GAM) which replaces the sum of the linear covariates by a sum of smooth functions which are iteratively estimated by the local scoring algorithm. Opsomer (2000) derived recursive asymptotic bias and variance expressions for the backfitting estimators on local polynomial regression smoothers.

This paper focuses on estimating the short bivariate time series given the contemporaneous effects of predictors. The parameters in the bivariate additive model will be estimated using a backfitting framework. The vector autoregression at order one (VAR(1)) is used to estimate the output autocorrelation coefficient ρ. The Generalized Additive Model (GAM) is used for regressing the sparse components with the bivariate output series Y_t. It is of interest to characterize the underlying empirical distribution function of the estimates. After achieving initial parameter estimates, the residuals is used for the sieve bootstrap procedure to give the final parameter estimates. The performance of the model is evaluated through a simulation study and application to the short data about the University of the Philippines on teaching, research, and extension programs for the last two decades (1995 - 2015).

2. Methodology

The postulated model [2] is compared to the $VAR(1)$ process defined as

$$Y_t = Y_{t-1}\rho + E_t, \qquad [1]$$

where ρ is a 2 × 2 output autocorrelation matrix coefficient of the immediate past Y_{t-1} of a given bivariate time series data $Y_t = (y_{1t}, y_{2t})$. E_t is a two-dimensional white noise process with time invariant positive definite covariance matrix $E(e_t e_t{}') = \Sigma_E$.

The postulated model accounts the additional contribution of the contemporaneous effects of some input series. This study is proposing an estimation procedure for an additive bivariate model

$$Y_t = Y_{t-1}\rho + \sum_{k=1}^{p}\sum_{j=1}^{m} f_{k,j}\left(X_{k,t-j}\right) + A_t\,, \quad j = 1,\dots,m, k = 1,\dots,p, t = 1,\dots,T \qquad [2]$$

where

$\sum_{k=1}^{p}\sum_{j=1}^{m} f_{k,j}\left(X_{k,t-j}\right) = \sum_{k=1}^{p}\sum_{j=1}^{m} \beta_{kj} x_{k,t-j}$

$\boldsymbol{Y_t}$ is the bivariate response variable with columns $y_h = (y_{h1}, y_{h2}, \dots, y_{hT})', h = 1,2$

$\boldsymbol{\rho}$ is the autocorrelation coefficient of $\boldsymbol{Y_{t-1}}$, i.e. $\rho = \begin{bmatrix}\rho_{11} & \rho_{12}\\ \rho_{21} & \rho_{22}\end{bmatrix}, 0 \le \rho_{iq} \le 1, i, q = 1,2$

$\boldsymbol{X_{t-j}}$ is the covariate matrix up to the m^{th} lag, i.e. for each p X variable,

$$x_{k,t-j} = \begin{bmatrix} x'_{p,t-1} \\ x'_{p,t-2} \\ \vdots \\ x'_{p,t-m} \end{bmatrix}, k = 1,\dots,p$$

$\boldsymbol{\beta_{kj}}, k = 1,\dots,p,\ j = 1,\dots,m$, is the coefficient matrix of $\boldsymbol{X_{t-j}}$

$\boldsymbol{A_t}$ is the error matrix with columns $a_1, a_2 \sim N(0, {\sigma_r}^2), {\sigma_r}^2 > 0, r = 1,2$ and $Cov(a_1, a_2) = \sigma_{12} \neq 0$

The postulated model states that at a given time point t, the variation in the bivariate output series can be attributed to a combination of additive components and some error component A_t . The first component is the VAR(1) or the linear function of the first lag of the bivariate output series. The contribution of the input series is summarized by the second component which is a function of the contemporaneous effects of the input series. The wide array of the m lagged values of p input series contained in the covariate matrix X_{t-j} is of dimension, $T \times p * m$, where $T < p * m$.

The proposed bivariate additive model [2] illustrates how the contemporaneous effects of input variables could simultaneously affect a pair of response variable at time t. This is commonly encountered on analyzing economic indices that are usually of limited length and are affected by the contemporaneous effects of numerous factors.

The proposed two-step estimation procedure for the bivariate additive model includes (1) identifying the sparse principal components using SPCA, and (2) performing a nonparametric regression with the sieve bootstrap procedure. The sparse principal component analysis (SPCA) procedure by Witten et al. (2009) will be used for dimension reduction, and the sieve bootstrap is incorporated in the algorithm to produce consistent parameter estimates. The SPCA includes: selection of the tuning parameter for PMD, computation of single-factor PMD model, and computation of K factors of PMD. The estimation of the model with the sparse principal components (SPCs) derived from the first step is detailed in the second step of the procedure which includes: (1) Algorithm 2 (initial parameter estimation

through a backfitting algorithm) incorporating the VAR(1) model estimation illustrated in Algorithm 1 and the GAM procedure which involves the implementation of the backfitting algorithm and the General Local Scoring Algorithm, and (2) Algorithm 3 (sieve bootstrap on the bivariate residual matrix).

Algorithm 1: $\boldsymbol{VAR(1)}$ to estimate the output vector autocorrelation coefficient $\boldsymbol{\rho}$

The vector autoregression (VAR) model is one of the most successful, flexible, and easy to use models for the analysis of multivariate time series. The VAR model is useful in describing the dynamic behavior of economic and financial time series (Zivot and Wang, 2006).

1. From the proposed bivariate additive model in [2], let $\begin{pmatrix}\varepsilon_{1t}{}'\\ \varepsilon_{2t}{}'\end{pmatrix} = \sum_{k=1}^{p}\sum_{j=1}^{m} f_{k,j}\left(X_{k,t-j}\right) + A_t$ and consider the bivariate $VAR(1)$ model
$$\begin{pmatrix}y_{1t}\\ y_{2t}\end{pmatrix} = \begin{pmatrix}\rho_{11} & \rho_{12}\\ \rho_{21} & \rho_{22}\end{pmatrix}\begin{pmatrix}y_{1t-1}\\ y_{2t-1}\end{pmatrix} + \begin{pmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\end{pmatrix}$$
That is,
$$y_{1t} = \rho_{11}y_{1t-1} + \rho_{12}y_{2t-1} + \varepsilon_{1t}$$
$$y_{2t} = \rho_{21}y_{1t-1} + \rho_{22}y_{2t-1} + \varepsilon_{2t}$$
where $\rho_{12}, \rho_{21} \neq 0$ and $cov(\varepsilon_{1t}, \varepsilon_{2t}) = \sigma_{12} \neq 0$.
2. Estimate ρ_{jj} by fitting $AR(1)$ model separately for each of y_{1t-1} and y_{2t-1} and $\rho_{jk}, j, k = 1,2, j \neq k$ by the correlation of y_{1t-1} with y_{2t-1}. Substitute these preliminary estimates in the VAR model to compute for the residual vector per time period.

Suppose that the model is expressed as $Y_h = \sum_{j=1}^{p} S_j\left(X_j\right) + \varepsilon$, for h = 1,2, where $s_{j,}$,j = 1, … , p are the smooth functions of the sparse principal components of exogenous variables. The backfitting algorithm adapted from Hastie and Tibshirani (1986) is performed as detailed in the following.

Algorithm 2: Initial Parameter Estimation through Backfitting Algorithm

1. Estimate ρ by VAR(1) as in Algorithm 1: $\boldsymbol{Y_t} = \boldsymbol{Y_{t-1}\rho} + v_t$, where $v_t = \sum_{k=1}^{p}\sum_{j=1}^{m} f_{k,j}\left(X_{k,t-j}\right) + A_t$.
2. Compute residual vector U = $\boldsymbol{Y_t}$ − $\boldsymbol{Y_{t-1}\hat{\rho}}$. This contains information about the covariate effects.
3. Implement the GAM on the SPCs produced in Step 1 with U. Obtain the bivariate fitted values F_t.
4. A new response vector is computed,
$$Y^*{}_t = Y_t - F_t.$$
This will set aside the covariate effects to focus on estimating ρ.
5. Repeat Steps (1) to (4) until minimal changes at 0.01% (convergence) are observed in the values of the estimates in $\hat{\rho}$ and F_t .

This yields the residual matrix $R = Y_t - Y_{t-1}\hat{\rho} - F_t$ that will be used for the sieve bootstrap procedure in the succeeding algorithm.

Algorithm 3: Sieve Bootstrap on the Bivariate Residual Matrix

The Gram-Schmidt process is the generation of a series of quantities by means of scalar products of vectors. Winch (1996) proved that these series of quantities are identical with those that arise in the solution of the normal equations by compact elimination methods. The Gram-Schmidt orthogonalization is applied on the covariance matrix $\hat{\Sigma}$ of the residual matrix R and the orthogonal transformation L is derived.

1. From the residual matrix $R = Y_t - Y_{t-1}\hat{\rho} - F_t$ in Algorithm 2, derive the orthogonal transformation L of the covariance matrix $\hat{\Sigma}$ of R by the Gram-Schmidt orthogonalization [refer to Appendix]. L is the square root matrix of $\hat{\Sigma}$. That is, $\hat{\Sigma} = L'L$.
2. Generate a bivariate U matrix of length T such that $U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$, where $u_1, u_2 \sim iid\ U(0,1)$.
3. Resample rows of U for B times ($B \geq 200$) each with length T. Multiply the orthogonal transformation L to each of the B resampled , with $Var(LU_{(b)}) = L'Var(U_{(b)})L = L'LI = L'L = \hat{\Sigma}, b = 1, \dots, B$.
4. From each of $LU_{(b)}$, recreate B time series $Y_{t_{(b)}} = Y_{t-1_{(b)}}\hat{\rho} + F_t + LU_{(b)}$, for b = 1, … , B, where $\hat{\rho}$ is the initial estimate of ρ and Ft is the estimate of the smooth functions of SPC obtained in Step 1.
5. Estimate the parameters until convergence on each recreated time series as in Algorithm 2.
6. Obtain $\hat{\rho}^{(1)}, \hat{\rho}^{(2)}, \dots, \hat{\rho}^{(B)}$ on each of the recreated time series $Y_{t_{(1)}}, Y_{t_{(2)}}, \dots, Y_{t_{(B)}}$.
7. Average of $\hat{\rho}^{(b)}$, b = 1, … , B is the final estimate of ρ.

Lagged values up to m^{th} of each of the generated $x_{k,t}, k = 1, \dots, p$, is included in the covariate matrix $X_{t-j}, t = 1, \dots, T$ and j = 1, … , m. That is,

$$X_{t-j} = [x_{1,t-1} \quad x_{1,t-2} \quad \dots \quad x_{1,t-m} \quad x_{2,t-1} \quad \dots \quad x_{2,t-m} \quad \dots \quad x_{p-1,t-m} \quad x_{p,t-1} \quad \dots \quad x_{p,t-m}].$$

Each of the $x_{1,t}$, $x_{2,t}$, … , $x_{p,t}$ baseline input time series has the same mean of 20 and has the same variance of 4. The baseline input series is simulated by the AR(1) process:

$$(1 - 0.5B)(x_t - 20) = a_t, \qquad \text{where } a_t \sim N(0,1). \qquad [3]$$

The bivariate time series data Y_t is be generated using its immediate past value Y_{t-1} with the output autocorrelation matrix ρ, the function of the covariate matrix X_{t-j} , and with the error matrix $A_t \sim N_2(0, \Sigma_A)$, where $\Sigma_A = \begin{bmatrix} {\sigma_{11}}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_{22}}^2 \end{bmatrix}$ and $\sigma_{12} \neq 0$. A burn-in period of 1000 is considered for the initialization of values.

To test the sensitivity of the model with the misspecification error, the variance of the error matrix is made 3 or 6 times larger. To achieve robust

results, 200 bootstrap replicates are considered in the estimation procedure and applied to 100 data replicates generated for each scenario. Simulation results are assessed according to the bias of estimates in the autocorrelation coefficient $\rho = \begin{bmatrix}\rho_{11} & \rho_{12}\\ \rho_{21} & \rho_{22}\end{bmatrix}$ and the MAPE on each scenario. The output autocorrelation to be estimated in the simulation studies is set to $\rho = \begin{bmatrix}0.10 & 0.15\\ 0.95 & 0.20\end{bmatrix}$.

The MAPE is computed for each of the 100 data replicates as in [4]. The MAPE per scenario is also computed as in [5] to measure the predictive performance of the postulated procedure. MAPE 1 and MAPE 2 represents the MAPE of the first and second column vector of the bivariate output series, respectively.

$$\boldsymbol{MAPE_{replicate,h}} = \left(\frac{1}{T}\sum_{i=1}^{T}\left|\frac{Y_{h,t} - \widehat{Y_{h,t}}}{Y_{h,t}}\right|\right) \times 100, h = 1,2 \qquad [4]$$

[5]

$$\boldsymbol{MAPE_{scenario,h}} = \frac{1}{100}\sum_{replicate=1}^{100} MAPE_{replicate,h}$$

Instead of the actual bias, the Absolute Percentage of Bias (APB) is computed for each estimate to simplify the presentation of over estimation or under estimation of the output autocorrelation. Given the actual [6]
output autocorrelation $\rho = \begin{bmatrix}\rho_{11} & \rho_{12}\\ \rho_{21} & \rho_{22}\end{bmatrix}$, we have

$$Absolute\ Percentage\ of\ Bias\ of\ \widehat{\rho_{p,q}} = \left|\frac{\rho_{p,q} - \widehat{\rho_{p,q}}}{\rho_{p,q}}\right| \times 100, \qquad p,q = 1,2$$

The VAR(1) estimation on the bivariate output series will be used as a benchmark in assessing the bias of estimates of the output autocorrelation matrix ρ. The MAPE under VAR(1) is also computed to compare the predictive performance of the proposed estimation procedure.

Each of the specified length of time series (t) has corresponding pairs of number of input time series (p) of (m) lags. The covariate matrix X_{t-j} , j = 1, … , m considered for each scenario comes from one of the following sets of t,p, and m values: t = 20, (p, m) = (5, 6), (8, 4); for t = 30, (p, m) = (6, 13), (13, 6), (14, 5); and for t = 50, (p, m) = (5, 16), (6, 12), (13, 6), (16, 5).

3. Result

It is observed that APB of the estimates of the proposed estimation procedure is larger when t = 30,50 compared for cases when t = 20. Most of the APB of estimates of the proposed procedure is less compared with the APB of estimates of VAR(1) over the different series lengths. The procedure consistently produces low MAPE across all the varying length of time series t while VAR(1) produces high MAPE for time series of shorter length. The

procedure remains robust even for shorter t. The estimated MAPE by the procedure is recorded to be at most 15 % which is observed at t = 30.

Lower APB is observed in the estimates when $p < m$ for t = 20 while range of APB remains the same for t = 30,50 when $p < m$ or $p > m$. It is also observed that the APB of autocorrelation coefficient components of small values (ρ_{11} = 0.10 , ρ_{12} = 0.15,ρ_{22} = 0.20) are relatively larger than with the APB of larger coefficient (ρ_{21}=0.95). The MAPE is robust with the choice of p and m over the different series length.

Infusing the misspecification error leads to minimal changes in the expected MAPE. The expected MAPE is also relatively higher for t=30. The estimated MAPE of the proposed procedure is robust and remains superior over VAR(1) in the presence of misspecification error. The procedure consistently produces low MAPE across the specified values of t while VAR(1) has the highest MAPE when t = 20. The APB of the estimates produced by the proposed procedure are also robust in the presence of misspecification error except when $p < m$ for t = 50 wherein changes in APB are relatively higher. Moreover, the standard error of the estimates are lower for t = 20 compared to the standard error of the estimates produced by VAR(1).

Generally, simulation results show that the postulated model is fairly robust to misspecification error. Furthermore, the predictive ability of the estimated model is better compared to VAR(1) over different lengths of time series.

The proposed estimation procedure is applied in a series of short annual data set from 1995 to 2015. The goal of the analysis is to determine how the contemporaneous effects of the total number of graduates from the University of the Philippines (UP) system (excluding UP Manila), the budget allocated for the UP system, and the board exam passing rate of UP for various Professional Regulation Commission (PRC) licensure examinations simultaneously affect the Real Gross Domestic Product (GDP) and Gross Value Added (GVA) in Agriculture, Hunting, Fishing, and Forestry from 1995 to 2015.

The contemporaneous effects of the input series are accounted by taking up to the fourth lagged values of the input series. The first sparse principal component is considered in the analysis as it explains 83.32% of the variability of the input series as shown in Table 1.

The estimated output autocorrelation measures the relationship between the bivariate components $y_{1,t}(log(real\ GDP))$ and $y_{2,t}(\log(\ GVA))$ with their past values and with the past value of the other output component ($y_{1,t}\ to\ y_{2,t-1}\ and\ y_{2,t}\ to\ y_{1,t-1}$). The derived components of the estimate $\hat{\rho}$ as shown in Table 2 explains the following relationship given the contemporaneous effects of the input series up to the fourth lag: a 36.49 % increase in the growth rate of real GDP at t results from a proportionate increase in GDP growth rate at $t-1$; a 36.29 % increase in GDP growth rate at t results from a proportionate increase in GVA growth rate at $t-1$; a 35.04 % increase in GVA growth rate at t results

from a proportionate increase in GDP growth rate at $t-1$; and a 35.14 % increase in GVA growth rate at t results from a proportionate increase in GVA growth rate at $t-1$. $VAR(1)$ produced negative estimates that contradict the expected relationship between the output series.

The proposed estimation procedure's predictive performance is also superior than $VAR(1)$ as shown in Table 3. The proposed procedure has the advantage of including the contemporaneous effects of the input series.

Table 1. The proportion of variance explained by the first K sparse principal components (SPCs) of the reduced covariate matrix with p=10 *and* m=4.

10	83.32 %, 87.20 %, 89.52 %, 90.34 %, 91.64 %, 93.64 %, 94.72 %, 95.74 %, 95.89 %, 96.52 %

Table 2. Estimates of the output autocorrelation coefficient with their corresponding standard errors by the proposed procedure.

	Estimates							
Procedure	$\hat{\rho}_{11}$	s.e.	$\hat{\rho}_{12}$	s.e.	$\hat{\rho}_{21}$	s.e.	$\hat{\rho}_{22}$	s.e.
Proposed	0.3649	0.4586	0.3629	0.4595	0.3504	0.4513	0.3514	0.4519
VAR(1)	-0.0292	0.0292	-0.0330	0.0313	-0.0291	0.0462	-0.0365	0.0451

Table 3 . Estimated MAPE of the proposed procedure and VAR(1)

Estimated MAPE (Proposed procedure)		Estimated MAPE (VAR 1)	
MAPE 1	MAPE 2	MAPE 1	MAPE 2
4.13 %	4.21 %	62.19 %	81.35 %

4. Discussion and Conclusion

An estimation procedure is developed for the postulated model of short bivariate time series with high dimensional inputs. The additive bivariate model is postulated for a pair of correlated series explained by its immediate past and by the contemporaneous effects of some input series. This is to characterize short series that are simultaneously influenced by the contemporaneous effects of some input series over time.

Simulation scenarios affirm that the proposed estimation procedure produces more accurate predictions and better estimates than VAR(1). The proposed estimation procedure has relatively better predictive performance than VAR(1) as reflected in the MAPE (less than 15%) given the varying length of series, number of input series, and lags of input series. Simulation results also show that the predictive performance of the proposed procedure is robust to misspecification error (when variance is three or six times larger). Minimal changes (between 1 % to 7 %) in the estimates as reflected in the absolute percentage bias is observed across the scenarios as the misspecification error is induced in the series.

The proposed estimation procedure is fairly robust as assessed and evaluated by the MAPE and APB. The estimation procedure produces low MAPE over the different series length and has estimates that are robust to misspecification error compared to VAR(1).

The dimension reduction method of SPCA has allowed the contribution of the high dimensional inputs to be incorporated in the estimation process. The embedded methods in the backfitting algorithm helped in yielding robust estimates and in providing a good predictive ability for the estimation procedure. The nonparametric regression with the modified sieve bootstrap method for the bivariate series with correlated components helped in producing consistent estimates. The combined nonparametric methods of backfitting embedded with VAR(1) and GAM plus the residual based bootstrap approach helped in providing better estimates.

References

1. Bühlmann, P., 1997, Sieve bootstrap for time series, Bernoulli, 3(2):123-148.
2. Bühlmann, P., 1998, Sieve Bootstrap for Smoothing in Nonstationary Time Series, The Annals of Statistics, 26(1):48-83.
3. Buja, A., Hastie, T., Tibshirani, R., 1989, Linear smoothers and additive models (with discussion), Annals of Statistics, 17:453-555.
4. Efron, B., 1979, Bootstrap methods: another look at the jackknife, The Annals of Statistics, 7: 1-26.
5. Friedman, J. H., Stuetzle, W., 1981, Projection pursuit regression, Journal of American Statistical Association, 76: 817-823.
6. Hastie, T., Tibshirani, R., 1986, Generalized Additive Models, Statistical Science, 1(3):297-318.
7. Opsomer, J., 2000, Asymptotic Properties of Backfitting Estimators, Journal of Multivariate Analysis, 73:166-179.
8. Opsomer, J., Ruppert, D., 1997, Fitting a bivariate additive model by local polynomial regression, The Annals of Statistics, 25(1):186–211.
9. Shen, H., Huang, J., 2008, Sparse principal component analysis via regularized low rank matrix approximation, Journal of Multivariate Analysis, 99: 1015-1034.
10. Winch, D. E., 1966, A criticism of the Gram-Schmidt Orthogonalization process applied to spherical harmonic analysis, Journal of Geophysical Research, 71(21): 5165–5170.
11. Witten, D., Tibshirani., R., Hastie, T., 2009, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 10(3):515–534.
12. Zivot, E., Wang, J., 2006, Vector Autoregressive Models for Multivariate Time Series In -Modeling Financial Time Series with S-PLUS®, New York: Springer, pp. 385-429.

Statistical matching for modeling of count data

Honeylet T. Santos
School of Statistics, University of the Philippines, Diliman, Quezon City
Valenzuela City, Philippines

Abstract

Statistical matching deals with methods of combining different data from different sources to get information on variables not observed in a single source. With the goal of estimating a Poisson regression model, this study explores statistical matching techniques and estimation procedures involving bootstrap. Simulation studies confirmed that Poisson regression imputation and MCMC imputation produce comparable results. It also showed that bootstrap within method performs well regardless of the matching method used.

Keywords

Imputation; Bootstrap; MCMC

1. Introduction

Recent technological development has created new sources and ways of harnessing data. Vast data sources are now available. Despite the proliferation of new and traditional data sources, there remains difficulties in utilizing them. A researcher may want to model a variable from Data Source A using a variable from Data Source B. This poses a problem because the information required are not contained in the same data source. There is thus a need to explore how these data sources can be combined through statistical matching. In D'Orazio et al. (2006), it is discussed that statistical matching deals with the problem of combining sources of information under the assumptions that (a) common variables are observed in different data sources and (b) observations from different data sources do not overlap.

Literature on statistical matching methods, for instance D'Orazio et al. (2006), often focuses on continuous and categorical variables. Hence, this study will focus on statistical matching of count data because count data can be found in many data sources.

Poisson is often assumed to be the distribution of count response variable in a generalized linear model (GLM). In Agresti (2013), Poisson loglinear models use the canonical link of a Poisson

GLM which is the log link. The Poisson loglinear model can be represented as $\log \mu = \alpha + \beta_1 x_1 + \cdots + \beta_p x.$ (1.1)

where μ is the mean of the response, α is the intercept, $\beta_.$ is the coefficient, and $x_.$ is the explanatory variable for $i = 1, \ldots, p$.

Aside from Poisson regression via Maximum Likelihood Estimation (MLE), bootstrap methods will also be considered in the estimation of model parameters. Efron (2000) argued that an advantage of bootstrapping is its broad application. Moreover, Efron and Tibshirani (1986) discussed that the bootstrap methods can be used in measuring statistical accuracy of estimators even with more complicated forms. They showed in a sampling experiment that bootstrap estimates for standard error of correlation coefficient, which does not have a simple form, are nearly unbiased.

It is possible that a true model involves a count response variable predicted by variables X_1 and X_2 that either have high correlation or low correlation. However, a predicament wherein only one of the $\boldsymbol{X}$ variables was observed in the data sources may occur in practice. Hence, simulations in this study will focus on cases in which only one of the common $\boldsymbol{X}$ variables is available in the data sources. Furthermore, other considerations that might take place in practice will be taken into account.

These include total sample size of concatenated data sources, ratio of data sources to total sample size, and effect of X_1 and X_2 on count variables to be matched.

The objective of this study is to combine data from different sources through statistical matching techniques with the end goal of developing a count regression model. Specifically, this study aims to: (1) develop a statistical matching technique to create synthetic count data, (2) estimate count regression models based on synthetic data, and (3) characterize the estimation procedure through simulation studies.

The matching and estimation procedure will be evaluated using absolute relative bias (RBIAS) and mean absolute error (MAE). RBIAS will measure the accuracy of the estimates obtained while MAE will measure the predictive ability of the estimated model.

2. Methodology

Statistical matching problem involves integrating different data sources to create a synthetic dataset. For the purpose of this study, two independent data sources – Data Source A with n_A observations and Data Source B with n_B observations – will be considered. Common variable X is observed in both data sources while specific variable Z is missing in Data Source A and specific variable Y is missing in Data Source B. The data sources are random samples from the same population. The assumption is that combining these two independent data sources will yield a larger random sample Data Source A ∪ B with $n = n_A + n_B$ observations from the same population. Consequently, the observation units in data Source A and data Source B are disjoint.

2.1. Matching Methods

Poisson regression imputation

1) Fit separate Poisson regression models to Data Source A and Data Source B with specific variables as response and common variable as predictor.
2) Impute missing values in Data Source A and Data Source B based on models in Step 1.
3) Concatenate Data Source A and Data Source B to form synthetic dataset with X, Y, and Z values.

Random Hot Deck imputation

1) Randomly choose an observation from Data Source B (donor file) for each observation in Data Source A (recipient file).
2) Impute missing values in Data Source A based on values of matched observations from Data Source B. Completed Data Source A will serve as synthetic dataset.

Markov chain Monte Carlo imputation

1) Fit a model to Data Source A with Y as response using a random walk Metropolis algorithm with posterior distribution of a Poisson regression model and improper uniform prior for the coefficient. The starting value of the coefficient used in the algorithm is its maximum likelihood estimate. The number of burn-in iterations is 1,000 and the number of Metropolis iterations is 10,000. Similarly, fit a model to Data Source B with Z as response.
2) Impute missing Z values in Data Source A using the model of Data Source B in Step 1. Impute missing Y values in Data Source B using the model of Data Source A in Step 1.
3) Concatenate Data Source A and Data Source B to form synthetic dataset with X, Y, and Z values.

In summary, Poisson regression imputation will be used to predict missing values in the data sources by fitting corresponding Poisson models, while random hot deck procedure will not require specification of a model. It will impute missing values by simply matching a random observation from donor file Data Source B to each observation in recipient file Data Source A. Alternatively, MCMC imputation will be used to get the model of the data sources with corresponding specific variables as responses and common variables as predictors. It involves simulation from a posterior distribution of Poisson regression models. Missing values of Y and Z will then be predicted based on these models.

2.2. Estimation of the Model

The goal of this study is to estimate a model of variable Y using variable Z after matching. The model is Poisson loglinear model characterized by the following:

$$\mu = \exp(\alpha + \beta z), i = 1,2, \dots, n \quad (2.1)$$

where μ_i is the mean of response y_i, α is the intercept, β is the coefficient, z_i. is the i^{th} observation of variable Z, and $n = n_A + n_B$.

After imputation, corresponding models (2.1) will be fitted to synthetic datasets to estimate the coefficients. Variable Y will be the response and variable Z will be the predictor. Moreover, coefficients will also be estimated using bootstrap methods described below.

Bootstrap within one synthetic dataset

1) Create synthetic dataset using matching methods described above.
2) Resample with replacement the same number of observations from synthetic dataset created in Step 1.
3) Fit a model (2.1) to the resampled dataset in Step 2. Get the coefficient estimates.
4) Repeat Step 2 and Step 3 200 times.
5) Get the average of the 200 coefficient estimates. The average will serve as the coefficient estimate of model (2.1).

Bootstrap across synthetic datasets

1) Concatenate the original data sources – Data Source A with missing Z values and Data Source B with missing Y values.
2) Resample with replacement the same number of observations from the concatenated dataset in Step 1.
3) From the resampled dataset in Step 2, group the observations with Z missing. This will serve as the new Data Source A. Similarly, group the observations with Y missing. This will serve as the new Data Source B.
4) Create synthetic dataset using matching methods.
5) Fit model (2.1) to the dataset in Step 4. Get the coefficient estimates.
6) Repeat Step 2 to Step 5 200 times.
7) Get the average of the 200 coefficient estimates. The average will serve as the coefficient estimate of model (2.1).

Note that synthetic datasets created using Poisson regression imputation and MCMC imputation are concatenated file Data Source A ∪ B, while synthetic datasets created using random hot deck imputation involve only Data Source A with imputed Z values from Data Source B. Hence, the corresponding synthetic datasets will be used in estimation of model (2.1).

A total of nine of model coefficients will be estimated. For each type of synthetic dataset, there are three coefficient estimates. These are: (1) Poisson

regression model estimates without bootstrapping, (2) Poisson regression model estimates using bootstrap within, and (3) Poisson regression model estimates using bootstrap across.

Absolute relative bias (RBIAS) and mean absolute error (MAE) will be used to evaluate the performance of matching and estimation procedures. Formula for RBIAS and MAE are shown below:

$$RBIAS\ of\ \hat{\beta} = \left|\frac{\beta - \widehat{\beta}}{\beta}\right| \quad (2.2)$$

where β is the true value of the coefficient and $\hat{\beta}$ is the coefficient estimate; and

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (2.3)$$

where y_i is the i^{th} observation, $\hat{y}_i$ is the predicted i^{th} observation, and n is the number of observations.

2.3. Simulation

To evaluate the performance of matching and estimation methods, datasets with complete information are generated first. The complete dataset consists of two common variables X_1 and X_2 with either high correlation or low correlation, and specific variables Y and Z. Highly correlated X_1 and X_2 are generated as follows:

$$X_1 \sim N(2,1); X_2 = 0.99x_1 + \delta\ where\ \delta \sim N(0,1) \quad (2.4)$$

Moreover, X_1 and X_2 with low correlation have the following characteristics:

$$X_1 \sim N(2,1); X_2 = 2 + 0.01x_1 + \delta\ where\ \delta \sim N(0,1) \quad (2.5)$$

Subsequently, specific variables Y and Z are Poisson distributed with log means that are either linear or nonlinear functions of X_1 and X_2. For the linear case, means of Y and Z are generated as follows:

$$\boldsymbol{\mu_Y} = \exp(\lambda_1 \boldsymbol{x_1} + \lambda_2 \boldsymbol{x_2}); \boldsymbol{\mu_Z} = \exp(\gamma_1 \boldsymbol{x_1} + \gamma_2 \boldsymbol{x_2}) \quad (2.6)$$

where $\boldsymbol{\mu_Y}$ – vector of means of Y; $\boldsymbol{\mu_Z}$ – vector of means of Z;
$\boldsymbol{x_j}$ – vector of observations of $X_j, j = 1,2$;
λ_j and γ_j – respective coefficients of $\boldsymbol{x_j}, j = 1,2$.

Furthermore, coefficients of the $\boldsymbol{X}$ variables were made to vary. Specifically, the effect of the $\boldsymbol{X}$ variables on Y can either be: (a) X_1 dominating with $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$ or (b) X_1 and X_2 have equal effects with $\lambda_1 = \lambda_2 = 0.45$. Similarly, the effect of the $\boldsymbol{X}$ variables on Z can either be: (a) X_1 dominating with $\gamma_1 = 0.7$ and $\gamma_2 = 0.1$ or (b) X_1 and X_2 have equal effects with $\gamma_1 = \gamma_2 = 0.4$.

For nonlinear case, means of Y and Z are generated as follows:

$$\boldsymbol{\mu_Y} = \exp[\exp(\lambda_1 \boldsymbol{x_1}) + \exp(\lambda_2 \boldsymbol{x_2})]; \boldsymbol{\mu_Z} = \exp[\exp(\gamma_1 \boldsymbol{x_1}) + \exp(\gamma_2 \boldsymbol{x_2})] \quad (2.7)$$

where $\boldsymbol{\mu_Y}$ – vector of means of Y; $\boldsymbol{\mu_Z}$ – vector of means of Z;
$\boldsymbol{x_j}$ – vector of observations of $X_j, j = 1,2$;
λ_j and γ_j – respective coefficients of $\boldsymbol{x_j}, j = 1,2$.

In this case, the coefficients of the $\boldsymbol{X}$ variables in generating the means of Y are: (a) $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$ or (b) $\lambda_1 = \lambda_2 = 0.14$. The coefficients in

generating the means of Z are: (a) γ_1 = 0.15 and γ_2 = 0.05 or (b) $\gamma_1 = \gamma_2$ = 0.12.

After generating a complete dataset which will be the benchmark for comparisons, the next steps will be done to the same complete dataset to simulate scenarios. For simulation summary, please refer to Table 1.

1) One of the $\boldsymbol{X}$ variables will be discarded. Either X_1 or X_2 will be used as common variable.
2) Random missing values will be assigned to variable Z. The percentage of missing values will be part of scenario cases. These are 10%, 30%, 50%, 70%, and 90%.
3) The dataset will be separated into Data Source A and Data Source B. Observation units with Z missing will comprise Data Source A, while the rest of the observation units will comprise Data Source B. Hence, if the percentage of missing values in variable Z is 10%, then Data Source A with with n_A number of observations will comprise 10% of the total sample size n while Data Source B with n_B observations will comprise 90% of the total sample size n.

Subsequently, matching procedures will then be applied to Data Source A with missing Z values and Data Source B with missing Y values. Then, estimation procedures will be used to estimate the coefficients of model (2.1). Each simulation scenario will have 100 replicates.

Table 1. Simulation Summary

Settings	**Scenarios**
Sample Size	200, 500, 1000
Correlation of X_1 and X_2	High Correlation, Low Correlation
Effect of X_1 and X_2 on Y	X_1 dominates, X_1 and X_2 equal effect
Effect of X_1 and X_2 on Z	X_1 dominates, X_1 and X_2 equal effect
Percentage of Source A: Percentage of Source B to total sample size	10:90, 30:70, 50:50, 70:30, 90:10
Common variable used	X_1 only or X_2 only
Log Mean of Y and Z	Linear function of $\boldsymbol{X}$, Nonlinear function of $\boldsymbol{X}$

3. Result

RBIAS measures the accuracy of the estimates obtained while MAE measures the predictive ability of the estimated model.

1. When Log Mean of Y and Z are linear functions of X_1 and X_2

Sample size

When the common variable used is X_1, Poisson regression imputation and MCMC imputation produce comparable RBIAS. As sample size increases,

RBIAS decreases. Also, the RBIAS under these two matching methods are lower compared to those of random hot deck imputation.

The three estimation procedures yield similar results under Poisson regression and MCMC imputation matching methods, although Poisson regression without bootstrap and bootstrap within methods yield the lowest RBIAS. Poisson regression without bootstrap produces the lowest RBIAS when the true coefficient of X_1 has dominating effect on Y in the data generating process.

Generally, bootstrap within method produces lowest RBIAS when true coefficients of X_1 and X_2 have equal effect on Y. Furthermore, under random hot deck imputation, bootstrap across method produces the largest RBIAS which are far from values produced by the other two estimation methods.

When the common variable used is X_2 and X_1 has dominating effect on both Y and Z, resulting RBIAS are larger. This is true for Poisson regression imputation and MCMC imputation which generally have similar results, although MCMC imputation has the lowest RBIAS in most cases. In general, bootstrap within and bootstrap across estimation methods under these two matching methods have lower RBIAS.

Furthermore, under Poisson regression imputation and MCMC imputation the differences of RBIAS values among the three estimation procedures are not large. However, under random hot deck imputation, bootstrap across method yields large RBIAS values far from values yielded by using Poisson regression without bootstrap and bootstrap within methods. Bootstrap within reduces RBIAS in scenario cases when X_1 and X_2 have equal effect on both Y and Z while X_1 and X_2 have low correlation.

Regardless of the common variable, the three estimation procedures yield similar results although bootstrap within and bootstrap across always have lower MAE values than those of Poisson regression without bootstrap. For small sample (size 200), bootstrap within produces lowest MAE values. This is true for all matching methods. Also, under random hot deck imputation, bootstrap within always yields lowest MAE values compared to those produced by using Poisson regression imputation and MCMC imputation. Moreover, comparable results were produced by using Poisson regression imputation and MCMC imputation methods, while larger MAE values were produced by using random hot deck imputation.

Ratio of Data Source A and Data Source B to total sample

When common variable used is X_1, RBIAS produced by the three estimation procedures are comparable when matching methods are Poisson regression imputation and MCMC imputation. Generally, when X_1 has dominating effect on Y, Poisson regression imputation yields the lowest RBIAS.

Also, lower RBIAS results from using Poisson regression imputation for scenario cases that satisfy all of the following conditions: (a) X_1, X_2 have equal

effect on Y, (b) X_1 has dominating effect on Z, and (c) X_1 and X_2 have high correlation. The only exception is when the ratio of Data Source A: Data Source B is 90:10. The rest of the scenario cases where X_1, X_2 have equal effect on Y have lower RBIAS when bootstrap within and bootstrap across were used.

Under random hot deck imputation, bootstrap across method produces RBIAS that are much larger than those of the other two estimation methods. For ratios 10:90 and 90:10, bootstrap within produces lower RBIAS for the following cases regardless of the common variable used: (1) X_1 dominating on Y and Z, high correlation of X_1 and X_2 and (2) X_1 dominating on Y, X_1 and X_2 have equal effect on Z, low and high correlation of X_1 and X_2.

RBIAS are largest when the matching variable used is X_2 when X_1 has dominating effect on both Y and Z. Bootstrap across does not perform well when using random hot deck imputation regardless of the common variable.

With regards MAE values, the three estimation methods produce similar results regardless of the common variable used. Additionally, MAE values produced under random hot deck imputation are higher than those of the other two matching methods.

In terms of predictive ability of estimated models, bootstrap within and bootstrap across perform better than Poisson regression without bootstrapping. This is true regardless of the matching methods used. Particularly, under random hot deck imputation, bootstrap within always has the lowest MAE values. Moreover, bootstrap within has lowest MAE values across matching methods for extreme ratios 10:90 and 90:10.

2. When Log Mean of Y and Z are nonlinear functions of X_1 and X_2

Only MAE will be discussed in this section because the model where the complete data were generated from is different from the model used to fit the data.

Regardless of the common variable and matching method used, Poisson regression, bootstrap within, and bootstrap across estimation methods produce comparable results. Although results of the estimation methods are comparable within a matching method, random hot deck imputation has larger MAE values compared to those of its Poisson regression imputation and MCMC imputation counterparts.

Furthermore, under Poisson regression imputation and MCMC imputation, bootstrap within and bootstrap across estimation procedures yield lower MAE values compared to those produced without bootstrapping. Subsequently under random hot deck imputation, bootstrap within always produce lowest MAE values.

4. Discussion and Conclusion

Simulation results presented above summarized the effects of various likely settings in statistical matching with the objective of estimating a count regression model. Bootstrap-based approaches for estimation of synthetic data created from various imputation methods are in general at par with or better than the benchmark regression approach. It might be of interest for future studies to include a donation class to improve the estimates of the model when synthetic datasets are created using random hot deck imputation.

Simulations show that MCMC imputation and Poisson regression imputation produce comparable results when it comes to accuracy of estimates. Particularly, MCMC imputation yields lowest RBIAS. In terms of predictive ability of estimated models, Poisson regression imputation and MCMC imputation still produce comparable results although Poisson regression imputation has most of the lowest MAE values. Under random hot deck imputation, RBIAS and MAE values are larger compared to those produced by the other two matching methods.

Moreover, under Poisson regression imputation and MCMC imputation, the three estimation procedures yield similar results. Accuracy of estimates and predictive ability are comparable. In terms of predictive ability, the three estimation procedures also yield similar results when using Poisson regression imputation and MCMC imputation although bootstrap within and bootstrap across produce lower MAE values compared to Poisson regression without bootstrapping.

The predictive ability of the models estimated using bootstrap within is good for all sample sizes and ratios of data sources (to total sample) in the scenario settings. Moreover, regardless of the matching method used, bootstrap within produces low RBIAS and MAE values. For instance, if random hot deck is really necessary as the matching method because of the nature of the data and no donation classes were specified, bootstrap within can yield lower RBIAS and MAE values compared to Poisson regression without bootstrapping and bootstrap across method. It is interesting for further study to incorporate donation classes in random hot deck imputation and evaluate the performance of bootstrap within and bootstrap across estimation methods.

References

1. Agresti, A. (2013). Categorical Data Analysis, 3rd edition. New Jersey, USA. John Wiley & Sons. 2. D'Orazio, M., Di Zio, M., Scanu, M. (2006). Statistical Matching Theory and Practice. West Sussex, England. John Wiley & Sons Ltd.
2. Efron, B. (2000). The Bootstrap and Modern Statistics. Journal of the American Statistical Association, 95(452): 1293-1296.
3. Efron, B., Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science, 1(1): 54-77

The social-demographic effects on earning and its differentials in public and private sector

Noor Ismawati Mohd Jaafar[1], Nor Hanizah Abu Hanit[2], Norisan Mohd Aspar[2]

[1]Social Wellbeing Research Centre, University of Malaya

[2] Department of Statistics Malaysia

Abstract

Despite noticeable demographic shift in Malaysia, national statistics shows that employees remain as the biggest group in the Malaysian working population. As more women joined the labour force with better educational attainment, the gender earning gap has been identified as one of demotivation factor for women to remain working. Focusing on the monthly earning of paid employees, this paper investigates the effect of socio-demographic factors on earning. Considering one of the aim of the employment policies enforced in the country at the public sector is to reduce gender discrimination, the paper compares the result for paid employees working in public and private sectors by gender. Analysis done with Malaysian Salaries and Wages Survey 2010 – 2017 shows that earning increases with age. Chinese earns significantly higher as compared with Bumiputera at both sectors and highly educated employees significantly earn more. Being unmarried is no longer a significant factor among female in the public sector for certain years.

Keywords

Paid employees; Principal occupation; Citizen; Socio-demographic factors.

1. Introduction

Between 2010 and 2017, the Malaysian population has increased from 28 588.6 thousand to 32 022.6 thousand persons (Department of Statistics Malaysia, 2017d). Throughout this observation period, majority of Malaysian population was within the age of 15 to 64 years with approximately 70 per cent of the population was within the working age group. This trend can be seen for each year since 2010 (Department of Statistics Malaysia, 2017a, 2017b, 2017c). Within the working population, employed persons comprised as the largest group not only in the working population but as well as in the overall population. During the same observation period, the female participation rate in the labour market had increased from 46.8 per cent in 2010 to 54.7 per cent in 2017. Nonetheless the rate in 1990 peaked at age 24 but decreased with age (Siti Rohani Yahya, 1993) and the rates of those with tertiary education were also lower as compared with their male counterpart (Roslilee AB Halim et al., 2016).

Monetary reward was identified as one of the top motivation factor (Wiley, 1997) to stay in employment, however, inter-gender gap factor in earning could be one of the strong drive in pulling the female employees from remaining in the labour force (Nor'Aznin Abu Bakar and Norehan Abdullah, 2017). In recognising the role of women in national development, a quota of at least 30 per cent of decision-making positions was allocated to women in the public sector (Economic Planning Unit, Prime Minister's Department, 2006). The private sector and NGOs were also encouraged to increase the participation of women in managerial and key positions. Initiatives were planned in increasing the number of talented women returning to the workforce (Economic Planning Unit, Prime Minister's Department, 2015). The initiatives to produce more women leaders in the public sector had shown great progress with women accounting for 35.6 per cent of top management positions in 2017 as compared to 32.5 per cent in 2015, beyond the 30 per cent target (Economic Planning Unit, Prime Minister's Department, 2018). Malaysia gender gap index (MGGI) for 2017 was 0.697 (Department of Statistics Malaysia, 2017e) comprising four selected domains i.e. Economic participation & opportunity (0.659), Educational attainment (1.092), Healths & survival (0.957) and Political empowerment (0.061).

As the earning in public and private sector differs significantly in Malaysia, the paper investigates the effect of socio-demographic factors on earning for both sector by gender. Even though the analysis was done for four sub samples, the result and discussion of this paper only covers female paid employee and limited to the social-demographic variables.

2. Methodology

The analysis is done based on Salaries and Wages Survey conducted in 2010 until 2017 administrated by the Department of Statistics Malaysia. The population of the survey is paid employees who live in private living quarters. The details on principal occupation and demographic background of all paid employees in the household were collected for the respective reference month. Paid employees is defined as individuals who are working full-time, working at least 6 hours a day or at least 20 days a month, contract workers in the government sector, individuals who receive monthly regular and periodic allowances or volunteers who receive fixed allowances (Department of Statistics Malaysia, 2016). The survey covers both urban and rural areas for all administrative districts within all states in the country. The sample was represented by 175,994 female employees aged between 15 and 64 in the respective years of observation. The whole sample comprised of 29.8 per cent of employees in the public sector and 70.2 per cent in the private sector. The sample excluded non-citizen.

The analysis was done with earning as the only continuous numerical variable with age, ethnicity, education attainment and marital status as the socio-demographic variables. The analysis also includes social-economic and geographical variables namely occupation, industry, state and area of residency as additional to the analysis.

Earning is represented by total monthly salaries & wages received which includes basic salaries or wages before deduction of income tax and Employee Provident Fund contribution, allowances, commissions, overtime and other payment in kind from primary occupation. Bonus is not included in this monthly measurement. For educational attainment, no formal education refers to those who had never attended school in any of the education institution. Primary refers to those who went to school between age 6 and 12 years, secondary for those who had been to education institution until age 17 and tertiary for those who obtained certificate higher than Sijil Pendidikan Malaysia (SPM). SPM certificate is awarded to those who passed the national examination (Department of Statistics Malaysia, 2016), normally taken in the final year of secondary level. The field of study is only applicable to those who are with tertiary education. Occupation is categorised into nine categories following Malaysia Standard Classification of Occupations 2008 (MASCO-08). Industry is coded in 20 categories based on the Malaysia Standard Industrial Classification (MSIC). The geographical characteristic is represented by the state and area of residency. Following the administrative boundaries of the country, state refers to respondent's usual place of stay during the survey. Out of 16 states, there are three federal territories. Area of residencies refers to the urban-rural classification of the survey respondent's living quarters (LQ).

The modified Mincer-Typed earning regression is utilised to determine the impact of socio-demographic variables in estimating earning for each year separately for both sector by gender. The regression equation applied is as follows,

$$\ln Y = a_0 + a_1 AGE + a_2 AGE^2 + a_3 ETHNIC_i + a_4 MARITALS_i + a_5 EDUCATION_i + \\ + a_6 STRATUM + a_7 STATE_i + a_8 OCCUPATION_i + a_9 INDUSTRY_i + e_i$$

where

$\ln Y$ = The logarithm of monthly earning;

a_i = Estimated coefficients;

AGE = Age on last birthday;

AGE^2 = Age square;

$ETHNIC_i$ = A series of dichotomous variables indicating the respondent is Bumiputera, Chinese, Indian, or other ethnic group;

$MARITALS_i$ = A series of dichotomous variables indicating marital status of the respondent;

$EDUCATION_i$ = A series of dichotomous variables indicating level of educational attainment;

$STRATUM$ = A dichotomous variable indicating whether the respondent lives in urban area;

$STATE_i$ = A series of dichotomous variables indicating state of residency;

$OCCUPATION_i$ = A series of dichotomous variables indicating occupation;

$INDUSTRY_i$ = A series of dichotomous variables indicating industry of employment;

e_i = The disturbance terms.

3. Results

In total, 32 regressions were generated for employees in public and private sector by gender. Due to limitation of the paper, Table 1 shows the regression output for female employees in both sectors for each separate year. In 2010, the model indicates that all 10 predictors explain 56 per cent of the variance in the sample of 6 218 female employees working in public sector. Age and Age Squared are highly significant (*p-value*<1%) determinants of earning. Chinese earns significantly (*p-value*<10%) as compared with Bumiputera. On the other hand, Indians earn lower and other ethnic earns higher but the difference is not significant. Data in 2010 of the female employees in public sector also indicates that those who are not married earn significantly lower (*p-value*<1%) as compared with the married employees. However the coefficient of never married is higher as compared with widowed/separated/divorced. The model also shows those having secondary or tertiary education significantly (*p-value*<1%) earn more as compared with those with no formal education or with primary education. The coefficient of secondary is lower than tertiary. This suggests that higher education attainment positively affects earning.

Comparison between years shows that age, age square and educational attainment remain significant determinants (*p-value*<1%) of earning among female employees in public sector. The coefficients of the Chinese's female employees in public sector are positive every year but the significant level varies throughout the year. The coefficients of Indians and other ethnic groups also vary throughout the year. It is observed that the Indians and other ethnic groups earning are significantly different as compared with Bumiputera in 2013. As shown in Table 1, being unmarried is significantly (*p-value*<1%) negatively effecting earning. Those who are never married remain to earn significantly (*p-value*<1%) lower as compared to those who are married until

2015. In 2016, the coefficient changes to become positive but is no longer significant. Widowed/Separated/Divorced earns lower as compared with those who are married but becomes no longer significant from 2013 until 2016.

The regressions output of the female employees in private sector indicates similar impact for age, age square and educational attainment (*p-value*<1%). In contrast with the female employee in public sector, Chinese employees in private sector earn significantly (*p-value*<1%) higher as compared with Bumiputera in all years of observation. Table 1 also shows that in majority of the years, the unmarried employees in this sector earn significantly (*p-value*<1%) less as compared employees who are married.

4. Discussion and Conclusion

The regressions show the positive relationship between earning and age. Chinese earns more as compared with Bumiputera and highly educated employees also earn more. This observation can be seen in both public and private sectors regardless of gender. The data also indicates that unmarried female employee earns significantly lower as compared with those who are married. However, being single is no longer significant determinants in earning of female employees in public sector since 2016 and being widowed/separated/divorced is not significantly effecting earning between 2013 and 2016. This provides numerical evidence of the successfulness of the gender equality policy in employment at Malaysian public sector. While marital status is no longer a significant determinant of earning in public sector, however it is not true for female employees in the private sector. Ethnic disparity in earning is very much visible in private sector. Further study should be done focusing on female employees by linking their earning with skills.

Table 1: The regression output of paid female employees in the public and private sector by year.

Variable	2010	2011	2012	2013	2014	2015	2016	2017
Public Sector								
Constant	5.665(51.5)***	5.599(52.8)***	5.816(46.8)***	6.337(45.2)***	5.994(46.1)***	6.36(47.6)***	6.082(39.4)***	5.93(57.8)***
Age	0.00(-7.6)***	-0.001(-10.2)***	0.062(14)***	0.051(11.8)***	0.062(14.7)***	0.059(14)***	0.059(13.6)***	0.068(20.4)***
Age^2	0.054(12.7)***	0.062(15.6)***	-0.001(-9.2)***	0.00(-7.5)***	-0.001(-9.7)***	0.00(-8.9)***	0.00(-8.6)***	-0.001(-14)***
Chinese	0.033(1.7)*	0.005(0.3)	0.04(1.9)*	0.104(3.2)***	0.029(1.5)	0.046(2.5)**	0.045(2.3)**	0.076(5.3)***
Indian	-0.034(-1.3)	-0.015(-0.6)	-0.005(-0.2)	0.063(2.1)**	0.033(1.1)	0.01(0.4)	-0.036(-1.2)	0.018(1)
Others	0.024(0.3)	0.039(0.3)	-0.035(-0.3)	-0.276(-2.7)***	-0.024(-0.2)	-0.055(-0.6)	0.034(0.4)	0.009(0.2)
Never married	-0.038(-3)***	-0.039(-3.4)***	-0.035(-2.7)***	-0.127(-8)***	-0.056(-4.2)***	-0.036(-2.9)***	0.003(0.3)	-0.011(-1.1)
Widowed/Separated/Divorced	-0.057(-2.7)***	-0.068(-3.3)***	-0.073(-3.3)***	-0.026(-0.7)	-0.017(-0.8)	-0.009(-0.4)	-0.029(-1.4)	-0.039(-2.6)***
Secondary	0.664(13.9)***	0.594(12.4)***	0.48(8.3)***	0.247(8.5)***	0.254(9.6)***	0.203(6.8)***	0.293(3.9)***	0.33(6.3)***
Tertiary	0.885(18)***	0.814(16.6)***	0.758(12.8)***	0.414(13.5)***	0.456(15)***	0.418(12.5)***	0.475(6.2)***	0.534(10.1)***
The R Square	0.56	0.56	0.53	0.53	0.59	0.56	0.54	0.51
Sample size	6218	6417	6367	6356	5733	5788	5314	10239
Private Sector								
Constant	5.653(109.9)***	5.741(117.2)***	5.863(122)***	6.108(121.3)***	6.112(128.9)***	6.238(132.6)***	6.166(125.5)***	6.771(173)***
Age	-0.001(-14.6)***	0.00(-13.9)***	0.042(16.5)***	0.034(13.1)***	0.038(15.8)***	0.034(14.4)***	0.035(14.4)***	0.029(15.9)***
Age^2	0.049(18)***	0.044(16.9)***	0.00(-13.7)***	0.00(-11)***	0.00(-13.6)***	0.00(-11.8)***	0.00(-11.5)***	0.00(-11.4)***
Chinese	0.191(18.5)***	0.205(21.4)***	0.182(18.8)***	0.211(20.8)***	0.197(20.7)***	0.192(20.5)***	0.206(21.1)***	0.128(18.3)***
Indian	-0.031(-2.2)**	-0.045(-3.3)***	0.018(1.3)	-0.004(-0.3)	-0.004(-0.3)	0.014(1.1)	-0.015(-1.1)	-0.024(-2.5)**
Others	0.013(0.3)	0.038(0.8)	0.095(2.1)**	-0.005(-0.1)	0.092(2)**	0.047(1.1)	0.047(1.1)	0.024(1)
Never married	-0.01(-1)	-0.035(-3.7)***	0.206(21.1)***	-0.039(-3.9)***	-0.032(-3.5)***	-0.032(-3.5)***	-0.028(-3)***	-0.019(-2.7)***
Widowed/Separated/Divorced	-0.12(-7.9)***	-0.063(-4.2)***	-0.049(-3.6)***	-0.042(-3)***	-0.014(-1)	-0.052(-3.9)***	-0.043(-3.2)***	-0.035(-3.5)***
Secondary	0.225(15.4)***	0.268(19.2)***	0.207(15.2)***	0.2(14.4)***	0.203(15.1)***	0.195(13.9)***	0.175(12.2)***	0.174(15.6)***
Tertiary	0.435(23)***	0.505(28.2)***	0.443(24.8)***	0.406(22.6)***	0.414(24)***	0.419(23.9)***	0.377(21.1)***	0.384(28.6)***
The R Square	0.61	0.61	0.62	0.6	0.61	0.6	0.6	0.51
Sample size	13308	14426	14227	13255	14564	14086	13185	24329

Source: Department of Statistics Malaysia, Salaries and Wages Survey 2010-2017.

Note: Coefficient (t-statistics)

Other variables used but not shown, comprise of area and state of residency, occupation, and industry.

Significant at *10%, **5% and ***1%.

References

1. Department of Statistics Malaysia. (2011). *Population Distribution and Basic Demographic Characteristics, 2010 Population and Housing Census*. Putrajaya: Department of Statistics Malaysia.
2. Department of Statistics Malaysia. (2016). *Salaries and Wages Report 2016*. Putrajaya: Department of Statistics Malaysia.
3. Department of Statistics Malaysia. (2017a). Current population estimates, Malaysia, 2016-2017 [Press release]. Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=a1d1UTFZazd5ajJiRWFHNDduOXFFQT09
4. Department of Statistics Malaysia. (2017b). *Labour Force Survey (LFS) Time Series Statistics by State, 1982-2016*. [Online Resource]. Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/ctimeseries&menu_id=NHJlaGc2Rlg4ZXlGTjh1SU1kaWY5UT09
5. Department of Statistics Malaysia. (2017c). Employed persons by status in employment, Malaysia/states, 1982 – 2016. [Online Resource]. Retrieved from https://www.dosm.gov.my/v1/uploads/files/3_Time%20Series/Labour_Force_Survey_Time_Series_Statistics_by_State_1982-2016/TABLE%2013.pdf
6. Department of Statistics Malaysia. (2017d). Current Population Estimates, Malaysia 2018. [Online Resource]. Retrieved from https://newss.statistics.gov.my/newss-portalx/ep/epFreeDownloadContentSearch.seam?cid=141613
7. Department of Statistics Malaysia. (2017e). Statistics on Women Empowerment in Selected Domains, Malaysia, 2018. [Online Resource]. Retrieved from file:///C:/Users/User/Downloads/Statistics%20On%20Women%20Empowerment%20In%20Selected%20Domains,%20Malaysia,%202018.pdf
8. Department of Statistics Malaysia. (2017f). *Salaries and Wages Report 2017*. Putrajaya: Department of Statistics Malaysia.
9. Economic Planning Unit, Prime Minister's Department. (2006). Ninth Malaysia Plan, 2006-2010. [Online Resource]. Retrieved from http://www.pmo.gov.my/dokumenattached/RMK/RM9_E.pdf
10. Economic Planning Unit, Prime Minister's Department. (2010). Tenth Malaysia Plan, 2011-2015. [Online Resource]. Retrieved from http://www.pmo.gov.my/dokumenattached/RMK/RMK10_E.pdf
11. Economic Planning Unit, Prime Minister's Department. (2015). Eleventh Malaysia Plan, 2016-2020. [Online Resource]. Retrieved from https://www.pmo.gov.my/dokumenattached/speech/files/RMK11_Speech.pdf

12. Economic Planning Unit, Prime Minister's Department. (2018).Mid-term Review of the 11th Malaysia Plan, 2016-2020, New Priorities and Emphases. [Online Resource]. Retrieved from https://www.talentcorp.com.my/clients/TalentCorp_2016_7A6571AE-D9D0-4175-B35D-99EC514F2D24/contentms/img/publication/Mid-Term%20Review%20of%2011th%20Malaysia%20Plan.pdf
13. Nor'Aznin Abu Bakar and Norehan Abdullah. (2017). Labour Force Participation of Women in Malaysia. Paper presented at International Economic Conference on Trade & Industry.
14. Roslilee AB Halim, Nurul Nadia ABD Aziz, Mawarti Ashik Samsudin. (2016). Malaysian female graduates: Marriage, motherhood and labour force participation. International Journal of Multidisciplinary Research and Development. 3(1) p. 109-114
15. Siti Rohani Yahya. (1993). Patterns and Trends of Labor Force Participation of Women. Paper submitted to HAWA under the Women Development Project, Kuala Lumpur
16. Wiley, C. (1997). What motivates employees according to over 40 years of motivation surveys. International Journal of Manpower, Vol. 18 Issue: 3, pp.263-280.
17. Dr Amjad Rabi (2017). UNICEF Malaysia Country Office, United Nations Children's Fund (UNICEF), Malaysia (2017). Malaysia 2050: Economically Productive and Socially Inclusive, UNICEF Malaysia Working Paper Series WP/2017/001.
18. Statistics Finland (2014). *Women and Men in Finland in 2014*

Developing and validating risk prediction model for re-offending of individuals with a severe mental illness (Psychosis)

Olayan Albalawi[1,2]; Handan Wand[1]; Tony Butler[1]
[1] The Kirby Institute, University of New South Wales, Sydney, Australia.
[2] Department of Statistics, Science Faculty, Tabuk University, Saudi Arabia.

Abstract

Objective- To develop and validate simplified risk score models for predicting the risk of re-offending among those individuals who were diagnosed with a severe mental illness (psychosis) prior to the first offence between 2001 and 2012 in New South Wales.

Methods- A cohort of 7,743 individuals were diagnosed with a severe mental illness (Psychosis) prior to the first offence in New South Wales from 2001 to 2012. Individuals were randomly assigned to either a development (67%) or an internal validation dataset (33%). The primary outcome was a reoffending status. Cox regression models were used to create a risk prediction algorithm from the development dataset. It was internally validated using standard statistical measures.

Results- In the risk prediction model, six factors were identified as significant of re-offending: age at the first offence, Indigenous status, type of a severe mental illness, contact with mental health service after the first offence, the outcome of the first offence and the type of the first offence. A score of ≥ 10 was selected as the optimum cut-point with 72% (43%) and 89% (19%) sensitivity (specificity) for development and validation datasets, respectively.

Conclusion - A new risk score was predictive of re-offending for those diagnosed with a severe mental illness and could help in local care and clinical research setting.

Keywords

Risk Prediction; offending; Severe mental illness

1. Introduction

Risk prediction models are frequently used in clinical and public health settings in order to identify those who are at risk of a disease of interest.[1-4] Besides unstructured clinical assessments by mental health experts, more than 100 structured tools have been developed and routinely used in clinical and justice system settings to predict the probability of future offending.[5] The majority of these tools have been primarily designed to predict the likelihood of future criminal behaviour based on evaluations of large numbers of cognitive and antisocial behaviours.[6-8] For example, the Comprehensive Health Assessment Tool (CHAT)[9] was developed as a standardised,

assessment tool and is frequently used in the Youth Justice System in the United Kingdom. It is comprised of three broad assessments: (1) physical and mental health, (2) substance abuse, and (3) neurodisability. These assessments were designed to be completed within 12, 21 and 30 working days, respectively. A systematic review[10] evaluated the predictive ability of more than 70 risk tools to assess the risk of criminal behaviour among approximately 25,000 people;[10] however, concerns were raised regarding the predictive accuracy of these tools as well as their practical use in real life. Due to their multicomponency and complex nature, these tools require multidisciplinary evaluations and may take several hours to be administered; therefore, they cannot be used universally (i.e. for everyone at clinical or criminal justice settings) to identify those at high risk of offending. In addition, these tools have been developed for certain populations such as psychiatric patients and those with poor cognitive and antisocial behaviours; therefore, their generalisability was also reported to be questionable.[10]

The primary objective of this study is to develop and validate risk prediction models to quantify an individuals' risks of reoffending using a data-linkage study for individuals who were diagnosed with a severe mental illness (psychoses) prior to the first offence in NSW, Australia.

2. Methodology

The study design and population has been described in detail elsewhere.[11] Data linkage was used to identify individuals who were diagnosed of psychosis from the NSW Ministry of Heath's Admitted Patient Data Collection (APDC) and the Emergency Department Data Collection (EDDC) based on International Classification of Diseases (ICD) nine and ten, and also who were committed the first offence Research's Re-offending Database (ROD) between 2001 and 2012.

7,743 men (%) and women (%) from a retrospective linkage study were included in this study. Individuals were randomly assigned to either a development (67%) or an internal validation dataset (33%). The outcome of interest in this study was an incident of reoffending. All individuals were followed from their first offence date until their reoffending, death or the 31st of December 2015, whichever occurred first. We developed the risk prediction model in two stages: first, we used a split-sample method in order to develop a risk equation using a weighted-scoring system. The study population was randomly allocated to either the development (67%) or internal validation (33%) sample dataset. We used a range of variables as potential predictors of the reoffending. These included, gender, country of birth, marital status, Indigenous status, Socio-Economic Indexes for Areas (SEIFA), psychosis type, age at the first offence, the outcome of the first offence, the type of the first offence and the status of contact with mental health services. Cox regression

models were used to create a prediction model for incidence of offending in the development and validation dataset. We first analysed the univariate associations between the independent variables and criminal convictions. Backward elimination was then used to reach the final multivariate model, in which factors with the largest p value were sequentially deleted until only significant predictors remained ($p < 0.05$). We then created a weighted scoring system by rounding all regression coefficients up to the nearest integer (i.e., the smallest integer greater than the estimate). This method was based on the β-coefficients [or log of the Hazard Ratios (HRs)] rather than HRs rather than HRs and is considered to be a more robust estimate.[2, 12, 13] After we identified the final gender specific models, we created integer weights for each variable by multiplying the β-coefficients by 10. These integer-weights were added to create the final scores for each individual. The discriminate power of the variables was assessed using the standard statistical techniques such as area under the receiving operating curve (AUC); while the diagnostic characteristics of various cut-points were evaluated using sensitivity and specificity in both datasets (i.e. development and derivation). The primary objective of this analysis was to investigate the accuracy and discriminative power of the models for the combination of the risk factors that we have included in the model. In an additional analysis, subject-specific scores were split into quintiles [1st to 5th]. Crude incidence rates (95% CIs) were calculated across the quintiles of the risk scores calculated separately for the development and validation datasets. HRs were also presented across the increasing quintiles of the scores.

3. Results

A total of 7,743 individuals were included in the study. Table 1 below summarises the characteristics of the study population. The development and validation datasets randomly selected included 5162 (67%) and 2,581 (33%), respectively. There was comparable regarding to group, age group at the first offence, gender, indigenous status, married status, country of birth, psychosis type, the status of contact with mental health service after the first offence, SEFIA and the type of the first offence.

Table 1: characteristics of the study population

Characteristics	Development N=5,162 (67)	Validation N=2,581 (33)	Total N=7,743 (%)
Group			
Treatment order	1,333 (26)	663 (26)	1,996 (26)
Punitive sanctions	3,829 (74)	1,918 (74)	5,747 (74)
Age at the first offence			
<20 years	202 (4)	101 (4)	303 (4)
20-29 years	1,583 (31)	792 (31)	2,375 (31)
30-39 years	1,722 (33)	851 (33)	2,573 (33)
40+ years	1,655 (32)	837 (32)	2,492 (32)
Gender			
Male	3,763 (73)	1,880 (73)	5,643 (73)
Female	1,399 (27)	701 (27)	2,100 (27)
Indigenous status			
No	4,624 (90)	2,346 (91)	6,970 (90)
Yes	538 (10)	235 (9)	773 (10)
Marital status			
Married	640 (12)	319 (12)	959 (12)
Other	4,522 (88)	2,262 (88)	6,784 (88)
Country of birth			
Australia	3,551 (69)	1,792 (31)	5,343 (69)
Other or unknown	1,611 (31)	789 (31)	2,400 (31)
Psychosis type			
Affective psychosis	513 (10)	283 (11)	796 (10)
Schizophrenia and related psychoses	3,575 (69)	1,776 (80)	5,351 (69)
Substance related psychosis	1,074 (21)	522 (20)	1,596 (21)
Contact with mental health services			
No	904 (18)	413 (16)	1,317 (17)
Yes	4,258 (82)	2,168 (84)	6,426 (83)
SEFIA (advantaged area)			
No or unknown	3,199 (62)	1,587 (61)	4,786 (62)
Yes	1,963 (38)	994 (39)	2,957 (38)
First offence type			
Violent	2,206 (43)	1,082 (42)	3,288 (42)
Non-violent	2,956 (57)	1,499 (58)	4,455 (58)

Cox regression were used to identify significant predictors of the reoffending in the development data sets. Our final Cox regression model identified 6-factors as being significant predictors of the reoffending; Indigenous status, psychosis type, age at the first offence, the outcome from the first offence, the type of the first offence and the status of contact with mental health service after the first offence. Results were remarkably similar when analyses were repeated using the validation datasets. (i.e. 33% of the study population). All the 6- risk factors that were identified as significant

predictor of the reoffending in the development datasets were also determined to be significant when the data were restricted to the validation datasets. Subject-specific risk scores were calculated by adding the estimated scores associated with each risk factor identified in the final prediction models (i.e. in Table 2).

Table 2: Developing the risk scoring algorithm for reoffending (overall):

Characteristics	Development Dataset (N= 5,162, 67%)					Validation Dataset (N=2,581, 33%)				
	%	Adjusted HR	p-value	βX10	Score	%	Adjusted HR	p-value	βX10	Score
Group										
Diverted	26%	**1**		**0.0**	**0**	26%	**1**		**0.0**	**0**
Punitive	74%	1.23 (1.11 – 1.36)	<0.001	2.1	**2**	74%	1.37 (1.18 – 1.59)	<0.001	3.1	**3**
Age at the first offence										
<20 years	4%	1.70 (1.38 – 2.10)	<0.001	5.3	**5**	4%	1.62 (1.23 – 2.16)	0.001	4.8	**5**
20-29 years	31%	1.52 (1.37 – 1.69)	<0.001	4.2	**4**	31%	1.60 (1.38 – 1.86)	<0.001	4.7	**5**
30-39 years	33%	1.42 (1.29 – 1.59)	<0.001	3.5	**4**	33%	1.31(1.13 – 1.52)	<0.001	2.7	**3**
40+ years	32%	**1**		**0.0**	**0**	32%	**1**		**0.0**	**0**
Indigenous status										
No	90%	**1**		**0.0**	**0**	91%	**1**		**0.0**	**0**
Yes	10%	2.09 (1.87 – 2.34)	<0.001	7.4	**7**	9%	2.00 (1.70 – 2.36)	<0.001	6.9	**7**
Psychosis type										
Affective psychosis	10%	**1**		**0.0**	**0**	11%	**1**		**0.0**	**0**
Schizophrenia and related psychoses	69%	1.16 (1.01 – 1.36)	0.041	1.5	**2**	69%	1.24 (1.01 – 1.52)	0.040	2.2	**2**
Substance related psychosis	21%	1.19 (1.01 – 1.40)	0.043	1.7	**2**	20%	1.29 (1.03 – 1.62)	0.028	2.5	**3**
Contact with mental health services										
No	17%	2.50 (2.27 – 2.74)	<0.001	9.1	**9**	16%	2.40 (2.09 – 2.75)	<0.001	8.8	**9**
Yes	83%	**1**		**0.0**	**0**	84%	**1**		**0.0**	**0**
First offence type										
Violent	43%	**1**		**0.0**	**0**	42%	**1**		**0.0**	**0**
Non-violent	57%	1.21 (1.11 – 1.31)	<0.001	1.91	**2**	58%	1.22 (1.09 – 1.38)	0.001	2.0	**2**

Table 3 below shows an increasing linear trend observed between the participants' risk scores (in quintiles) and the corresponding HRs when the lowest scoring category (i.e. 1st quintile) was the reference. Overall the reoffending rates ranged from 5.0 per 100 PY (95% CI: 4.5, 5.6) (in 1st quintile) to 23.6 per 100 PY (95% CI: 21.9, 25.4) (in 5th quintile) (for the development dataset) and 5.7 per 100 PY (95% CI:4.9, 6.7) (in 1st quintile) to 25.0 per 100 PY (95% CI:22.5, 27.8) (in 5th quintile) (for the validation dataset).

Table 3:Incidence rates (95% CI) at quintiles of the risk score

	Overall (overall incidence rate: 9.6 per 100 PY (9.3,9.9)					
	Development data: Incidence rate (95% CI)			**Validation data:** Incidence rate (95% CI)		
Overall	**9.6 (9.2,10.0) per 100 PY‡**			**9.5 (9.0,10.1) per 100 PY‡**		
Score[1]	Incidence rate (95% CI)	Adjusted HR (95% CI)	p-value	Incidence rate (95% CI)	Adjusted HR (95% CI)	p-value
1st quintile	5.0 (4.5,5.6)	1		5.7 (4.9,6.7)	1	
2nd quintile	6.9 (6.3,7.7)	1.38 (1.19,1.61)	<0.001	6.2 (5.4,7.0)	1.13 (0.92,1.38)	0.238
3rd quintile	9.1 (8.5,9.7)	1.81 (1.59,2.06)	<0.001	9.3 (8.1,10.6)	1.63 (1.33,1.99)	<0.001
4th quintile	14.8 (12.8,17.0)	2.65 (2.22,3.17)	<0.001	10.7 (9.4,12.1)	1.87 (1.53,2.27)	<0.001
5th quintile	23.6 (21.9,25.4)	4.26 (3.73, 4.87)	<0.001	25.0 (22.5,27.8)	3.96 (3.29,4.76)	<0.001

We have performed additional analyses in order to measure the performance of the risk prediction models using different cut-points of the total risk score in the development and validation datasets (Table 4). The purpose of these analyses was to determine the optimum cut-point to classify those at highest risk of the reoffending with statistically acceptable robustness. Discriminative powers of the risk scores was only 0.57% and 54% for development and validation data sets, respectively. However,

Table 4:Performance of the risk scoring algorithm for different cut points:

	Overall			
	Development Data W‡=57% (95%CI: 56%, 58%)		**Validation Data** AUC‡=54 % (95%CI: 53%, 55%)	
	Median (IQR)	**Mean (SD)**	**Median (IQR)**	**Mean (SD)**
Risk score	9 (7-12)	**10 (5)**	10 (7-13)	11 (5**)**
Cut points†	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
≥4	98%	5%	99%	3%
≥8	84%	30%	94%	11%
≥9	81%	34%	91%	16%
≥10	72%	43%	89%	19%
≥11	69%	47%	87%	21%
≥12	58%	56%	83%	25%
≥14	56%	58%	78%	29%

≥16	52%	60%	76%	30%
≥20	41%	64%	72%	31%
≥24	35%	66%	67%	33%

‡AUC: Area Under the Curve for score as continuous variable; †cut-points identified based on the distribution of the score

High cut-points resulted in high sensitivity with poor specificity across all models. For example, the sensitivity of the scoring tool was estimated to be 98% for a cut-point ≥4 (5% specificity), while a cut-point ≥10 yielded 72% sensitivity and 43$ specificity in the development set; sensitivity and specificity were estimated as 89% and 19% when the analyses were repeated using validation dataset. Based on these diagnostic measures, the cut-point ≥ 10 was selected to classify individual as being at high-risk for the reoffending.

4. Discussion and Conclusion

The result from our study show that a combination of a set of risk factors can be used to develop risk prediction models which can quantify an individual's risk of the reoffending with acceptable statistical accuracy. We developed and validated a 6-item risk prediction model, using data from 7,743 individuals from a data-linkage study in New South Wales. Most of the risk factors included in this study have been well recognised and consistently reported to have a strong association with re-offending, with broader implications. In the current study, having not contact with mental health service after the first offence had the highest impact on the reoffending with the highest risk scores. Those diagnosed with Schizophrenia and related psychoses or Substance related psychosis, their age was less than 20 years at the time of the first offence, not diverted outcome from the first offence were also at risk of the reoffending. However, to improve the Discriminative powers, we require to include more covariates in the data set such as education status, number of previous conviction and their types.

In conclusion, our risk predication models can potentially identify individuals at high risk of the reoffending by targeting very specific profiles and conditions. This approach may have significant implications for justice health system and health treatment planning within the clinical setting. Risk prediction tools such as the one developed here could be seen have several possible applications in local health care setting, justice system setting and clinical research setting.

References

1. Diamond, P.M., et al., *Screening for Traumatic Brain Injury in an Offender Sample: A First Look at the Reliability and Validity of the Traumatic Brain Injury Questionnaire.* Journal of Head Trauma Rehabilitation, 2007. **22**(6): p. 330-338.
2. Schmidt, M.I., et al., *Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study.* Diabetes Care, 2005. **28**(8): p. 2013-2018.
3. Wand, H., et al., *Developing and validating a scoring tool for identifying people who inject drugs at increased risk of hepatitis C virus infection.* Epidemiology Research, 2011. **2**(1).
4. Toll, D., et al., *Validation, updating and impact of clinical prediction rules: a review.* Journal of clinical epidemiology, 2008. **61**(11): p. 1085-1094.
5. Higgins, N., et al., *Assessing violence risk in general adult psychiatry.* The Psychiatric Bulletin, 2005. **29**: p. 131-133.
6. Archer, R.P., et al., *A survey of psychological test use patterns among forensic psychologists.* Journal of Personality Assessment, 2006. **87**: p. 84-94.
7. Khiroya, R., T. Weaver, and T. Maden, *Use and perceived utility of violence risk assessments in English medium secure forensic units.* The Psychiatric Bulletin, 2009. **33**(129-132).
8. Singh, J.P. and S. Fazel, *Forensic risk assessment: A metareview.* Criminal Justice and Behavior, 2010. **37**(9): p. 965-988.
9. Chitsabesan, P., et al., *The development of the comprehensive health assessment tool for young offenders within the secure estate.* The Journal of Forensic Psychiatry & Psychology, 2014. **25**(1).
10. Fazel, S., et al., *Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis.* the bmj 2012. **345**(e4692).
11. Albalawi, O., et al., *Court diversion for those with psychosis and its impact on re-offending rates: results from a longitudinal data-linkage study.* BJPsych Open, 2019. **5**(1).
12. Gonen, M., *Analyzing Receiver Operating Characteristic Curves with SAS*. 2007, Cary, NC: SAS Institute Inc.
13. Sackett, D., et al., *Evidence-Based Medicine: How to Practice and Teach EBM. 2nd Edition*. 2000, Churchill Livingstone: London.

Covariate-Adjusted response-adaptive designs for semi-parametric survival responses

Ayon Mukherjee
Lead Statistician, Novartis Pharmaceuticals, Hyderabad, India

Abstract

Covariate-adjusted response-adaptive (CARA) designs use the available responses to skew the treatment allocation in a clinical trial in favour of the treatment found at an interim stage to be best for a given patient's covariate profile. There has recently been extensive research on various aspects of CARA designs with the patient responses assumed to follow a parametric model. However, the range of application for such designs become limited in real-life clinical trials where the responses infrequently fit a certain parametric form. On the other hand, the parametric assumption yields robust estimates for the covariate-adjusted treatment effects. To balance these two requirements, designs are proposed without any distributional assumptions about the survival responses, relying only on the assumption of proportional hazards for the two treatment arms. To meet the multiple experimental obectives of a clinical trial, the proposed designs are developed based on optimal allocation approach. The optimal designs are based on biased coin procedures, with a bias towards the better treatment arm. These are the doubly-adaptive biased coin design (DBCD) and the efficient randomised adaptive design (ERADE). The derived treatment allocation proportions for these designs converge to the expected target values, which are functions of the Cox regression coefficients that are estimated sequentially with the arrival of every new patient into the trial. Based on simulation studies, it is found that the ERADE is preferable to the DBCD when the main aim is to minimise the variance of the observed allocation proportion and to maximise the power of the Wald test for a treatment difference. However, the former procedure being discrete tends to be slower in converging towards the expected target allocation proportion. Other comparative merits of the proposed designs have been highlighted and their preferred areas of application are discussed. It has been found that the proposed designs are a suitable alternative to traditional balanced randomisation designs in terms of their power, provided that response data are available during the recruitment phase to enable adaptations to the designs.

Keywords

Censored response; Optimal allocation; Power; Variability; Unbalanced randomization.

1. Introduction

There is great interest in the possibility that clinical trials can be designed with adaptive features that may make the studies more efficient, more likely to demonstrate an effect of the drug if one exists, or more informative. An adaptive clinical trial is a trial that evaluates a treatment by observing patient responses on a prescribed schedule, and modifies parameters of the trial protocol in accordance with those observations. The adaptation process generally continues throughout the trial. Clinical trials are often designed with adaptive features to force balance in sequential allocation of patients between two or more competing treatments. In order to address the ethical criteria of a clinical trial, it is also used to force imbalance by allocating a greater number of study subjects to the better-performing treatment.

A clinical trial is a complex experiment on humans with multiple and often competing experimental objectives. Here, several treatments for a disease are compared with the purpose of obtaining information on the performance of the treatments. Since it involves human patients, there is an ethical concern to treat as many patients as possible with the best treatment. At the same time, there must be some allocation of patients to the worse treatment arm for making useful statistical inferences about treatment comparisons.

An increase in the number of patients receiving better treatments leads to sequential experiments in which data are analysed and new allocations are made in the light of the estimated parameters. However, the advocates of traditional balanced randomization designs argue with the fact that having a balanced allocation of patients across the treatment arms helps estimate treatment effects efficiently. Since clinical trials involve human patients, balanced allocation can be a serious problem as one would be more inclined to be treated with the better treatment and such balanced allocation allocates almost half of the patients to the worse treatment arm. To balance these competing goals of ethics and statistical efficiency in a clinical trial, response-adaptive designs have been developed and used. A response-adaptive design uses the available response and treatment allocation histories to skew the treatment allocation probabilties in favour of the treatment arm found best thus far in the trial. However, human patients are heterogeneous and therefore one needs to take into account such concomitant information when allocating a particular patient to a treatment arm.

Covariate-adjusted response-adaptive designs balance the competing goals of assigning a greater number of study subjects to the better treatment and achieving high statistical efficiency in estimating treatment effects in the presence of covariates, while maintaining randomness in treatment assignments. Investigators are often aware of important baseline covariates that may have a strong influence on patient responses and they may wish to adjust the randomization procedure for these covariates. Rosenberger and

Sverdlov (2008) gave an overview of different techniques for handling covariates in the design of clinical trials and distinguished between two main approaches. These are covariate-adaptive randomization and covariate-adjusted response-adaptive (CARA) randomization procedures. CARA randomization is applicable to clinical trials where nonlinear and heteroscedastic models determine the relationship among responses, treatments and covariates, and when multiple experimental objectives are pursued in the trial. The goals of a CARA procedure may be to skew allocation in the direction of the treatment that is clinically best given a patient's covariate profile, while maintaining the power of a statistical test for treatment differences. These designs rely on correctly specified parametric models. Although the exponential model for survival responses has been previously considered by Sverdlov, Rosenberger and Ryzenik (2013) for developing CARA designs, to extend the scope of application of such designs in real-life clinical trials, the survival models that have been considered here relaxes on any distributional assumption of the patient responses but relies only on a lighter assumption of proportional hazards of patients between the two treatment arms.

2. Background

The exponential model used by Sverdlov, Rosenberger and Ryeznik (2013) to develop CARA randomization procedures for survival trials is useful for its historical significance and mathematical simplicity, but it is less likely to fit data in practice. This is because the exponential distribution corresponds to a "lack of memory" model and that it has only one parameter making the expected remaining lifetime for a patient at any point in the trial, to be a constant. The methods discussed in this paper extends the applicability of CARA designs even further by encompassing situations where the designs are applicable for survival responses irrespective of its theoretical distribution, provided the hazards of the event considered at any given time point is proportional and time-independent for any two patients in the trial.

In a medical context, the hazard rate is also known as the force of mortality and it represents a continuous version of a death rate per unit time. It is always convenient in survival analysis to describe the distribution of the survival responses in various different but inter-related ways. For a continuously distributed survival time T let f(t) be the density function and F(t) be the distribution function. The hazard function $h(t) = [f(t)/S(t)]$ can be interpreted as the instantaneous failure rate, where S(t) is the survivor function and gives the probability for a patient to survive beyond a given time point t. The survivor function can also be written as;

$$S(t) = e^{\left(-\int_0^t h(u)du\right)} = e^{[-H(t)]} \quad (1)$$

where H(t) is known as the integrated hazard or the cumulative hazard. It can therefore be seen that for the distribution of T to be proper i.e, for its density to integrate to one, $H(t) \rightarrow \infty$. If this is not true, the implication is that the individual may never die; though in some contexts this may not be an unreasonable approximating assumption (as, for example, where children may be cured of a childhood tumour and live "indefinitely" in relation to the time scale of the study). Normally, statisticians would want to insist on the distribution of T to be proper.

The hazard function gives the event rate at a given time t, conditional on having survived to time t. If the hazard rate is increasing, the risk of death (or failure) is also increasing with time because the ratio of the hazard rates will be the same as the ratio of the risk functions.

The Lehmann family, also known as the Proportional Hazard family is an important family of distributions in modelling surival times. If ξ is an arbitrary constant, the form of the Lehmann Family can be generated by:

$$S_k(t;\xi) = [S_k(t)]^{\xi}; f_k(t;\xi) = \xi[S_k(t)]^{\xi-1} f_k(t); h_k(t;\xi) = \xi h_k(\xi t) \quad (2)$$

This model can clearly be used to model the log hazard and is the basis of the important Proportional Hazard models where the covariates act additively on the logarithm of the hazard function. The exponential distribution and the Weibull distribution belongs to the Lehmann family.

Sir David Cox in the year 1972 had effectively used this concept to provide a semi-parametric approach for modelling time to event data where the survival experience of patients in different groups can be compared after adjusting for the effects of other variables which has a significant effect on the patients' survival responses. His approach had been extremely popular and the paper in 1972 about the Cox proportional Hazard model has become the most cited paper in the statistical literature. Unlike the Accelerated Life models which assume a particular parametric distribution for the survival time of the patients, the Cox proportional hazard model does not make any strong assumption about the functional form of the survival times but make a lighter assumption about the hazard ratio between two individuals at a particular time point being constant. Since the model makes no assumption about the functional form of the survival times, the parameter estimates are not based on the the probability of the observed outcomes given the parameter values. Instead of attempting to construct a full likelihood, Sir David Cox considered the conditional probability that, given that exactly one individual in the risk set R_i , with covariate vector z_j , dies at time t_i , it is the j^{th} individual that does so. Let x_j denote the treatment indicator for the j^{th} patient such that $x_j = 1$, if the patient is assigned to treatment A, and $x_j = 0$, if the patient is assigned to treatment B. Associated with patient $j = 1, \ldots, n$ is a $(p + 1)$ vector of baseline covariates $z_j = (1, z_{1i}, \ldots\ldots, z_{pi})^T$ and a risk set R_i which is

defined as the set of individuals still at risk at time t_i , where t_i is the i^{th} ordered event time. Let the hazard for the j^{th} individual, with covariate vector z_j , be $\mathrm{h}(t|z_j) = \mathrm{h}(\mathrm{t}|z=0)e^{\beta^T z_j}$ where $\mathrm{h}(\mathrm{t}|z=0)$ denotes the baseline hazard function.This calculates the hazard rate when all the covariate values for a patient is set to zero. Throughout this paper it is assumed that the survival responses follow a continuous time model, so that only one event occurs at any one time. Therefore, the conditional probability is given by;

$$P[individual\ j\ dies\ in[t_i,(t_i+\delta t)]|one\ death\ in\ t_i] = \frac{e^{\beta^T z_j}}{\sum_{k\in R_i} e^{\beta^T z_k}} \quad (3)$$

Thus the baseline hazard cancels out from the expression. This is the essence of the analyis : to evaluate the conditional probability the hazard at the event times ti only needs to be considered. The product of these conditional probabilities over all the ordered event times t_i is termed the partial *likelihood*, where j(i) is the index of the individual who dies at the i^{th} time :

$$PartialLikelihood(PL) = \prod_{i=1}^{n} \frac{e^{\beta^T z_{j(i)}}}{\sum_{k\in R_i} e^{\beta^T z_k}} \quad (4)$$

It can be seen from equation (4) above that the individual times ti do not appear in the expression of partial likelihood. This can be justified by the argument that in the absence of a parametric form for the hazard, there is no information about its value between successive t_i : it could quite possibly be zero. It follows that the partial likelihood is a function of only the ranks of the times and it would be unchanged if the time scale were transformed by any monotonic transformation.

3. The Proposed CARA Designs

Let β_A and β_B be the population characteristics representing the treatment effects of A and B, respectively. During the initial phase of the trial, one uses some restricted randomization procedure to allocate the initial $2m_0$ patients equally among treatments A and B, where m_0 is a positive integer. This ensures that at least m_0 patients are allocated to each treatment, and m_0 is chosen so that estimates of the parameters (β_A, β_B) can be obtained from this initial sample. At stage m , one computes the partial likelihood estimates ($\hat{\beta}_{AM}, \hat{\beta}_{BM}$) based on the responses of the first m patients, eliminating the effects of the prognostic factors. When the $(m+1)^{\text{th}}$ patient enters the clinical trial with covariate vector $z_m + 1$, this patient is randomized to treatment A with probability $c(\hat{\beta}_{AM}, \hat{\beta}_{BM}, z_m + 1)$ where $0 \le c(.) \le 1$ is an allocation function which bridges the past allocation pattern, response histories and the covariate vector of the m patients to the $(m+1)^{\text{th}}$ allocation with the covariate vector $z_m + 1$. This allocation is chosen with the intention of skewing the treatment allocation probability in favour of the

better treatment arm. A critical assumption made here is that the survival times and the censoring times are independent. Since patients arrive sequentially in the clinical trial and are observed until the end of the trial, the type of censoring considered here is the generalized Type I right censoring scheme. This section discusses the method of derivation of this allocation function using two optimal allocation approaches which target the derived allocation proportions.

Since clinical trials are complex experiments with multiple experimental objectives, a formal optimization procedure can be used to develop the CARA randomization procedures. Treating the baseline hazard as arbitrary makes the design more dependent on the observed data as compared to that of the designs based on parametric models. Such designs therefore increases its applicability in real-life clinical trials.The hazard function plays an important role in any survival trials. Let $\epsilon_k(\boldsymbol{z})$be the probability of event before censoring for a patient with treatment k and with covariate vector z,and $\sqrt{V^{-1}(t_{(i)}, \hat{\beta}^T}$ be the principal square root of the inversed weighted variance matrix of the covariates among the individuals at risk at time $t_{(i)}$ .One way to meet most of the multiple experimental objectives in a clinical trial is to minimize the overall hazard for a patient with a given covariate, subject to the constraint of keeping the asymptotic variance of the difference between the estimated hazard functions for the two treatment groups to be constant. This is done by

$$min : n_A h_A(t|z_j) + n_B h_B(t|z_j)$$

$$subject\ to : z_{jA}^T J^{-1}(\hat{\beta}) z_{jA} e^{2\hat{\beta}^T z_{jA}} + z_{jB}^T J^{-1}(\hat{\beta}) z_{jB} e^{2\hat{\beta}^T z_{jB}} = \mathrm{k}$$

where $J(\hat{\beta})$ is the observed inforation matrix for the Cox regression coefficients. The optimal allocation proportion for treatment A is given by:

$$\pi_{A1}(\beta_A, \beta_B, z) = \frac{\sqrt{\epsilon_B(\mathbf{z}) h_B(t|z_j) z_{jA}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jA} e^{2\hat{\beta}^T z_{jA}}}}{\sqrt{\epsilon_B(\mathbf{z}) h_B(t|z_j) z_{jA}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jA} e^{2\hat{\beta}^T z_{jA}}} + \sqrt{\epsilon_A(\mathbf{z}) h_A(t|z_j) z_{jB}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jB} e^{2\hat{\beta}^T z_{jB}}}} \tag{5}$$

One can use other metrics of treatment difference and obtain different optimal allocations. For instance, minimizing the overall trial size, subject to the constraint of keeping the asymptotic variance of the difference between the estimated hazard functions for the two treatment groups to be constant, leads to the Neymann allocation. The Neymann allocation function for treatment A is given below :

$$\pi_{A2}(\beta_A, \beta_B, z) = \frac{\sqrt{\epsilon_B(\mathbf{z}) z_{jA}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jA} e^{2\hat{\beta}^T z_{jA}}}}{\sqrt{\epsilon_B(\mathbf{z}) z_{jA}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jA} e^{2\hat{\beta}^T z_{jA}}} + \sqrt{\epsilon_A(\mathbf{z}) z_{jB}^T V^{-1}(t_{(i)}, \hat{\beta}^T) z_{jB} e^{2\hat{\beta}^T z_{jB}}}} \tag{6}$$

The doubly-adaptive biased coin design (DBCD) or the efficient randomized adaptive design (ERADE) can be used to allocate the patient and target these derived allocation proportions. After the allocation of the two treatments to m patients and observing their responses, let $N_A(m)$ and $N_B(m) = m - N_A(m)$ denote the numbers of patients assigned to each of the two treatments. When the $(m + 1)^{\text{th}}$ patient enters the clinical trial with covariate vector z_{m+1}, let $\hat{\pi}_m = \pi_A(\hat{\beta}_{AM}, \hat{\beta}_{BM}, z_{m+1})$ represent the estimate of $\pi_A(\beta_A, \beta_B, z)$ based on the responses observed from the m patients, adjusted for the covariate z_{m+1} of the incoming patient. Using the DBCD procedure, the $(m + 1)^{\text{th}}$ patient can be assigned to treatment A with probability $g_{m+1}\left(\frac{N_A(m)}{m}, \hat{\pi}_m\right)$, where $\frac{N_A(m)}{m}$ is the proportion of patients who have been assigned to treatment A after m allocations. Let $\hat{\rho}_m = \sum_{i-1}^{m} \frac{\pi_A(\hat{\beta}_{Am}, \hat{\beta}_{Bm}, z_i)}{m}$ be an estimate of the average target allocation of patients to treatment A, based on the data for the first m patients. The mathematical form of the allocation rule for the $(m + 1)^{\text{th}}$patient entering the clinical trial with covariate vector z_{m+1} ,to be assigned to treatment A is given in Sverdlov,Rosenberger and Ryzenik (2013), whereas the mathematical form for the ERADE allocation rule is given by

$$g_{m+1}\left(\frac{N_A(m)}{m}, \hat{\pi}_m\right) = \begin{cases} \alpha\hat{\pi}_m \text{ if } \frac{N_A(m)}{m} > \hat{\rho}_m, \\ \hat{\pi}_m \text{ if } \frac{N_A(m)}{m} = \hat{\rho}_m, \\ 1 - \alpha(1 - \hat{\pi}_m) \text{ if } \frac{N_A(m)}{m} < \hat{\rho}_m, \end{cases} \tag{7}$$

where $0 \leq \alpha < 1$ is a constant that reflects the degree of randomization. This gives a family of CARA designs that are fully randomized and also asymptotically efficient. The ERADE can be viewed as a generalisation of Efron's biased coin design for any desired allocation function, which may depend on the unknown parameters. If the response distribution belong to the exponential family, the ERADE for any $\alpha \in [0,1)$ is fully efficient. The parameter α controls the degree of randomness of the design. The performance of the various randomization procedures targeting each of the derived allocation proportions is discussed in the next section after performing an extensive simulation study.

4. Simulation Results

The simulation results compare different designs according to three experimental scenarios, the first being the neutral treatment effect, which refers to the hypothetical experimental scenario where treatments A and B are equally effective. In the case of comparing a new treatment with a control, this scenario refers to the situation where the new treatment is as good as the

existing control. The next is the positive treatment effect, which refers to the hypothetical experimental scenario where treatment A is more effective than treatment B, or the new treatment performs better than the control. The third scenario focuses on the negative treatment effect, which refers to the hypothetical experimental scenario where treatment B is more effective than treatment A, or, in the case of comparing a new treatment with a control, this means that the new treatment is not as effective as the control. The procedure used here is a fully sequential one that recalculates the randomization probabilities for every new patient over 5,000 simulation runs for 400 patients, each arriving sequentially into the trial. The censoring scheme considered here is generalized type I right-censoring. The patient's arrival pattern is simulated from a uniform(0,365) distribution, whereas the response of a patient is added to the recruitment time of the patient, and patients whose outcomes have not been observed by a specified time are said to be generalized type I right-censored. The length of the recruitment period is 365 days and the overall trial duration is taken to be 581.66 days. Following Rosenberger, Vidyashankar and Agarwal (2001), a covariate structure of three independent covariates have been generated. These are Gender (Bernoulli, p = 0.5), Age (Uniform[30,75]) and Cholestrol Level (Normal [200,400]). The survival time of a patient with covariate vector $z = (1, z_1, z_2, z_3)^T$ in treatment group k is simulated from the Weibull distribution with scale parameter $\mu_k(z) = \exp(\beta_k^T \boldsymbol{z})$ and shape parameter $\gamma_k = 1.07527$. Since there are three predictive covariates in the model, the direction and magnitude of the treatment difference will vary for the patients, depending on their observed covariate values.

In survival trials, the delay time for a patient is the patient's survival or censoring time. To facilitate CARA designs with delayed responses, it is required that, at the i^{th} patient's randomization time, only data from those patients who have responded before the i^{th} patient's arrival are used for computing the randomization probability for the i^{th} patient. In practice, an assumption of immediate responses is infeasible due to inherent delay in time-to-event outcomes. For the implementation of the CARA designs, initially $2m_0$ patients have been equally allocated to the two treatment arms using a Efron's Biased Coin design. Here, m_0 is a positive integer, and, following Sverdlov, Rosenberger and Ryeznik (2013), it is chosen to be 40 for each treatment arm. For appropriate implementation of the ERADE designs Hu, Zhang and He (2009) recomended that it is reasonable to choose α, the degree of randomness of the design, to be between 0.4 to 0.7. When α is smaller, the ERADE is more deterministic and has a smaller variability. Hu, Zhang and He (2009) showed that the ERADE response-adaptive designs give similar results when $\alpha = 0.5$ as compared to when $\alpha = 0.67$. Here, for the implementation of the ERADE designs, α is chosen to be 0.55, whereas for the implementation

of the DBCD design α is chosen to be 2 following Sverdlov, Rosenberger and Ryeznik (2013) .

The operating characteristics of the proposed adaptive designs as well as the balanced randomization designs have been compared through simulations. It has been found that, when one treatment performs better than the other, all of the proposed CARA designs generate skewed allocations on an average towards the better treatment arm according to covariate-specific treatment effects. This result in fewer events in the trial without compromising much on the statistical efficiency as compared to the balanced randomization designs. The degree of skewness also varies according to the background model that the design is based on. The skewness of the treatment allocation proportions in favour of the better treatment arm thus establishes the ethical gain of using the CARA designs as compared to the traditional balanced randomization procedures.A slight delay in the response does not affect the convergence of the CARA designs to their target allocation proportions. It has been established that even without considering a distriutional assumption for the survival responses and just assuming proportional hazard of patients with respect to time, such ethical gain persists without heavily compromising on the statistical power of the Wald test for the difference of the covariate-adjusted treatment effects. It has been established by simulation results that such ethical gain is achieved most with the DBCD based CARA design, as the allocation proportions converge quickly to the target values. However, the ERADE being the most efficient design provides a method that has the highest power for testing the treatment differences. Therefore, ERADE is preferred over the DBCD if efficiency is the sole criterion for an experimenter. However, to keep a balance between ethics and efficiency, the DBCD procedure outshines the ERADE and the traditional balanced randomization procedure.

5. Discussion and Conclusion

In the present paper, CARA randomization procedures for two-armed survival trials has been considered for semi-parametric survival responses. This means that the procedures are developed without any distriutional assumption of the survival responses but only with a lighter assumption about the hazards of an event at any given time point to be proportional and time-independent for any two patients in the trial. The procedure is based on two distinct approaches to optimality: the doubly-adaptive biased coin design and the efficient randomised adaptive design. The relative merits of the proposed design options, and the respective contexts favouring their preferred applicability, have been highlighted.

The skewness of the treatment allocation proportions in favour of the better treatment results in some reductions in the number of events in the study and establishes the ethical gain obtained by using CARA designs. It has

been established that such ethical gain prevails without significantly compromising on the statistical power of the Wald test for the difference in the partial likelihood based covariate-adjusted treatment effects. Extensive simulation studies have established that the balance between ethical gain and statistical efficiency in estimating treatment effects is achieved most with the CARA design using the doubly-adaptive biased coin design. However, the efficient randomised adaptive design works best if the variability of the allocation procedure is the sole criterion for evaluation.

The mertis of the proposed designs have also been highlighted by redesigning an existing clinical trial from Sverdlov, Rosenberger and Ryeznik (2013), which is intended to be discussed during the talk in the congress.

References

1. Hu, F. and Zhang, L-X. (2004), Asymptotic properties of doubly-adaptive biased coin designs for multitreatment clinical trials, The Annals of Statistics 32, 268–301.
2. Hu, F., Zhang, L-X, and He, X. (2009). Efficient randomized adaptive designs, The Annals of Statistics 37, 2543–2560.
3. Rosenberger, W.F. and Sverdlov, O. (2008), Handling covariates in the design of clinical trials, Statistical Science 23, 404–419.
4. Rosenberger, W.F., Vidyashankar, A.N., and Agarwal, D.K. (2001), Covariate-adjusted response-adaptive designs for binary response, Journal of Biopharmaceutical Statistics 11, 227–236.
5. Sverdlov, O., Rosenberger, W.F., and Ryeznik, Y. (2013), Utility of covariate-adjusted response-adaptive randomization in survival trials. Statistics in Biopharmaceutical Research 5,38–53.

The Impact of Climate on Tourism in the GCC

Amira Al-Salhi
GCC-Stat Centre, Muscat, Sultanate of Oman

Abstract

Many tourist destinations are highly seasonable, shaped in part by the climate conditions. The GCC region has become an important tourist destination, with about 59 million tourists each year. The region has distinct seasons and inbound tourism tends to be influenced by climate conditions. This paper assesses the seasonal variation in tourist numbers in the GCC, in the context of the variation in climatic factors such as temperature, humidity, and rainfall. A case study on Salalah, Oman, which has specific climate conditions, is also presented to highlight the importance of climate as a driver for tourism.

Keywords

Temperatures; rainfall; humidity; seasonal movements; monsoon.

1. Introduction

Tourism refers to specific types of trips that take a traveler outside his/her usual environment for less than a year and for the main purpose other than to be employed by a resident entity in the place visited (UNWTO, 2008). The tourism sector is one of the economically important sectors, which is a source of national income for many countries around the world. Therefore, countries give it great importance and strive to develop it. Tourism occupies third place in terms of its contribution to the international economy after chemicals and fuels with total exports of US $ 1.6 trillion in 2017 (UNWTO, 2018).

Climate changes are one of the world's greatest challenges because of their impact on various facilities and human dimensions. Tourism is one of the economic sectors most sensitive to the potential impacts of climate changes, such as agriculture, environment, and water. The tourism sector in the Arab world is closely related to the landscape, environmental features, and cultural characteristics, directly or indirectly.

Seasonal movements of tourists are affected by climate changes, such as temperature, rainfall, and humidity. Seasonal movements of tourists can be measured by distributing the number of tourists by seasons per year. Kulendran and Dwyer (2012) identified the factors that cause seasonal movements in tourism flows: Natural and Institutional factors. Natural factors relate to the climate changes in the tourist destination such as sunshine, the maximum and minimum temperatures, wind, fog, rainfall, humidity, snow, etc.

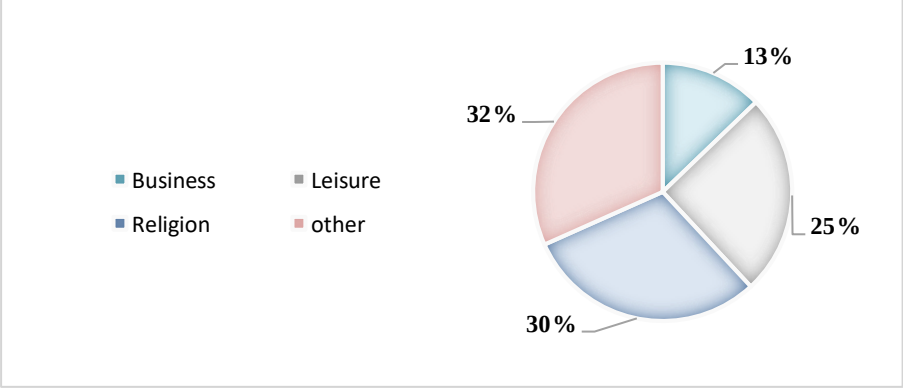
Institutional factors include the calendar effects (religious festivals such as Ramadan, Christmas, etc.) and timing decisions (school and industry vacations, etc.).

This study is primarily concerned with research on the relationship between Natural factors (especially the role of temperature, humidity, and rainfall) and seasonality of tourism in the GCC countries.

Tourism in GCC

The Gulf Cooperation Council (GCC) is a political and economic alliance of six countries in the Arabian Peninsula: Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates. At the level of tourism indicators for the GCC countries, the total number of inbound tourists to the GCC countries in 2017 reached about 59 million tourists which 59% of them are foreign visitors (from outside GCC countries). As Figure 1 shows, about 30% of inbound tourists to the GCC countries in 2017 came for the purpose of religion in Saudi Arabia and 25% came for leisure.

Figure 1: Distribution of inbound tourists' arrival to the GCC according to the main purpose of the trip, 2017



Source: GCC -Stat centre

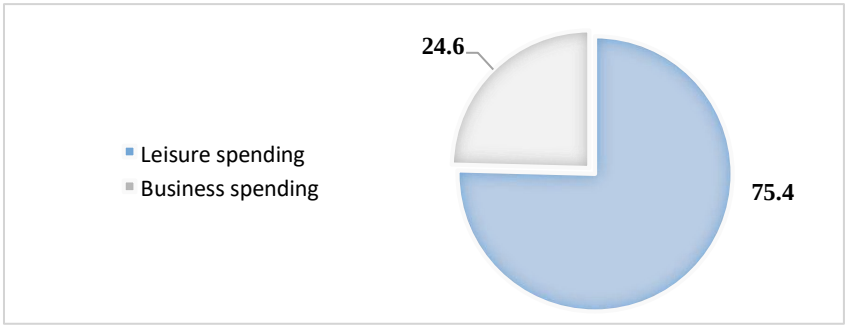
The economic impact of tourism in GCC countries

The tourism industry plays a very large role in the global economy, it contributed 10.4% to global GDP in 2017 and it the world's fastest growing sector at 4.6% (WTTC, 2018).

In GCC, the direct contribution of Travel and Tourism to GDP in 2017 was USD54.7bn (3.8% of GDP).This is forecast to rise by 3.9% to USD56.8bn in 2018 (WTTC, 2018).

Figure 2 shows that leisure travel spending (inbound and domestic) in GCC generated 75.4% of direct Travel and Tourism GDP in 2017 compared with 24.6% for business travel spending. leisure travel spending is expected to grow by 3.7% in 2018, and rise by 4.6% in 2028 and business travel spending is expected to grow by 2.6% in 2018, and rise by 5.4% in 2028 (WTTC, 2018).

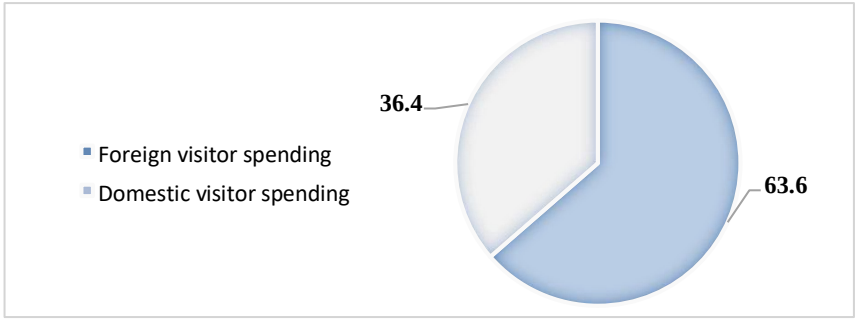
Figure 2: Travel and tourism's contribution to GDP: business vs leisure, 2017



Source: World tourism and travel council

Figure 3 shows that domestic travel spending in the GCC generated 36.4% of direct Travel and Tourism GDP in 2017 compared with 63.6% for foreign visitor spending. Domestic travel spending is expected to grow by 3.3% in 2018, and rise by 3.5% in 2028 and visitor exports are expected to grow by 3.5% in 2018, and rise by 5.5% in 2028 (WTTC, 2018).

Figure 3: Travel and tourism's contribution to GDP: domestic vs foreign, 2017



Source: World tourism and travel council

Tourism statistics project in GCC-Stat center

GCC-Stat (the regional statistics center of the GCC) was established in 2011, to build and strengthen the statistical and institutional capacity of countries to meet the statistical requirements at the GCC level, build a culture of statistics, and strengthen the correct use of data and information in decisionmaking and policy formulation. In 2015, the field of Tourism Statistics was added as one of the priority areas of the Strategic Plan for Common Statistical Work (2015-2020) (GCC-Stat center, 2015). As part of this, GCC-Stat was mandated to propose a strategic plan and a detailed work program for the development of tourism statistic. As part of this, the center evaluated the status of tourism statistics in the GCC countries. The evaluation showed that some countries, such as Oman were more advanced in the compilation of tourism statistics.

2. Methodology

This paper examines the impact of natural factors - temperature, rainfall, and humidity on seasonal variations in GCC tourism. The paper assesses the seasonal variation in tourist numbers, in the context of the variation in climatic factors such as temperature, humidity, and rainfall. A case study on Salalah, Oman, which has specific climate conditions, is also presented to highlight the importance of climate as a driver for tourism in the region. Seasonality in tourism is also caused by institutional conditions as mentioned before. In the case of the GCC, the major institutional tourism event is the annual Hajj. In 2018, about 1.3 million non-Saudi residents took part in the Hajj (General Authority of Statistics, 2018). Because the Islamic calendar is based on the lunar cycle, the dates vary each year. In recent years, the Hajj has occurred during the September quarter. In addition, inbound tourism can also be affected by Ramadan, which also varies each year. In recent years, Ramadan has been in the June quarter. While some tourists may visit for family reasons, generally inbound tourism declines during Ramadan.

3. Results

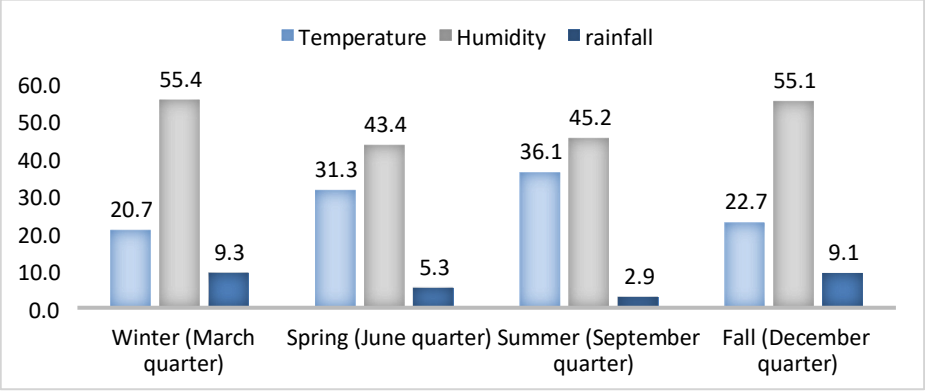
Seasonality of tourism flows in the GCC

The GCC countries are known for their desert climate – long hot summers (May – September) when the daytime temperatures can reach highs of 50°C. The winter months of January to March are cooler, with average daytime temperatures of 20°C.

Figure 4 shows the mean values of temperature, humidity, and rainfall for the different seasons in GCC from January 2007 to December 2017. Different quarters have different mean values for climate variables. The March quarter (winter) has the lowest temperature mean value and the highest amount of rainfall whereas September quarter (summer) has the highest temperature mean value and the June quarter (spring) has the lowest mean value for humidity percentage.

As Figure 4 shows, the winter and fall seasons experience some rainfall. However, there are still many sunny days, where the average days of rain in the year in the GCC is about 15 days only.

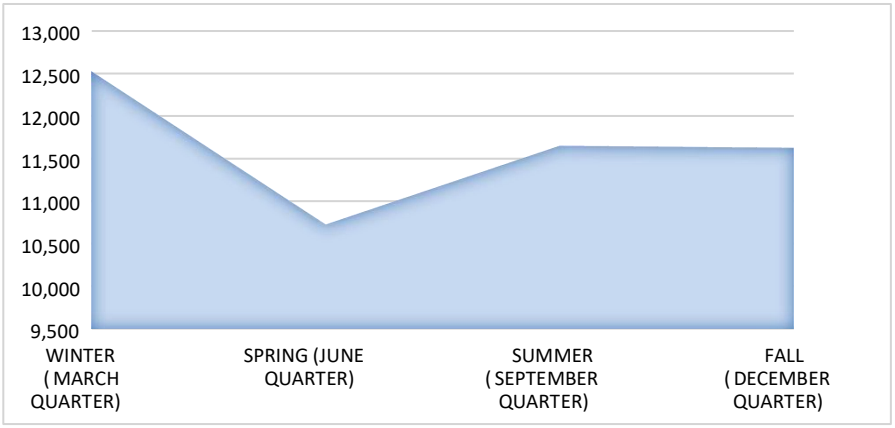
Figure 4: Average temperature (°C), average humidity (%), and average rainfall (mm) in GCC, 2007-2017



Source: National statistical centres in GCC

Tourism numbers increase in the winter months, as Figure 5 shows, while the June quarter has the lowest seasonal levels. The June quarter includes the start of the summer weather, and as noted above included Ramadan in 2017. That year, Ramadan was from May 24 to June 26.

Figure 5: Distribution of inbound tourists' arrival to GCC by season, 2017



Source: GCC-Stat centre

However, while the overall GCC has a desert climate, some parts of the GCC have a semi-tropical climate. One example of this is Salalah in Oman, the subject of the case study.

Case Study – Salalah, Sultanate of Oman

The climate of Salalah is quite different from the conditions in the rest of GCC, due to a combination of the summer monsoon and the local topography. The summer monsoon or Khareef extends from the end of July until the beginning of September. During that time, the city of Salalah and the hills are surrounded by white fog. Light rains drizzle to cool the air. During the Khareef season, temperatures never rise above 27 degrees Celsius (Oman Tourism website).

Table 1 shows the mean values of temperature, humidity, and rainfall for the different seasons in Salalah from January 2011 to December 2017. The September quarter has the lowest temperature mean value and the highest amount of rainfall due to the highest percentage of humidity whereas June quarter has the highest temperature mean value.

Table 1: Average temperature (°C), average humidity (%), and average rainfall (mm) in Salalah, 2011-2017

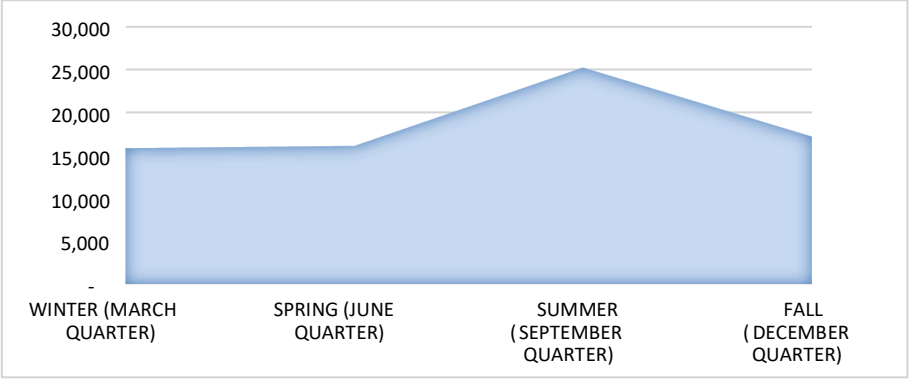
	Temperature (°C)	Humidity (%)	Rainfall (mm)
Winter (March quarter)	32.1	84.8	1.1
Spring (June quarter)	34.6	90.3	1.8
Summer (September quarter)	**30.3**	**91.3**	**10.0**
Fall (December quarter)	33.2	84.3	4.9

Source: National Centre for Statistics & Information

Salalah attracts many people from other parts of Oman, the GCC and other Arab countries during the monsoon/khareef season.

The total number of tourists to Salalah in fall season in 2017 reached about 644,931 tourists, with an average increase of 12% from 2011 to 2017. Around 94% of tourists to Salalah in 2017 came for leisure. Domestic visitors (from Oman) accounted for 71% of all tourists to Salalah in the fall season. Figure 6 shows the mean of tourist flows to Salalah by season from January 2011 to December 2017. The September quarter has the heights seasonal mean value and June quarter has the lowest seasonal mean value.

Figure 6: Average number of tourists' arrival to Salalah by season, 2011-2017



Source: National Centre for Statistics & Information

Figure 6 and table 1 show that there is a link between the seasonal variation in tourist arrivals and the climate variables. For example, The September quarter seasonal variation mean value is high, which is consistent

with low-temperature value and high rainfall amount in Salalah in this quarter. While, the June quarter seasonal variation mean value is low, which is consistent with high-temperature value.

4. Discussion and Conclusion

Tourism to the GCC appears to be influenced by the climate. Tourism numbers were higher in the winter season, when the climate conditions are favourable, and declines when the climate is less suitable for outdoor activities. The case study provides a specific example of how the climate is a direct driver for tourism in the GCC. Domestic tourists from Oman, along with tourists from other parts of the GCC and the Arab region, travel to one part of Oman, based on the specific climate conditions.

Most of the international literature (see for example Boken, 2010 and Kulendran and Dwyer, 2012) who have studied the impact and importance of climate on tourism have focused on the importance of beach weather or snow as attractors for tourism. This case study has highlighted a different climate attraction – mild monsoon weather with high humidity, but relatively low temperatures.

In this case, the weather at home is also an important driver – although in a different way to that summarised by Boken, who noted in her literature review that "Unfavourable climate or poor weather conditions act as a push factor for tourists to travel to warmer and drier locations." In this case, the hotter weather in the home locations, are likely to be a strong push factor to visit Salalah. Indeed local advertising for Salalah emphasises the temperature in Muscat (at 45°C) as a reason to visit. This paper has examined the relationship between Climate variables and seasonal variation in tourist numbers in the GCC. Across the world, Climate is important to tourism because it attracts tourists who expect favourable weather conditions in their holiday destination and as has been shown by the case study, it plays a major role in the marketing of holiday tourism to destinations.

This study has shown that these international patterns also apply to the GCC. At the GCC level, there is evidence that tourism numbers increase in the winter months and a decline in the summer when the temperatures are very high. The case study has shown a different example of climate tourism – when the summer monsoon/Khareef with high humidity, fog, and rain is a major attraction in a region dominated by desert conditions.

Other factors influence decisions on tourism destinations. The international literature points to factors such as income, the price of tourism, cost of transportation, and cost of living at the destination. These factors were not considered here, but may be considered for future studies, as GCC-Stat works with member states in the GCC to increase the range and availability of tourism statistics.

References

1. GCC-Stat Center. (2018). a Glimpse at Tourism Statistics in the GCC Countries for 2017. Retrieved from https://www.gccstat.org/en/statistic/publications/gulf-tourism-overview
2. World tourism and travel council. (2018). Economic Impact 2018. Retrieved from https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2018/world2018.pdf
3. Kulendran and Dwyer. (2012). Modelling Seasonal Variation in Tourism Flows with Climate variables. Retrieved from http://vuir.vu.edu.au/23288/1/s2.pdf
4. Boken. (2010). the importance of Climate and Weather for Tourism. Literature Review. Retrieved from http://www.lincoln.ac.nz/PageFiles/6750/WeatherLitReview.pdf
5. GCC-Stat Center. (2015). Strategic Plan for Common Statistical Work (2015-2020).
6. World Tourism Organization. (2018). UNWTO Tourism Highlights. Retrieved from http://www2.unwto.org
7. World Tourism Organization. (2008). International Recommendations for Tourism Statistics 2008. Retrieved from https://www.e-unwto.org/doi/book/10.18111/9789211615210
8. Oman Tourism. Dhofar Governorate. Retrieved from http://www.omantourism.gov.om/wps/portal/mot/tourism/oman/home/sultanate/regions/dhofar/!ut/p/a0/04_Sj9CPykssy0xPLMnMz0vMAfGjzOL9gwKD3fxcTQwMvN0NDTyN3F0DDA2DDU0NTfQLsh0VAfUTWlM!/
9. General Authority of Statistics. (2018). Censuses. Retrieved from https://www.stats.gov.sa/en/28
10. National Centre for Statistics and Information. Publication. Retrieved from https://www.ncsi.gov.om/Elibrary/Pages/LibraryContentView.aspx
11. General Authority of Statistics. Statistics Library. Retrieved from https://www.stats.gov.sa/en/node
12. Ministry of Development Planning and Statistics. Statistics. Retrieved from https://www.mdps.gov.qa/en/statistics1/pages/default.aspx
13. Central Statistical Bureau. Statistics and Bulletins. Retrieved from https://www.csb.gov.kw/Pages/Statistics_en?ID=18&ParentCatID=2
14. Information and eGovernment Authority. Publications. Retrieved from http://www.iga.gov.bh/en/category/publications
15. Federal Competitiveness and Statistics Authority. Statistics. Retrieved from http://fcsa.gov.ae/en-us/Pages/Statistics/Statistics-by-Subject.aspx#

Self-organizing ensemble of LSTM to enhance the air pollution estimation in Santiago of Chile

Javier Linkolk López-Gonzales[1,2], Rodrigo Salas Fuentes[3], Cristian Ubal[1], Orietta Nicolis[4], Romina Torres[4]

[1]Facultad de Ciencias, Instituto de Estadística, Universidad de Valparaíso
[2]Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión
[3]Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso
[4]Facultad de Ingeniería, Universidad Andrés Bello

Abstract

In highly populated and / or industrialized cities, they are affected by high levels of pollutant concentration in the air, which seriously affects the health of their inhabitants. In this article we propose a new ensemble technique of Long-Short Term Memory networks based on self-organizing maps to enhance the estimation if PM2.5 concentration in urban Santiago, Chile. Simulation results with time series collected in one of the Station shows very promising results.

Keywords

Air pollution; Ensemble of Artificial Neural Networks; Long-Short Term Memory; Self-Organizing Maps

1. Introduction

In recent years, the presence of haze has been frequent in the southern areas of Chile, mainly in Santiago. The main factor of air pollution is fine particulate matter (PM2.5, PM · 2.5μm in diameter). This was presented for the first time as an indicator for the ambient air quality standard by the American Environmental Protection Agency (EPA) in 1997 [1].

It was detected that the surface of these particles is linked to several toxic products, such as organic compounds, elements and bacteria [2]. Because they are so small, they only micrometers in size, and it is easy for them to cross the respiratory system and enter the lung or even enter the blood, which can cause asthma, respiratory inflammation and cancer [3]. Some studies reported correlations between PM2.5 and death ([4], [5], [6]). Taking into account the influence of PM2.5, the World Health Organization recognized PM2.5 as carcinogens for the first time in October 2013.

In this sense, improving air quality is one of the major environmental challenges of this century [7]. Many cities in Latin American countries are heavily polluted with PM2.5 and PM10, affecting human health and life quality. In Chile's capital, the Santiago Metropolitan Area (SMA), PM 2.5 concentrations systematically exceed values defined by World Health

Organization (WHO) standards and Chilean national air quality standards (NAQS) [8].

This pollution episodes are associated with weather conditions [9], mainly in the cold season, and there is little understanding of how the variation in particle matter differs between cities and how this is affected by the meteorological conditions [10].

On the other hand, the interest in Machine Learning has exploded in the last decade. Fundamentally, Machine Learning is the use of algorithms to extract information from raw data and represent it through some kind of mathematical model.

Under this context, the research aims to clusterize scenarios and estimation of higher PM2.5 pollution in the "La Florida" area, categorized as the zone with the highest pollution index in the Metropolitan Capital. For this, machine learning techniques will be used, such as self-organized maps (SOM) and long-term memory networks (LSTM).

2. Methodology

Self-Organizing Maps

The Self Organizing Maps (SOM) model was introduced by T. Kohonen [11]. The model preserves the topology mapping from the high-dimensional input space onto a low-dimensional display.

The Map **M** consists of an ordered set of prototypes $\mathbf{w_k} \in \mathbf{W} \subset \mathbf{R^d}$, $k=\mathbf{1...M}$, with a neighbourhood relation between these units forming a grid, where k indexes the location of the prototype in the grid. The most common used lattices are the linear, the rectangular and the hexagonal array of cells. In this work we will consider a rectangular grid where $K(\mathbf{w_k})=(i, j) \in \mathbf{N^2}$ is the vectorial location of the unit $\mathbf{w_k}$ in the grid, where i and j stand for the row and column of the prototype in the rectangular array [12].

When the data vector $x \in$ Rd is presented to the model **M**, it is projected to a neuron position of the low dimensional grid by searching the best matching unit (**bmu**), i.e., the prototype that is closest to the input.

The learning process of this model consists in moving the reference vectors towards the current input by adjusting the location of the prototype in the input space. The winning unit and its neighbours adapt to represent the input.

Long Short-Term Memory (LSTM)

The LSTM are a type of recurrent neural network with memory over a period of time by adding to "memory cell". As indicated in [13], LSTM RNNs addresses the problem of the misclassification in RNNs incorporating activation functions in their state dynamics. This "memory cell" is controlled mainly by "the entrance door", "the door of forgetfulness" and "the exit door".

The "entry door" activates the entry of information to the "memory cell", and "the forgetting door" selectively erases certain information in the cell memory and activates the storage to the next entry [14]. Finally, "the exit door" decides what information the memory cell will emit [15]. The LSTM network structure is illustrated in figure 1. At each time step, an LSTM maintains a hidden vector h and a memory vector m responsible for controlling updates and outputs of the state. More concretely, Graves et al. [16] define the computation at time step t as follows:

$$\mathbf{g}u = \sigma(\mathbf{W}u\mathbf{h}t-1+\mathbf{I}u\mathbf{x}t)$$
$$\mathbf{g}f = \sigma(\mathbf{W}f\mathbf{h}t-1+\mathbf{I}f\mathbf{x}t)$$
$$\mathbf{g}o = \sigma(\mathbf{W}o\mathbf{h}t-1+\mathbf{I}o\mathbf{x}t)$$
$$\mathbf{g}c = \tanh(\mathbf{W}c\mathbf{h}t-1+\mathbf{I}c\mathbf{x}t)$$
$$\mathbf{m}t = \mathbf{g}f 0\ \mathbf{m}t-1 + \mathbf{g}^{u} 0\ \mathbf{g}c$$
$$\mathbf{h}t = \tanh(\mathbf{g}^{o} 0 \mathbf{m}t)$$

where σ is the logistic sigmoid function, ⊙ represents elementwise multiplication, $\mathbf{W}u$, $\mathbf{W}f$, $\mathbf{W}o$, $\mathbf{W}c$ are recurrent weight matrices and $\mathbf{I}u$, $\mathbf{I}f$, $\mathbf{I}o$, $\mathbf{I}c$ are projection matrices [17].

In addition, they classify and predict based on time series data, since, there may be delays of unknown duration between important events in a series of time. It allows clearly remembering events selected from far away in the past, which contrasts with basic NRs, for which the memory of an event decays over time [18].

Self-Organizing Ensemble of LSTM

The proposal is aimed at how to find temporary pollution groups and also improve the estimation with LSTM networks, that is, a SOM + LSTM assembly would be developed.

It is important to mention that each neuron is a network. Virtually, it becomes a neuron properly, when evaluating the membership of the data, however, when predicting is a whole network.

Each node is a network that specialized in a dataset. The time series are grouped, and one of the neurons will be the one that best represents the time series. However, the same neuron becomes an LSTM, and being the one that best models the series, delivers the prediction. Inclusive, the neighbouring neurons, also contribute relevant information, by principle of the self-organization as neighbourhood of the winning unit.

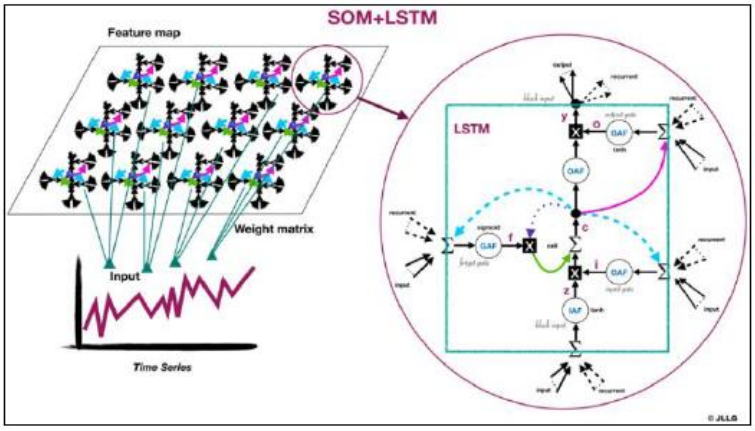


Figure 1: SOM ensemble of LSTM

3. Results

We present the results of a neural network SOM+LSTM model designed for the characterization and estimation of hourly pollution PM2.5 concentrations by one station in the Metropolitan area: "La Florida".

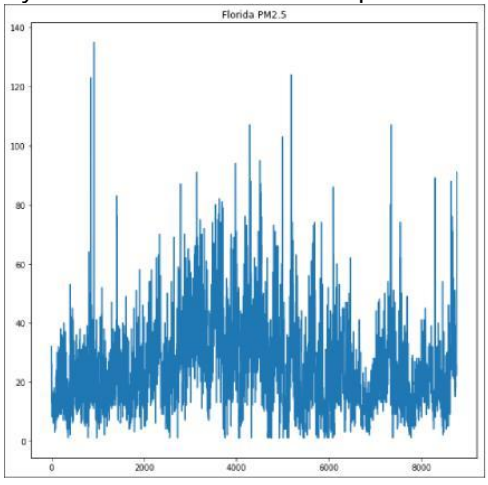


Figure 2: Time series of PM2.5 "La Florida"

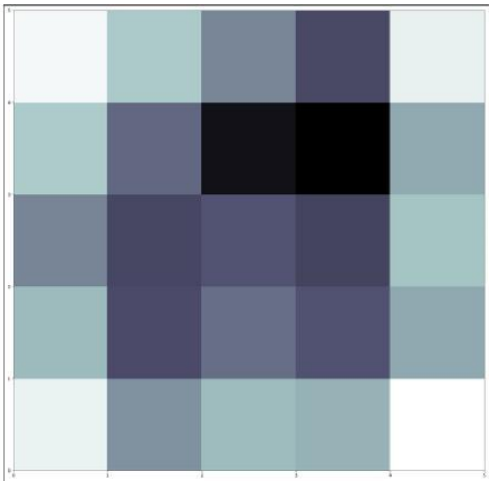


Figure 3: SOM of the data

In figure 2, the behaviour of the contamination data of the "La Florida" station during the 365 days is reported. The SOM is used to group the time series according to the pattern. Figure 3 reports neurons that have a group of similar time series corresponding to the pollution record per hour in a day.

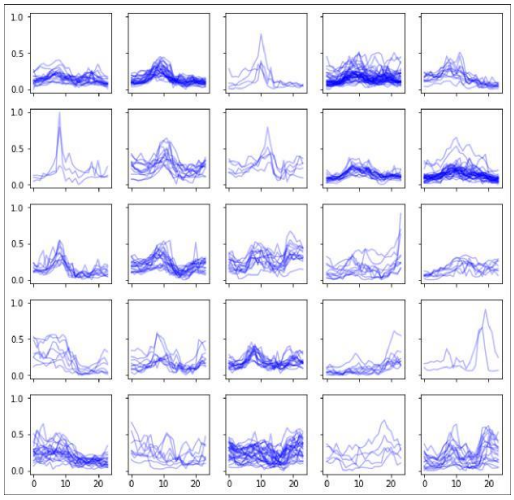


Figure 4: SOM-Clustering

According to Figure 4, an SOM is observed that groups the daily pollution series in "La Florida", clustering the days.

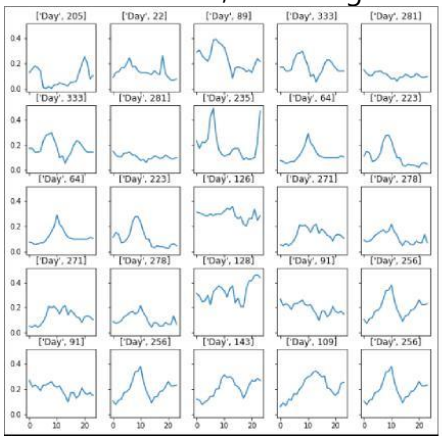


Figure 5: LSTM Generated - 25

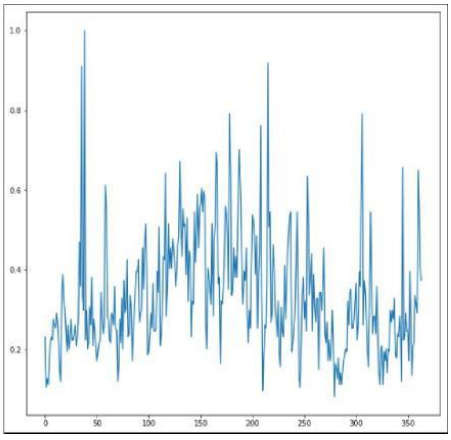
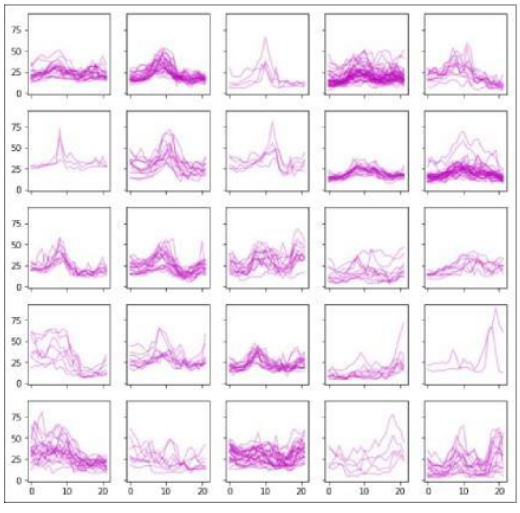


Figure 6: PM2.5-Generated



TSN	RMSE
Train Score neuron (4,0):	6.71
Train Score neuron (0,3):	5.79
Train Score neuron (4,4):	11.33
Train Score neuron (0,0):	6.25
Train Score neuron (1,4):	5.24
Train Score neuron (1,3):	5.08
Train Score neuron (0,4):	6.43
Train Score neuron (0,1):	5.65
Train Score neuron (2,3):	7.86
Train Score neuron (1,2):	11.10
Train Score neuron (2,1):	7.78
Train Score neuron (3,4):	17.97
Train Score neuron (1,0):	19.46
Train Score neuron (2,4):	5.59
Train Score neuron (3,3):	5.58
Train Score neuron (4,2):	8.74
Train Score neuron (1,1):	8.62
Train Score neuron (3,2):	7.11
Train Score neuron (3,0):	8.33
Train Score neuron (3,1):	9.98
Train Score neuron (2,0):	9.03
Train Score neuron (2,2):	11.86
Train Score neuron (4,3):	9.72
Train Score neuron (4,1):	9.67
Train Score neuron (0,2):	11.27

Figure 7: LSTM-Data Figure 8: TSN-RMSE

Finally, according to Figure 5,6 and 7, the LSTM are presented, where each square represents a neuron, that is, each neuron reflects an LSTM. This shows that there are 25 LSTMs that generate the 365 time series.

In parallel, in figure 8, the root of the mean quadratic error (RMSE) of each neuron is observed along with its Train Score (TSN), reporting the lowest RMSE in the neuron (1,3).

4. Discussion and Conclusion

Under the LSTM model, the estimation and adjustment report of the "La Florida" station had important results close to the real data. In the same way that [19] in this research it is shown that the LSTM has a good capacity of adaptation to estimate the concentration of PM2.5, even more when it merges with the SOM at the time of grouping.

The assembly between SOM and LSTM, allowed grouping the time series according to the determined pattern. This showed neurons that have a group of similar time series that corresponds to the pollution record per hour in a day.

For each neuron an LSTM network was adjusted, able to model its data well; thus, one LSTM is trained per neuron, but each LSTM learns several time series.

As future work, different assembled algorithms will be competed, both for estimation and clustering, with the assembly proposal that was made.

References

1. Liang, B., Li, X. L., Ma, K., and Liang, S. X, "Pollution characteristics of metal pollutants in PM2. 5 and comparison of risk on human health in heating and non-heating seasons in Baoding, China", Ecotoxicology and environmental safety, vol. 170, pp. 166-171, 2019.
2. Xing, Y. F., Xu, Y. H., Shi, M. H., and Lian, Y. X, "The impact of PM2. 5 on the human respiratory system", Journal of thoracic disease, vol. 8, no 1, p. E69, 2016.
3. Wang, C., Tu, Y., Yu, Z., & Lu, R, "PM2. 5 and cardiovascular diseases in the elderly: An overview", International journal of environmental research and public health, vol. 12, no 7, p. 8187¬8197, 2015.
4. Zhang, Q., Jiang, X., Tong, D., Davis, S. J., Zhao, H., Geng, G., and Ni, R, "Transboundary health impacts of transported global air pollution and international trade", Nature, vol. 543, no 7647, p. 705, 2017.
5. Atkinson, R. W., Kang, S., Anderson, H. R., Mills, I. C., and Walton, H. A., "Epidemiological time series studies of PM2. 5 and daily mortality and hospital admissions: a systematic review and meta-analysis", Thorax, 2014, p. thoraxjnl-2013-204492, 2014.
6. Zhao, H., Li, X., Zhang, Q., Jiang, X., Lin, J., Peters, G. P., and Zhang, L., "Effects of atmospheric transport and trade on air pollution mortality in China", Atmospheric Chemistry and Physics, vol. 17, no 17, p. 10367-10381, 2017.
7. A. P. K. Tai, L. J. Mickley, and D. J. Jacob, "Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change", Atmospheric Environment, vol. 44, no. 32, pp. 3976–3984, 2010.
8. O. Nicolis, C. Camano, J. C. Marin, and S. K. Sahu, "Spatio-temporal modelling for assessing air pollution in santiago de chile", in AIP Conference Proceedings, vol. 1798, no. 1. AIP Publishing, 2017, p. 020113.
9. H. Riojas-Rodriguez, A. S. da Silva, J. L. Texcalac-Sangrador, and G. L. Moreno-Banda, "Air pollution management and control in Latin America and the Caribbean: implications for climate change", Revista panamericana de salúd pública = Pan American journal of public health, vol. 40, no. 3, pp. 150–159, 2016.
10. M. A. Yáñez, R. Baettig, J. Cornejo, F. Zamudio, J. Guajardo, and R. Fica, "Urban airborne matter in central and southern Chile: Effects of meteorological conditions on fine and coarse particulate matter", Atmospheric Environment, vol. 161, pp. 221–234, 2017.
11. T. Kohonen, "The self-organizing map", Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480, 1990.

12. R. Salas, S. Moreno, H. Allende, and C. Moraga, "A robust and flexible model of hierarchical self-organizing maps for non-stationary environments", Neurocomputing, vol. 70, no. 16-18, pp. 2744– 2757, 2007.
13. Hochreiter, S., and Schmidhuber, J., "Long short-term memory", Neural computation, 1997, vol. 9, no 8, p. 1735-1780, 1997.
14. F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM", Neural Computation, vol. 12, no. 10, pp. 2451–2471, 2000.
15. F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks", Journal of Machine Learning Research, vol. 3, no. 1, pp. 115–143, 2003.
16. A. Graves, "Supervised Sequence Labelling With Recurrent Neural Networks", vol. 385. London, U.K.: Springer, 2012.
17. Karim, F., Majumdar, S., Darabi, H., and Chen, S., "LSTM fully convolutional networks for time series classification", IEEE Access, vol. 6, pp. 1662-1669, 2018.
18. F. Cady, The Data Science Handbook. John Wiley & Sons, 2017.
19. Yan, L., Zhou, M., Wu, Y., and Yan, L., "Long Short Term Memory Model for Analysis and Forecast of PM2.5", "International Conference on Cloud Computing and Security", pp. 623-634, Springer, Cham, 2018.

One-sided misclassification of a binary confounder and bias when estimating causal effects

Ronnie Pingel
Department of Statistics, Uppsala University, Sweden

Abstract
Measurement errors in the confounders is often neglected when adjusting for confounders in observational studies. This study is a contribution in that we study the case of non-differential mis¬classification of a binary confounder. For the case of linear models an expression is provided so that applied researcher may study whether the causal effect is under-estimated or over-estimated. Similar pattern occurs for log-linear models and risk ratios. The results regarding logistic regression and the odds ratio is less clear.

Keywords
Average treatment effect; Non-differential; Differential; Measurement error;

1. Introduction

The aim of causal inference in observational studies is to study the effect of a single variable (treatment, intervention exposure, etc.) on an outcome. In order to estimate causal effects in observational studies, the researcher needs to include all true confounders in the analysis to achieve an unbiased estimate. This is known and acknowledged by all researchers trying to make any causal claims based on such a study.

Well-known is also that it matters how the confounders are used in the statistical analyses, and that the adjustment for confounding can be made more robust by applying non-parametric or semi-parametric methods, e.g. matching on covariates or propensity score-based methods (Waernbaum, 2012; Stuart, 2010).

What to a certain degree is perhaps less emphasized, at least by applied researchers, is that observational studies of any design require accurate measurements of the confounders included in the statistical analysis. Still, measurement error is one of the main sources of bias (Greenland, 1983; Rothman, 2008; Willett, 1989). Furthermore, attention have mostly been on measurement error of the treatment or the outcome (Armstrong, 1994), and not the confounding variables.

Thus, although measurement error is a common feature of empirical data, especially when working with data from registers, the consequences of using confounders with measurement error are often neglected (Brakenhoff, 2018).

In this study we focus on binary confounding variables. Usually, measurement error of binary variables is called misclassification, which can be divided into different types of misclassification. It is common to distinguish between non-differential misclassification, i.e. misclassification that is random, and differential misclassification, that is misclassification that varies between groups.

There are some results in the literature covering the case of non-differential misclassification of a confounder, e.g. Greenland (1980) argued that non-differential classification of a binary confounder leads to attenuation bias. However, even though this was further studied by e.g. Ahlbohm (1992), Fox (2005), Greenland (1996), Savitz (1989), Fung (1984) no formal proofs were given. Moreover, none of these studies related the bias to causal quantities, instead the bias was related to parameters in statistical models.

Formal proofs for non-differential classification of a binary confounder were eventually provided by Ogburn and Vanderweele (2012). They showed that under a monotonicity assumption, i.e. the direction of causal effect is same across levels of the confounder, there is in fact an attenuation bias. Their result is an important contribution and they also relate the bias to parameters within the causal inference framework.

However, Ogburn and Vanderweele (2012) do not study misclassification in the differential case. Although Di Martino et al. (2014) discover that differential misclassified confounder can lead to unpredictable consequences and misleading results for logistic regression models when studying treatment quality of caesarean sections in different hospitals, they only study misclassification using empirical data and no theory.

The focus of this study is a certain type of misclassification denoted one-sided misclassification that is differential. It occurs when the classification is always correct for individuals coded as belonging to a certain category, but not necessarily for those not belonging to the category.

One-sided misclassification may occur due to under-reporting and it would be the case when all individuals being registered having a characteristic really have it, however, among those individuals not registered as having the characteristic some still have it. A typical confounder could be "dementia", where it would be rather safe to assume that all individuals registered as having dementia really has impaired cognitive ability, but there are individuals in the register having dementia but are misclassified as having "no dementia". Another example would be diabetes.

Further, and more generally, this study fits into the broader category of sensitivity analyses, that is, to assess the robustness of empirical evidence by examining how one's estimate varies when a key assumption is relaxed (Rosenbaum, 2002).

2. Methodology

To study misclassification of the aforementioned type, this paper uses formal mathematical proofs, simulations and an empirical analysis.

In the empirical analysis, a register-based evaluation of vocational rehabilitation, serves as an illustration. As commissioned by the Swedish government in 2011, the Public Employment Service (PES) and the Swedish Social Insurance Agency (SIA) implemented in 2012 a model involving enhanced cooperation between the two government agencies. The aim was to target individuals entering sick leave and identify the need of support in order to regain work ability.

The individuals' work-ability, both from a medical and a labour market viewpoint, was identified through a joint assessment meeting (JAM) where individuals during an evaluation period had meetings with PES and SIA. Because individuals were not randomly assigned to JAM, but instead assigned according to the decision of each individual's case-worker there is reason to believe that the characteristics of individuals assigned to JAM are not similar to those not assigned to JAM.

Thus, when evaluating the causal effect of being called to JAM, on for instance the total extent of sick leave, it is necessary to make the JAM group and non-JAM group comparable. To handle the selection bias, Fowler et al. (2017) (which was funded by Swedish government) designed a protocol for the estimation of the causal effect of JAM. The protocol includes several confounders, for instance year of birth, sex, country of origin, marital status, education level.

Furthermore, the two confounders, Mental and behavioural disorders (Chapter V in the ICD-10 classification) and Diseases of the musculoskeletal system and connective tissue (Chapter XVIII in the ICD-10 classification), were pointed out as extra important. Following the protocol, the effect of JAM on sick leave was evaluated in Fowler et al (2017).

However, although different sources of biases are acknowledged in Fowler et al. (2017), misclassification of Mental and behavioural disorders and Diseases of the musculoskeletal system and connective tissue is not mentioned. Yet, studies show that there is reason to believe that for instance mental disorders are underreported (e.g. Ezzat 2015; Takaynagi, 2014). It would therefore be important to study how misclassification affects the causal effect estimates in Fowler et al. (2017).

Moreover, because the degree of one-sided misclassification probably differs when comparing Mental and behavioural disorders and Diseases of the musculoskeletal system and connective tissue this empirical example would further lend itself to illustrate the bias of misclassification.

3. Results

We show that differential one-sided mis-classification can lead to the estimated causal effect being greater than or smaller than the true causal effect. This is opposed to Ogburn and VanderWeele (2012), who show for non-differential misclassification that the estimated causal effect always lies between the true and the crude causal effect.

For the case of linear regression, we derive an expression similar to well-known large-sample omitted variable bias approximation in linear regression. Thus, given some assumptions regarding the correlations between the measurement error and the other variables in the model, this expression can be used to say whether the effect is over-estimated or under-estimated.

Differential mis-classificiation is also studied for a binary outcomes using logistic and log-linear models for the odds ratio and the relative risk. No mathematical proofs are given for these models, but simulation studies show that the bias in the log-linear model for the relative risk, has the same pattern as linear regression. However, for logistic model and odds-ratio, the pattern is not as clear, probability due to the non-collapsibility property of the odds ratio.

Depending on the assumptions regarding the prevalence mental disorders, the re-analysis in this paper show that the surprising negative findings in in Fowler, 2017, regarding the effect of vocational rehabilitation is potentially explained by underreporting in Swedish registries.

4. Discussion and Conclusion

Measurement errors in the confounders is often neglected when adjusting for confounders in observational studies. This study is a contribution in that we study the case of non-differential mis-classification. For linear models an expression is provided so that applied researcher can study whether the causal effect is under-estimated or over-estimated. The results regarding the odds ratio is less clear.

This research is funded by The Institute for Evaluation of Labour Market and Education Policy, which is a research institute under the Swedish Ministry of Employment, situated in Uppsala, Sweden. IFAU's objective is to promote, support and carry out scientific evaluations.

References

1. Ahlbom, A. and Steineck, G. (1992). Aspects of misclassifacation of confounding factors. American journal of industrial medicine, 21(1):107{112.
2. Armstrong, B. K., White, E., and Saracci, R. (1994). Principles of exposure measurement in epidemiology. monographs on epidemiology and biostatistics. Oxford: Oxford University Press.
3. Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., and van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. Journal of clinical epidemiology, 98:89{97.
4. Di Martino, M., Fusco, D., Colais, P., Pinnarelli, L., Davoli, M., and Perucci, C. A. (2014). Differential misclassification of confounders in comparative evaluation of hospital care quality: caesarean sections in Italy. BMC public health, 14(1):1049.
5. Ezzat, V. A., Lee, V., Ahsan, S., Chow, A. W., Segal, O., Rowland, E., Lowe, M. D., and Lambiase, P. D. (2015). A systematic review of icd complications in randomised controlled trials versus registries: is our `real-world'data an underestimation? Open Heart, 2(1):e000198.
6. Fowler, P., de Luna, X., Johansson, P., Ornstein, P., Bill, S., and Bengtsson, P. (2017b). Study protocol for the evaluation of a vocational rehabilitation. Observational Studies, 3:1{27.
7. Fox, M. P., Lash, T. L., and Greenland, S. (2005). A method to automate probabilistic sensitivity analyses of misclassified binary variables. International journal of epidemiology, 34(6):1370{1376.
8. Fung, K. Y. and Howe, G. R. (1984). Methodological issues in case-control studies iii:|the e_ect of joint misclassification of risk factors and confounding factors upon estimation and power. International journal of epidemiology, 13(3):366{370.
9. Greenland, S. (1980). The effect of misclassification in the presence of covariates. American journal of epidemiology, 112(4):564{569.
10. Greenland, S. (1996). Basic methods for sensitivity analysis of biases. International journal of epidemiology, 25(6):1107{1116.
11. Greenland, S. and Kleinbaum, D. G. (1983). Correcting for misclassfication in two-way tables and matched-pair studies. International Journal of Epidemiology, 12(1):93{97.
12. Ogburn, E. L. and VanderWeele, T. J. (2012). On the nondifferential misclassifcation of a binary con¬founder. Epidemiology (Cambridge, Mass.), 23(3):433.
13. Rosenbaum, P. R. (2002). Observational studies. In Observational studies, pages 1(17. Springer. Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). Modern epidemiology.

14. Savitz, D. A. and Bar_on, A. E. (1989). Estimating and correcting for confounder misclassification. American Journal of Epidemiology, 129(5):1062{1071.
15. Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1.
16. Takayanagi, Y., Spira, A. P., Roth, K. B., Gallo, J. J., Eaton, W. W., and Mojtabai, R. (2014). Accuracy of reports of lifetime mental and physical disorders: results from the baltimore epidemiological catchment area study. JAMA psychiatry, 71(3):273{280.
17. Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. Statistics in medicine, 31(15):1572{1581.
18. Willett, W. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. Statistics in Medicine, 8(9):1031{1040.

Earning and shares of paid employees in different industries by citizenship

Noor Ismawati Mohd Jaafar[1], Nor Alkashah Arif Shah[2], Zaidatul Azreen Zulkiple[2]

[1]Social Wellbeing Research Centre, University of Malaya

[2]Manpower and Social Statistics Division, Department of Statistics Malaysia

Abstract

While studies had shown that non-local workers is giving a positive impact to the Malaysian productivity, the public raised many concern over varios issues. One of their worrieness is the employment lost by the local workers. Hence this paper investigate the distribution of paid employees in various industry and examine whethere they are differentiable by citizenship status. Based on Salaries and Wages Survey 2010-2017, the analysis also includes earnings. The data indicates that majority of the citizen employees worked in manufacturing sector, construction sector and three industries under other sectors. In contrast, almost all of the non-citizen works in agriculture sector, manufacturing sector, construction sector and another 2 industries under other sectors. The analysis also shows that the proportion of non-citizen employees were small and their earning is comparatively lower. However in certain industries, these group of employees earns significantly higher. The analysis also shows that the gap in proportion of employee and earning in agriculture was relatively small. It is concluded that distribution of the paid employee varies by industry and citizenship except in agriculture. The same conclusion can also be made on earning distributions.

Keywords

Earnings; Primary occupation; Monthly pay; Foreign worker

1. Introduction

The growing number of foreign worker can be seen not only in developed countries but also in developing countries. The foreign worker can be classify wether they have had low, medium or high skills (Huber et al., 2010). While the migrant workers is normally associate with low and medium skill workers, the expatriates is referred to those with the high skill-high pay-high education(Howell & Newman, 1959; Mandell, 1958; Thompson, 1959; Wallace, 1959; Schollhammer, 1969; as summarised in McNulty & Brewter, 2017). In 2017, 98.5 percent of the foreign paid employee in the Malaysian work force were those with low and medium skill. Eventhough the increase of these workers especially the migrant workers is a public concern in Malaysia, their existence had shown to give positive productivity effects on the economy

(Jordaan, 2018). Despite the highly perceived negative connotation in the country's foreign worker depency, Jordaan(2018), Nur`ain Achim et al. (2017) and Ramesh Kumar Moona Haji Mohamed et al. (2012) highlighted the concentration of them in few specific industries.

Effective from Jannuary 2013, the Malaysian Government enforced the Minimum Wages Order (MWO). The wage was set at a minimum of MYR900 in Peninsular Malaysia and MYR800 in East Malaysia. However, these wages is not applicable to domestic workers. The amounts were increased into MYR1000 and MYR920 starting from July 2016, respectively (Malaysia, 2016). This government intervention in earning by introducing minimum wage is widely perceived to cause a negative impact in the economy as it increase the labour cost.. Hence this paper presents the labour distribution of the local and foreign workers by industry and its earning pattern.. The analysis only consider primary occupation and does not include annualised monetary benefit. Analysis is done on selected years and industries. The analysis does not differentiate the employees by their skill level.

2. Methodology

Salaries & Wages Survey (SWS) is one of the modules in the Labour Force Survey which is carried out from January until December based on guidelines and recommendations of the International Labour Organization (ILO) with reference to An Integrated System of Wages Statistics. SWS main objective is to obtain the wage rate from the principal occupation through household approach collected from respondents aged 15 years and over and fulfil at least one of the following criteria; i. Employees who are working full-time; ii. Employees who did not work during the reference month but receiving salaries & wages and will definitely be called for work; iii. Working for at least 6 hours a day or at least 20 days a month for the usual occupation done every month; iv. Contract workers in the government sector; v. Individuals who receive regular and periodic allowances every month; and vi. Volunteers who receive fixed allowance. Wage rates information consisting of basic wages, cost-of-living allowances and other guaranteed and regularly paid allowances in cash or in kind and overtime payment. However, it excludes bonuses and gratuities, family allowances and social security payments made by employers. Principal occupation is defined as job with the longest number of hours or highest paid or longest period during the reference week if the person having more than one job.

Descriptive statistics is mainly applied in the paper. Three variables studied in the paper are industry, citizenship and earning. Industry is classified according to the The Malaysian Standard Industrial Classification (MSIC) 2008 based on the International Standard Industrial Classification of all Economic Activities (ISIC) Revision 4. MSIC divisions is done in 21 groups. The analysis

was done on the first 19 groups namely A - Agriculture, forestry and fishing, B - Mining and quarrying, C – Manufacturing, D - Electricity, gas, steam and air conditioning supply, E - Water supply; sewerage, waste management and remediation activities, F – Construction, G - Wholesale and retail trade; repair of motor vehicles and motorcycles, H - Transportation and storage, I - Accommodation and Food service activities J - Information and communication, K - Financial and insurance/takaful activities, L - Real estate activities, M - Professional, scientific and technical activities, N - Administrative and support service activities, O - Public administration and defence; compulsory social security, P – Education, Q - Human health and social work activities, R - Arts, entertainment and recreation and S - Other service activities. Earning refers to the basic salary or wage received per month for a primary job as declared by the respondent. In the analysis, proportion of employees by industry and median earning is combined to describe the earning patterns by industry and citizenship. The median earning is presented as the earning distributions were not normally distributed. In the scatter plot, the trend line is added assuming the distribution follows linear pattern.

3. Result

The sample consists of 458,515 respondents which represented by about 50,000 respondents per year except in 2017. Each year, there were less than 9 percent of non-citizen in the sample. On average, there were 8.3 percent of non-citizen employees in the sample. As indicated by Table 1, 65 percent of the employees in 2010 worked in C – Manufacturing, G - Wholesale and retail trade; repair of motor vehicles and motorcycles, O - Public administration and defence; compulsory social security, P – Education and F – Construction. However, the first four industries were the ones with double digit shares in 2010. The same trend can be seen in other years and can be seen for the whole sample and among the Malaysian citizen employees. On average these five industries constitutes 63 percent of the employees for the duration of study. However, the top five industries with the non-citizen employees are A - Agriculture, forestry and fishing, C – Manufacturing, F – Construction, I - Accommodation and Food service activities and G - Wholesale and retail trade; repair of motor vehicles and motorcycles. These five industries comprised of 90 percent of the employees in 2010 with overall average of 89 percent.

Table 1: Percentage distribution of employees in the sample by industry and year.

Industry	2010	2011	2012	2013	2014	2015	2016	2017
A	7%	5%	5%	5%	7%	6%	5%	5%
B	1%	1%	1%	1%	1%	1%	1%	1%
C	20%	20%	19%	18%	19%	18%	18%	18%
D	1%	1%	1%	1%	1%	1%	1%	1%
E	1%	1%	1%	1%	1%	1%	1%	1%
F	8%	9%	9%	8%	9%	9%	8%	8%

G	13%	14%	14%	14%	14%	14%	14%	14%
H	5%	5%	5%	5%	5%	5%	5%	5%
I	6%	6%	6%	6%	7%	7%	7%	7%
J	1%	1%	1%	1%	1%	1%	1%	1%
K	3%	3%	2%	2%	2%	2%	2%	2%
L	0%	0%	0%	0%	0%	0%	0%	1%
M	2%	2%	2%	2%	2%	2%	2%	2%
N	4%	5%	6%	6%	6%	6%	6%	6%
O	13%	12%	11%	12%	10%	10%	10%	10%
P	11%	10%	10%	11%	10%	11%	11%	10%
Q	3%	4%	4%	4%	4%	4%	4%	4%
R	1%	1%	1%	1%	1%	1%	1%	1%
S	1%	1%	1%	1%	1%	1%	1%	1%
Non-citizen	8.5%	7.9%	7.5%	7.3%	9.1%	8.8%	8.4%	8.5%
Sample size	52,066	54,972	54,210	51,668	52,713	51,852	47,625	93,410

Source: the unweighted SWS20010-2017.
Note: Column Percentage.

Figure 3 shows the scatter plot of the median earning and the share of labour for 19 industries in 2010, 2014 and 2017. The black dots represents the Malaysian and the blue dots represent the non-citizen. The first scatter plot shows that most of the employees in the country were Malaysian and they were generally earns more as compared with the non-citizen paid employees in 2010. Comparing the distribution one year after the enforcement of the minimum wage policy, the earning distribution is showing increasing pattern for both type of employees regardless of industry.

The plot also shows the earnings of non-citizen employees in 2010 in industry J - Information and communication was significantly higher (MYR5,000) as compared with employees in other industries. This group of employees remains as top five earners by industry in other years (MYR5,450 in 2017). The plots also show drastic change in earning of non-citizen employees in L - Real estate activities and M - Professional, scientific and technical activities from median earning of MYR750 and MYR1,600 in 2010 into MYR4,850 and MYR4,480 in 2017.

Among the Malaysian paid employees, the top earners works in M - Professional, scientific and technical activities, O - Public administration and defence; compulsory social security and P – Education. Among the three industries, those in P – Education was the highest earners in all years of observation.

It is also observed that the proportion of citizen version non-citizen employees in A- Agriculture, forestry and fishing is almost equal in 2010, 2014 and 2017. Figure 3 also indicates that their earning is almost identical. Expanding the comparison years into single years from 2010 to 2017 shows that gap in proportion of citizen and non-citizen employees in this industry did widening in between 2011 and 2013 where there ratio was 0.52:0.48 in 2011; 0.51:0.49 in 2012 and 0.62:0.38 in 2013. During this period, the earning

remains below MYR1,000 for both group of employees.

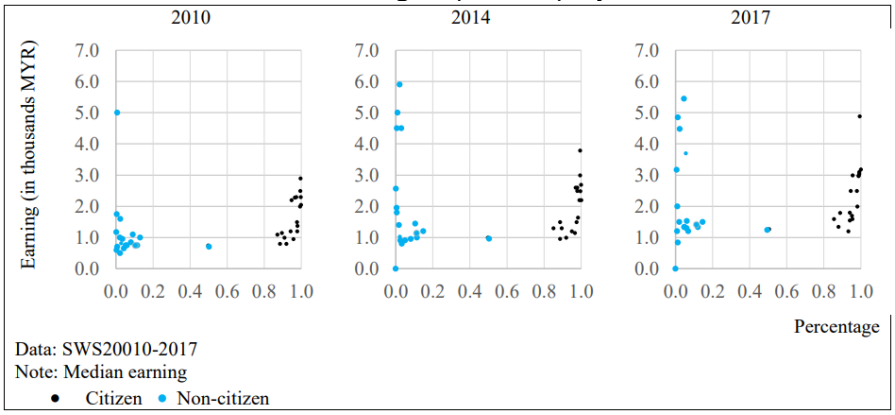


Figure 3: The distribution of earning with share of labor by citizenship by selected years

The detailed of the distribution in figure 3 among the top three industries by earning listed earlier is shown in Figure 4 and Figure 5. The scatter plots in Figure 4 represents the percentage-earnings for the top three industries with the highest earning among the non-citizen employees. Fgure 5 represents the top three industries among the Malaysian citizen employees. The eight dots in each plots correspond to specific years.

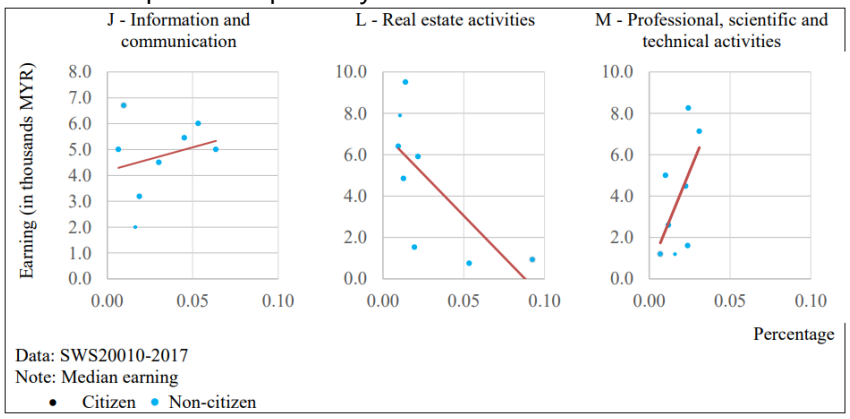


Figure 4: The distribution of earning with share of labor by citizenship by selected years.

The proportion of non-citizen employees in industry J - Information and communication remains around 1-2 percent but increased gradually before remained between 5 -6 percent approaching 2017. With the increase in proportion of employees, the shows an increasing pattern too. In L - Real estate activities, the proportion of non-citizen employees was decreasing over the years but their earning increase significantly since 2013. The same scenarios can also be seen among non-citizen employee who were working in M - Professional, scientific and technical activities.

Among the Malaysian citizen employees, those who were working in P – Education earns the highest followed by O - Public administration and defence;

compulsory social security and M - Professional, scientific and technical activities. At these three industries, almost all of the employees were Malaysian and their earning increases every year.

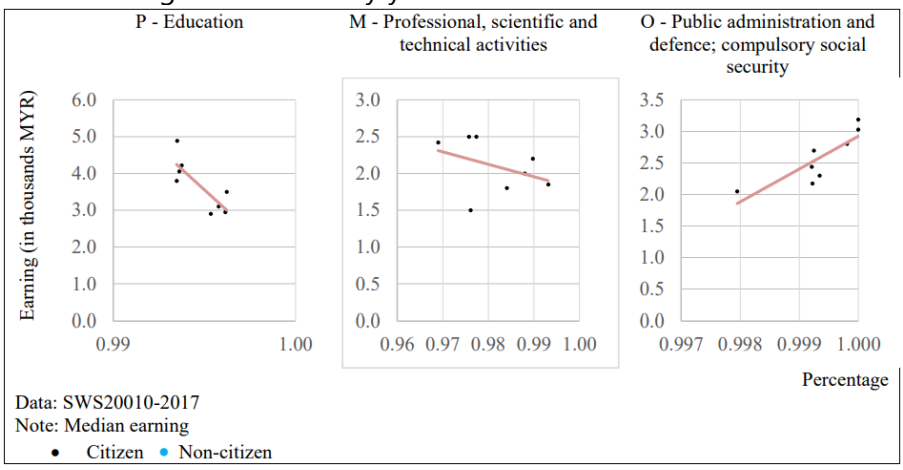


Figure 5: The distribution of earning with share of labor by citizenship by selected years.

4. Discussion and Conclusion

The paper described the distribution of the paid employees in the country based on industry, citizenship and monthly earning of their primary occupation. Based on data collected in 2010 until 2017, on average 63 percent of the citizen employees worked in five industries which comprised of manufacturing sector, construction sector and three industries under other sectors. Almost all of the non-citizen which comprised of less than 9 percent of the sample works in agriculture sector, manufacturing sector, construction sector and another 2 industries under other sectors. This percentage is higher as compared with UK at 12.8 percent in 2017(Rienzo, 2018), US at 14 percent in 2002 (Capps et al., 2003) and Canada at 23.8 percent in 2016 (2017). One the reason could be due the strict definition of employee and the its exclusions. In general the proportion of non-citizen employees were small and their earning is comparatively lower. However in certain industries, these group of employees earns significantly higher as compared with their colleges especially in three other sectors namely, J - Information and communication, L - Real estate activities and M - Professional, scientific and technical activities. The data also shows that the gap of citizen with non-citizen employee and earning in A - Agriculture, forestry and fishing is relatively small. The paper concludes that understanding the paid employees in the country should be done via industry as their earning distribution varies and differentiable by the citizenship. The paper also demonstrates that reducing the industry by four-sector categorization will not gives a reflective conclusion. Further analysis could include the occupation to see how strong the variable is in explaining

variability in earning within the same industry and running the analysis by skill level and differentiating the expatriates and the migrants.

References

1. C, Rienzo. (2018). Migrants in the UK Labour Market: An Overview. Retrieved from https://migrationobservatory.ox.ac.uk/resources/briefings/migrants-in-the-uk-labour-market-an-overview/
2. Howell, M., & Newman, S. (1959). How we should train for overseas posts. The International Executive, 1(1), 21-22.
3. Huber, P., Landesmann, M., Robinson, C., & Stehrer, R. (2010). Migrants' skills and productivity: A European perspective. National Institute Economic Review, 213.
4. J.J. Jordaan. (2018). Foreign workers and productivity in an emerging economy: The case of Malaysia. Review of Development Economics. 22(1) pp:148-173.
5. Malaysia. (2016). Minimum Wages Order 2016. Retrieved from http://minimumwages.mohr.gov.my
6. Mandell, M. M. (1958). Selecting Americans for overseas assignments. Personnel Administration, 21(6), 25-30.
7. Nur`ain Achim, Syezreen Dalina Rusdi and siti noratikah mohamad amin. (2017). The Employment of Foreign Workers: Issues and Implications towards Organization Performance. Conference: International Business Management Conference (IBMC 2017).
8. R. Capps, M. Fix, J.S.Passel, J. Ost, D. Perez-Lopez.(2003). A profile of the low-wage immigrant workforce. Immigrant families and workers facts and persepctive. Brief no 4.
9. Ramesh Kumar Moona Haji Mohamed, Charles Ramendran SPR and P. Yacob. (2012). The Impact of Employment of Foreign Workers: Local Employability and Trade Union Roles in Malaysia. International Journal of Academic Research in Business and Social Sciences October 2012, 2(10). ISSN: 2222-6990
10. Schollhammer, H. (1969). Compensation of international executives. Michigan State University Business Topics, 17(1), 19-30.
11. Statistics Canada. (2017). Labour in Canada: Key results from the 2016 Census. The daily.
12. Thompson, D. (1959). Contracts for Americans working abroad. The International Executive, 1(1), 19-20.
13. Wallace, W. (1959). How to maintain productive working relationships with overseas managers. The International Executive, 1(2), 17-18.

Identifying survey location using GIS: A case study on Malaysia employment survey in oil palm plantations, 2018

Mahdir Bahar, Adrian Austin Spiji, Norhayati Jantan
Department of Statistics Malaysia

Abstract

Malaysia Employment Survey in Oil Palm Plantations, 2018 covered both workers in plantations and smallholdings. Samples were selected through the list of plantations and smallholdings provided by the Ministry of Primary Industries. Due to the classification of the household frame in Malaysia Statistical Household Register (MSAR) which use enumeration blocks (EBs) rather than locality, information such as the names of the selected plantations and smallholdings were either in duplication or unavailable. This paper elaborates the role of Geographical Information System (GIS) in identifying the survey location to enable the sample of workers in oil palm plantations or smallholdings to be drawn.

Keywords

Enumeration Block; Plantations; Smallholdings; Land Title Registration

1. Introduction

MSAR is a database used as a sampling frame for household surveys conducted by the Department of Statistics, Malaysia. It comprised of information on the living quarters by EBs. Each EB on average contains 80 to 120 living quarters. In most household surveys, samples were drawn by two stage stratified sampling in which the first stage was the EBs. For each selected EB, the samples of living quarters (LQs) will be selected using the systematic sampling. However, for the Malaysia Employment Survey in Oil Palm Plantation, 2018, the statistical unit for the survey were oil palm plantation workers and not targeting the households. Due to duplication and availability issue, other methods were being considered in identifying the location of the selected plantations and smallholdings in order for the enumerators to carry out the survey. The samples of workers in this sector can only be selected after the locations are identified.

In this survey, EBs are divided into two categories for plantations which are Main EBs (MEB) and Paired EBs (PEB). First, GIS will locate the MEB and all adjacent EBs will be identified as PEB. This PEBs are required in this survey with the assumption that the plantation workers are residing in the living quarters within the same EB or neighbouring EBs. However, for sample of smallholders only EBs contained the smallholdings will be identified.

One of the considered method can be used is using GIS since many studies found that GIS is very helpful in site location determination. GIS is a computer-based system that aids in collection, maintenance, storage, analysis, and distribution of spatial (geographical) data and information (Cheng, Li, & Yu, 2007; Roig-Tierno, Baviera-Puig, Buitrago-Vera, & Mas-Verdu, 2013). It is suggested that the multi-layer maps and visualized spatial and non-spatial data can be used to find the optimal solution for the location identification problems (Cheng et al. (2007)). In addition to that, GIS-based model has been used to find solutions for minimum distance, maximum demand coverage, maximum income coverage, and optimal centre to determine a suitable location for a shopping mall. In a more recent study, Mardouki and Kordzadeh (2016) concluded that GIS are able to locate a suitable location for a children-specific store in Bannock County, Idaho. The essential parts of GIS implementation to plantation management shall include mapping where accurate verification of estate boundaries can be made, typically revealing discrepancies and areas occupied by others (David Miller, 2009).

The identification of survey locations using GIS depends heavily on the land titles of the plantations or smallholdings which was issued through land registration. Land registration is the process of official recording of rights in the land through deeds or title (on properties). It means that there is an official record (the land register) of right on land or deed concerning changes in legal situation or defined units in land or deed concerning changes in the legal situation or defined units of land. It gives an answer to the question "who" and "how" (Jaap Zevenbergen, 2004). Another similar term used with land registration is "cadastre" which is described as "an official record of information about land parcels, including details of their bounds, tenure, use and value" (McLaughlin/Nichols, 1989). A Land Title shall be deemed to alienate only the land within the boundaries as marked on the ground at the time of the survey on which the title is based (State of Sabah Land Ordinance, 1968). Registered land in Sabah can easily be traced by referring to the digital maps provided by Sabah Land and Survey Department in their website. Digital maps are maps that are in softcopy format and are stored electronically in computer files. Examples of these satellite imageries, digital photographic imageries, digitized or scanned analogue maps (Ayeni,O.O. & O.S. Adewale, 2002).

Matching processes between the registered lands with the EBs in MSAR will be carried out by the application of GIS Analysis Techniques which involves proximity analysis and overlaying analysis techniques. In a case study conducted in a tea plantation area, proximity analysis and overlaying analysis techniques were used to identify the areas which need protection (N N K Wellala, et al., 2012).

2. Methodology

The processes involved in identifying sample locations for this survey are as follows:

i. Preparation of the following lists

a. List of selected samples of plantations and smallholdings

Bil	Code	Smallholder	License No.	Address	Postcode	City	State	Status	Ownership	Smallholdings Location					
										Title No.	Lot No.	State	District	Mukim	Plantation Area (ha)
1	SH0055156	TALINGO BIN MANDASA	275310201000	KG. RUMIDI BARUPETI SURAT 167	90107	BELURAN	SABAH	AKTIF	SENDIRI	LA	PT89082483	SABAH	BELURAN / LABUK & SUGUT	LABUK/SUGUT	6.07
2	SH0055422	MUNANDAR B TAHIR	275807401000	NO.664, RICH PARK,LORONG IDAMAN,3, JALAN PANTAI	91000	TAWAU	SABAH	AKTIF	SENDIRI	CL	CL103042799	SABAH	TAWAU	TAWAU	4.09
3	SH0055689	JAWINAH BINTI DANSOL	276342601000	D/A IBRAHIM BIN MASDIK JAB. PERTANIAN TELUPID, PETI SURAT 10	89300	TELUPID	SABAH	AKTIF	SENDIRI	LA	PT73081263 KIABAU	SABAH	BELURAN / LABUK & SUGUT	LABUK/SUGUT	1.62
4	SH0055956	U TECK ONG	276805301000	W. D. T 429	90009	SANDAKAN	SABAH	AKTIF	SENDIRI	CL	CL20129 SANDALA	SABAH	SANDAKAN	SANDAKAN	2.91
5	SH0056223	WONG BOON KIM	277337501000	P.O.BOX NO.60222	91111	LAHAD DATU	SABAH	AKTIF	SENDIRI	CL.115326773	CL.115326773	SABAH	TAWAU	LAHAD DATU	5.94
6	SH0056489	YONG LEN NYUK	277827001000	W. D. T. 248	90009	SANDAKAN	SABAH	AKTIF	SENDIRI	LA	CL075350597 BT. 20	SABAH	SANDAKAN	SANDAKAN	3.11
7	SH0056756	LIEW NYUK MIN	278338901000	PETI SURAT 61651	91026	TAWAU	SABAH	AKTIF	SENDIRI	CL	CL105365535	SABAH	TAWAU	TAWAU	18.57
8	SH0057023	SHO CHE WAN	278847001000	PETI SURAT 46	90107	BELURAN	SABAH	AKTIF	SENDIRI	CL	CL085320190	SABAH	BELURAN / LABUK & SUGU	LABUK/SUGUT	4.99
9	SH0057289	ABU KASSIM BIN PABUN	279396101000	D/A. ABD. JAYA BIN GEDDEK PEJABAT PERTANIAN, PETI SURAT NO.3	91207	KUNAK	SABAH	AKTIF	SENDIRI	PT2012120043	PT2012120043	SABAH	TAWAU	SEMPORNA	6.07
10	SH0057556	SAKAYAN BIN ALI	279993501000	NO. 58, KG. MUHIBBAH, PERINGKAT 3	91100	LAHAD DATU	SABAH	AKTIF	TANAH NT (NATIVE TITLE)	NT.113044401	NT.113044401	SABAH	TAWAU	LAHAD DATU	6.07

Figure 1: Examples of Selected Plantations' List

Bil	Code	Plantations	License No.	Category	Address		Postcode	City	State	District	Plantation Area (ha)
1	OILPALM1771	FELCRA BHD KAW. SUNGAI YEH YEH	503581-602000	FELCRA	MDLD 6968, BANDAR SRI PERDANA,	PETI SURAT 61644,	91100	LAHAD DATU,	SABAH	SEMPORNA	102.06
2	OILPALM1801	FELCRA BHD KOP. SUNGAI MATANGGAR	503627-802000	FELCRA	LOT NO. 6, BGN SEDCO LIGHT INDUSTRIES,	PETI SURAT 487,	90107	BELURAN,	SABAH	LABUK/ SUGUT	184.69
3	OILPALM3920	FELCRA BHD KAW. KG. BAMBANGAN	568865-002000	FELCRA	LOT NO. 6, BGN SEDCO LIGHT INDUSTRIES,	PETI SURAT 487,	90107	BELURAN,	SABAH	SANDAKAN	90
4	OILPALM4679	LADANG MESEJ KG. TUKAR	608432-002000	Agensi Kerajaan	KG. TUKAR MESEJ		88100	PITAS	SABAH	PITAS	98
5	OILPALM0380	BINGKOR PROJECT	501438-002000	Agensi Kerajaan	P. O. BOX 2504,		89008	KENINGAU,	SABAH	KENINGAU	353
6	OILPALM1260	LADANG PINAWANTAI	502811-902000	Agensi Kerajaan	PETI SURAT 154		89108	KOTA MARUDU	SABAH	KOTA MARUDU	864.39
7	OILPALM1751	FELCRA BHD KAW. KUDAT	503550-602000	FELCRA	LOT NO.18, TINGKAT 1,TAMAN WTK, FASA 2,	PETI SURAT 256	89108	KOTA MARUDU,	SABAH	KUDAT	607.2
8	OILPALM2294	FELCRA BERHAD KG. KIABAU	504423-802000	FELCRA	PETI SURAT NO. 78	PEJABAT POS TELUPID, 89320 TELUPID, SABAH	89320	TELUPID	SABAH	LABUK/ SUGUT	612
9	OILPALM3125	FELCRA BHD. KAW. SIPITANG	528550-002000	FELCRA	PETI SURAT 671		89808	BEAUFORT	SABAH	SIPITANG	313.6
10	OILPALM4827	SKIM NABAWAN	615507-102000	Agensi Kerajaan	L.K.T.N.S - SKIM NABAWAN, SLDB REGIONAL	KM 6, JALAN APIN-APIN, P.O.BOX 2504,	89009	KENINGAU	SABAH	PENSIANGAN	741.41

Figure 2: Examples of Selected Smallholders' List

b. List of EBs in MSAR; and

c. List of land title registration numbers by plantations and smallholders from Malaysian Palm Oil Board (MPOB).

ii. Based on the list of selected sample of plantations and smallholders, the land title registration numbers were used as the main reference to identify the survey locations. The coordinate of the plantations and smallholdings

are obtained from the Sabah Lands and Surveys Department website (http://www.jtuwma.net/) using the search box provided.

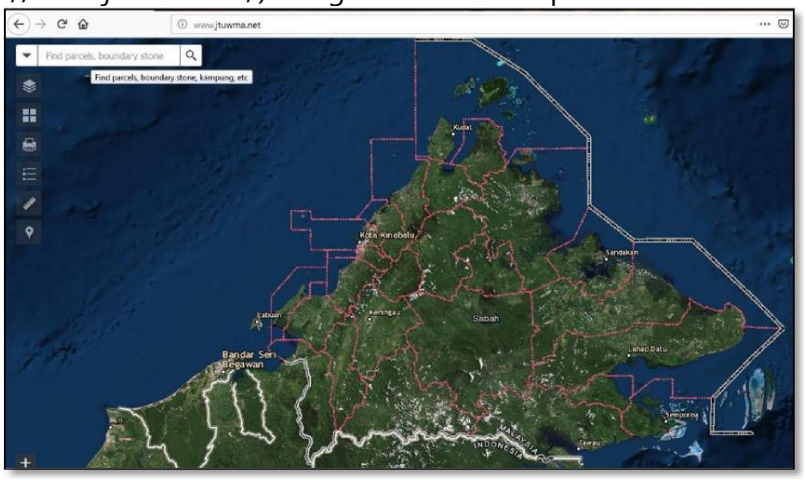


Figure 3: Website http://www.jtuwma.net/

iii. Using the coordinates, The Google Earth map can be layered with the EBs polygon which is created through GIS in KML format.

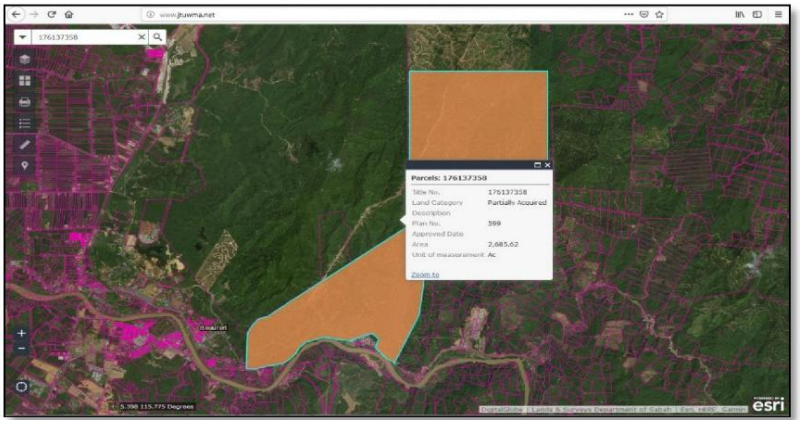


Figure 4: Website http://www.jtuwma.net/

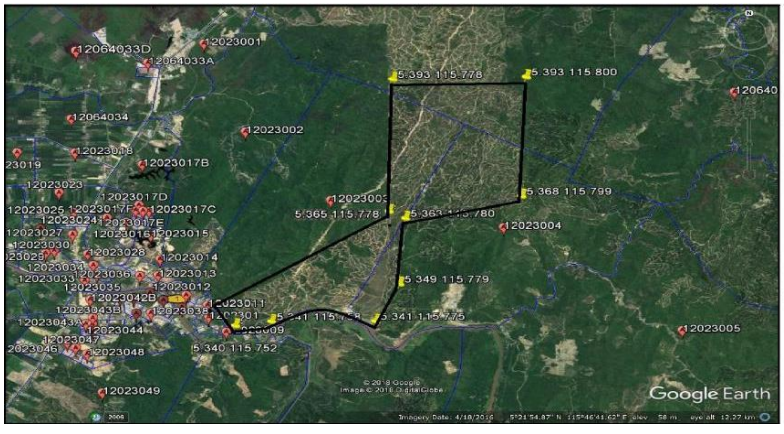


Figure 5: Google Earth (KML format)

iv. or sample of plantations, the EBs which touches the boundary of the particular land title is listed as a MEB. All adjacent EBs will be identified as potential PEB. Only one PEB will be chosen using simple random sampling method for each MEB.

v. For Smallholders, the EBs which touches the boundary of the particular land title is listed as selected EBs.

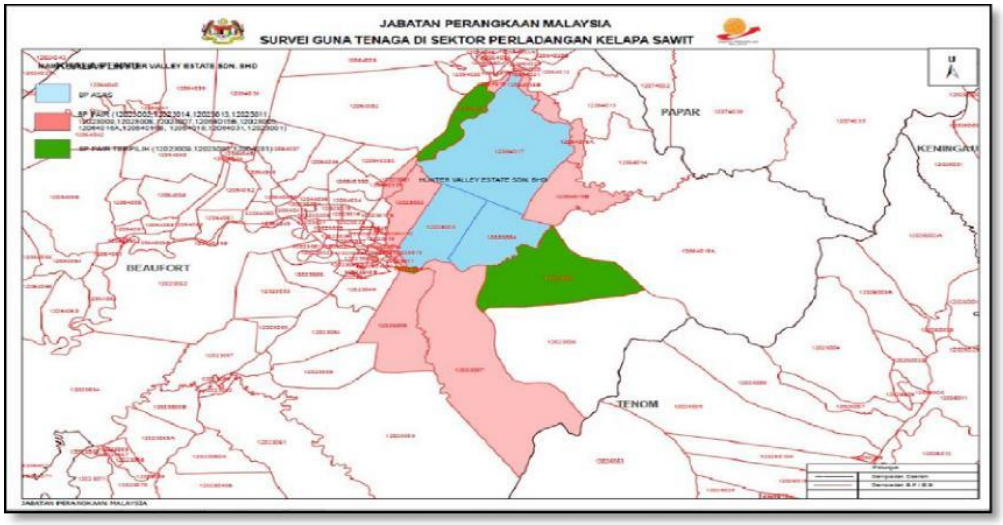


Figure 6: EBs boundaries in GIS output

In this example in Figure 6, the selected plantation falls in three (3) EBs which in blue colour (MEB). While the possible PEBs which are the adjacent EBs are in pink colour. Out of eight (8) possible PEBs, three (3) EBs are selected as PEBs which are in green colour. The process above will be repeated for all samples of plantations and smallholders that consist of land title registration numbers.

3. Result

Using this method, a total of 134 MEBs were identified with 127 PEBs for all 65 selected plantations. The number of PEBs less than the number of MEBs as a result of some MEBs are also PEBs to other selected plantations. A total of 261 EBs were covered to represent the population in oil palm plantations. GIS is able to identify the location of oil palm plantations provided that the list of selected plantations or smallholdings had the land title registration number. A total of 261 EBs (MEB and PEB) are covered for oil palm plantations in this survey.

Total sample of Plantations	Total Plantations Consist of Land Title Registration Number	Total of MEB	Total of PEB	Total EB
65	65	134	127	261

Table 1: The number of selected EBs for Oil Palm Plantations

For sample of smallholders, out of 143 smallholders only 84 EBs can be identified based on the land category. Meanwhile, 75 smallholders cannot be matched through the search application in JTUWMA (GIS). For unidentified smallholders, another method has been used as an alternative solution to identify the locality of the smallholdings such as reference to the Land Office.

Land Category	Total	BP Identified
Native Title	41	53
Commercial Lease	24	28
Field Register	3	3
Land Application	69	-
Lot	6	-
Total Sample of Small Holders	143	84

Table 2: The number of selected EBs for Smallholdings Identified using GIS

4. Discussion and Conclusion

Based on this study, it was found that the usage of GIS is helpful in identifying locality of survey location. At the same time, GIS also can be used as tools for effective planning in the other aspects of survey operation such as time management, logistic and number of enumerators. In this survey, GIS also facilitate in the identification of population settlement in the selected EBs to enable the sampling of plantation's workers using Adaptive Cluster Sampling method. Besides that, the model map for survey areas for the purpose of survey operation can also be obtained using the spatial analysis through GIS and Google Earth.

The success of the matching process in identifying the correct EBs are highly depends on the accurateness of the spatial and attributes data in the GIS database of the Sabah Land and Survey Department. Any discrepancies will lead to the inaccurate matching process and will produce wrong MEBs and PEBs. The whole process has to be repeated once the wrong MEBs and PEBs are identified for samples of plantations dan selected EBs for smallholders. The conventional methods are used to identify and locate the EBs for smallholdings without a land title registration number. This can be done but it will take a longer time in the identification stages. It is suggested that for similar study, the sampling frame should contain the complete registration information on the land plot.

GIS is a powerful set of tools in current perspective because GIS has the capabilities to provide information about a location. GIS technology helps in simplifying decision making based on its analysis capabilities with spatial data. It is an application developed to provide information through spatial visualization framework that can assist and support decisions taken in the management and use of man-made resources and environment.

Currently, there is a rapid development in geographic information system. The use of Google Earth in this study proved that GIS is able to identify the locality of the samples more accurate and faster compared to conventional method. In addition, the use of this GIS information can also be developed in other application such as a tracking device to facilitate the conduct of survey operation.

References

1. Cheng, E. W., Li, H., & Yu, L. (2007), A GIS approach to shopping mall location selection. Building and Environment, 42(2), 884-892.
2. Mardouki, Arina and Kordzadeh, Nima, "Site Location Determination Using Geographic Information Systems: The Process and a Case Study" (2016). SAIS 2016 Proceedings, 4.
3. Roig-Tierno, N., Baviera-Puig, A., Buitrago-Vera, J., & Mas-Verdu, F. (2013). The retail site location decision process using GIS and the analytical hierarchy process. Applied Geography, 40, 191-198.
4. Jaap Zevenbergen (2004). A Systems Approach to Land Registration and Cadastre. Nordic Journal of Surveying and Real Estate Research.
5. McLaughlin, J. and S. Nichols (1989), Resource Management: The Land Administration and Cadastral Systems Component, In: Surveying and Mapping, No. 2, p. 77-86.
6. State of Sabah Land Ordinance (Sabah Cap.68), p.14
7. David Miller. 2009. Implementation of GIS to Palm Oil Palnatation Management in Indonesia.
8. Ayeni,O.O. & O.S. Adewale (2002), GIS Queries for Population Data Analysis and Management.
9. N N K Wellala, J Gunatilake and H W Shyamalie (2012), Use of Geographic Information System in Tea Plantation Management: A Case Study at St. Coombs Estate, Talawakelle. Sri Lanka J. Tea Science Volume 77, Part (1/2); 70-82

Summary of bio-statistical consultation for clinical research support in Japan

Makoto Tomita
School of Data Science, Yokohama City University, Japan.

Abstract

Changes in the environment surrounding recent clinical research are obvious, and support in the field of statistical analysis is also diverse. We aim to feedback statistical consulting supported by our center by grasping the situation and introduce them.

Keywords

Clinical research support; statistical consulting; biostatistics; National University Hospital; academic research organization (ARO)

1. Introduction

University conducts research in various areas and has the potential of developing new drugs by nurturing the seeds of research. We will foster alliance network as an ARO (academic research organization) by sharing resources and experience that becomes difficult for one member university, thereby step up as an organization that will contribute to new drug development. In recent years, evidence-baced medicine (EBM) seems to have completely established even in the field of clinical research. Evidence in clinical research means that it is required to be scientifically the most reliable, and based on that judgment statistical plays a decisive and important role. It receives about 50 to 80 statistical consultations per year, it tends to increase with the age and it realizes that it is useful as a biostatistician.

The Ministry of Education, Culture, Sports, Science and Technology's Ministry of Health, Labor and Welfare established the "ethical guidelines on research on medical science for human beings" in 2014, the environment surrounding clinical data became more rigorous, data managers (DM) More and more, the presence is increasing.

A biostatistics who was assigned to the Center since 2009 has initially launched a web page of statistical consulting in June and gave a guide in the hospital by lamppost, etc. The service gradually became widely publicized and spreading in the word of mouth. I felt as if there were many repeaters from the same field, I thought that I would like to compile the situation from FY 2010 again and grasp the trend and transition.

Furthermore, at our university, the "Statistical Analysis Support Unit" targeting clinical researchers at all universities in the Center for Medical

Innovation Promotion has been launched to promote clinical research support so that more convenience for clinical researchers will be increased. Then, we summarize to feedback statistical consulting supported by our center by grasping the situation and introduce them.

2. Methodology

For statistical counseling engaged between 2009 and 2017, visualize the contents, performance and statistical methods for each department/research field using graphs.

3. Result

Abe (2015) is the Biostatistics Office of the Clinical Research Promotion Center, Keio University School of Medicine Hospital Clinical Research Promotion Center, and the number of statistics counseled and the number of the same thesis is shown in Figure 1. Meanwhile, the center compiled the number of statistical consultations from fiscal 2010 to the present and the number of papers published as a co - author by it (Fig. 2). Although both have an increase and decrease, it seems that the number of papers is also linked.

Furthermore, we focused on our Clinical Research Promotion Center and compiled the number of articles via statistical consulting for each department, department, research field, center, and show the transition. As seen in Figure 3 and 4, many clinics and departments are enumerated, but it was found that the proportion of intensive care unit and ophthalmology is particularly high. In addition, clinical research exceeded 80%, of which intervention studies accounted for 30%. On the other hand, Figure 5 shows the trends in cumulative number of hits every fiscal year from fiscal 2010. It can be seen that the department of medicine and other departments which received statistical consulting from early stage are gradually increasing.

4. Discussion and Conclusion

According to the department of medical examination/research field, there was bias, and it was found that the statistical consulting worked by the ophthalmology, respiratory department internal department, intensive care department and the results published as co-author are remarkable. In addition, statistical analysis on surgery/regenerative medicine research and development and practical use of medical care are also increasing in recent years. From 2017, we will also work as a supporter at the Statistical Analysis Support Division of the Medical Innovation Promotion Center throughout the campus as well as the Hospital affiliated with the Faculty of Medicine, and promote clinical research support so that more convenience for clinical researchers will be increased.

The number of consultations in statistical analysis and the number of research papers have shown a tendency to monotonically increase in the right direction, and it is expected that the likelihood of decreasing will be low in the future. An increase in the number of sudden consultations is thought to be comparatively suppressed due to the establishment of the statistical analysis support department, and it is vital that future efforts to efficiently and promptly promote efforts while grasping the trends and changes to date is there.

Statistical consulting was conducted at the university other than the Center for Clinical Research Promotion, but since this fiscal year the window has been integrated as a statistical analysis support unit, the environment for researchers is easy to inquire and the convenience has improved. In the same unit, the Web page was also maintained and provided lectures that can be taken free of charge in the university - statistical analysis of data, design of the number of cases, clinical research design - and a reception system was also introduced, which lecture was packed quickly It has become a great success.

When we compiled statistical consulting conducted at our Center this time, we found out that we are again receiving statistical counseling from many medical departments / research fields / centers. As each graph also shows, it is predicted that it will increase even in the near future, and we want to push forward as a biostatist.

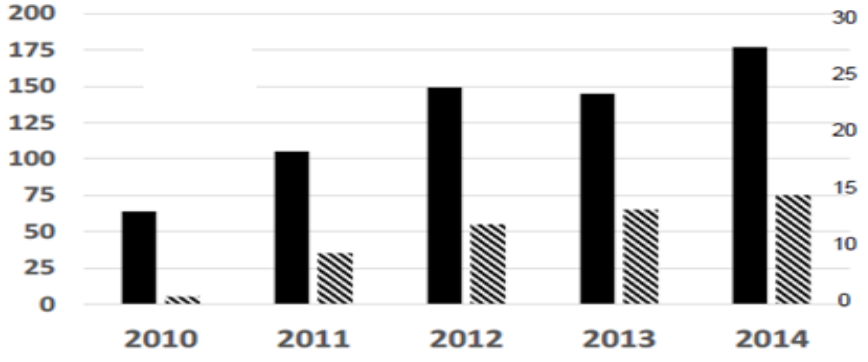


Fig 1. Keio University School of Medicine Hospital Clinical Research Promotion Center Statistical consultation statistics consultation number (black paint) in the biostatistics room and trends in the number of articles (shaded) published by the corresponding statistician as coauthor. (Abe, 2015)

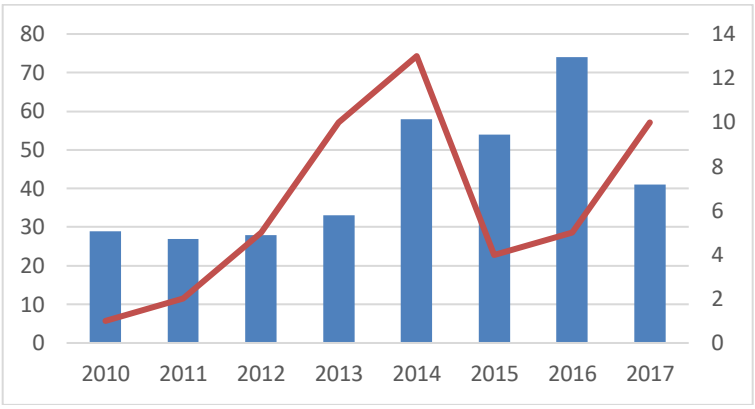


Fig 2. Statistics counseling statistics (bar graph) of statistical consulting at Tokyo Medical and Dental University Medical College Hospital clinical trial management center and transition of the number of papers published as co-authors by the corresponding statistician (line graph). (Tomita, 2017)

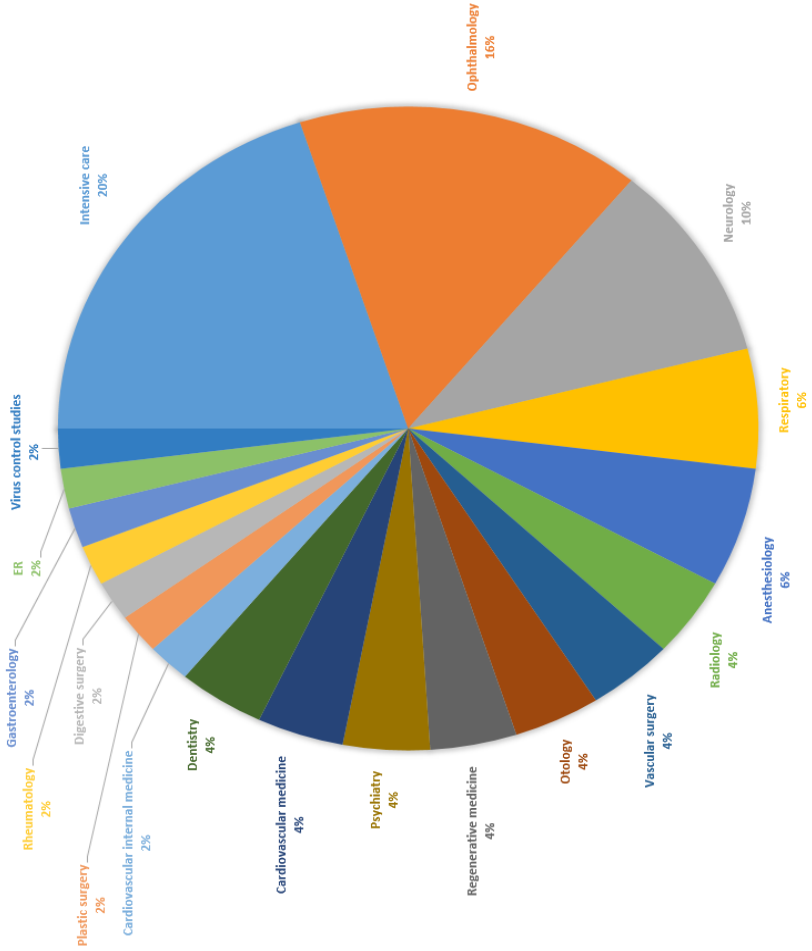


Fig 3. Percentages of the number of articles (2010 - 2017) published by statistical consulting at the clinical trial management center of Tokyo Medical and Dental University Medical College Hospital by medical department, research field, center.

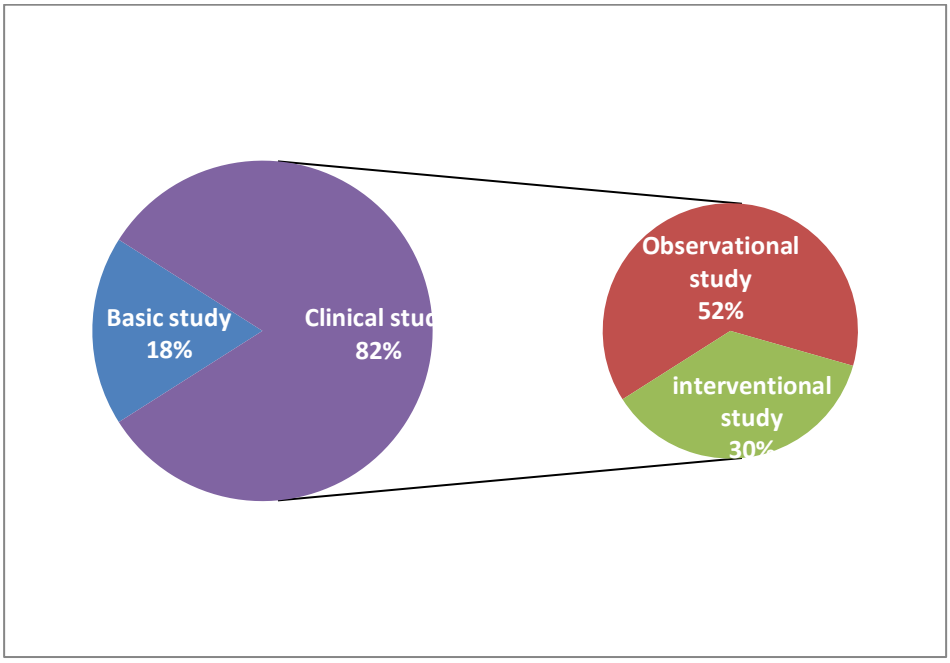


Fig 4. Percentage by basic research / clinical research of the number of articles (2010 - 2017) posted through statistical consultation at clinical trial management center of Tokyo Medical and Dental University hospital.

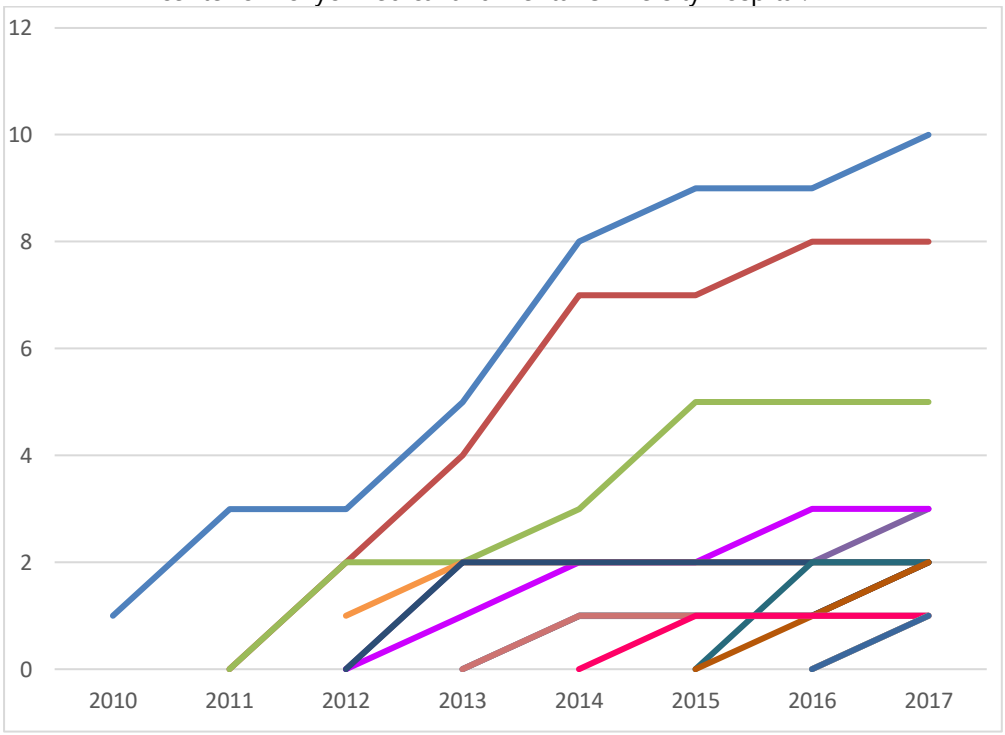


Fig 5. Changes in the cumulative number of papers by medical department, research field, center in papers (2010 - 2017) posted through statistical consulting

References

1. Tomita M. (2016) Current situations on clinical research support and the role of biostatisticians in Japan, Proceedings of the 2016 International Conference for JSCS 30th Anniversary in Seattle, 81-82, Seattle, WA.
2. Tomita M. (2017) Topics on clinical research and clinical trial efforts at "University Hospital Clinical Trials Alliance" and " National University Hospitals Clinical Research Promotion Initiative", Proceedings of Japanese Joint Statistical Meeting 2016, Kanazawa, Japan. (In Japanese)
3. Tomita M. (2017) Visualization and feedback of statistical consultation trends and results at clinical research center of Tokyo Medical and Dental University, Proceedings of 38th Annual Scientific Meeting of The Japanese Society of Clinical Pharmacology and Therapeutics, Yokohama, Japan. (In Japanese)
4. Tomita M. (2017) Real situations on clinical research support and the role of biostatisticians in Japan, Scientific Programme - 61st ISI World Statictics Congress, p. 25 (PO-A07), Marrakech, Morocco.

Improving paddy rice statistics using area sample frame technique

Anna Christine Durante ,Pamela Lapitan, David Megill, Lakshman Nagraj Rao
Asian Development Bank

Abstract

Agricultural surveys typically use a list frame as the main sampling frame to identify and access the elements of the target population. List frames are built on the basis of information from administrative records and data from recently conducted censuses. While the efficiency of survey implementation may be affected by the overall survey process, frame quality has a major effect on the quality of statistics produced. In most developing countries, a complete frame with updated and comprehensive holding information may not be available, leading to under-coverage and biased estimates. To circumvent these issues, the Asian Development Bank evaluated the usefulness of an area sampling frame developed using technological advances in satellite imagery and geographic information system techniques for estimating paddy rice area and production in a study that was piloted in three major rice producing provinces in Southeast Asia—Savannakhet, Lao PDR; Ang Thong, Thailand; and Thai Binh, Viet Nam. The study employed a three-stage stratified sampling technique for which the sampling frame used was developed using freely available MODIS and Landsat data on land cover and land use. Direct estimates of total paddy rice area and production are calculated from area frame using two methods––one involving measurement of plot size using a global positioning system instrument and the other utilizing a digitized map of farmer-identified plot boundaries on a high- resolution Google Earth image. A third method involving the calculation of ratio estimates using independent mesh-level measures is compared with the first two methods involving direct estimates, and with the estimates generated from administrative data from the countries. Our study finds that ratio estimation significantly improves the level of precision of paddy rice statistics. Significant deviations for paddy rice area, production, and yield are also observed between official statistics and the statistics generated through direct estimation. Nonetheless, the estimates are likely more reliable when compared to official estimates since satellite-based estimation methodology is transparent, reduces under-coverage and allowed for calculation of confidence interval. It is also necessary to improve the land-use stratification of the frame by using higher-resolution satellite images and a greater power of discrimination in the models used for defining the strata. Sentinel 1 and Sentinel 2, which have better resolution than existing freely

available satellite images at the time of the study, are likely candidates for future research.

Keywords
Agriculture; Area Frames; Sampling Methods; Satellite Images

1. Introduction

Timely and reliable agricultural statistics are critical for monitoring government agricultural development plans and mitigating the effects of extreme weather and climate change. The preparation of national accounts, evaluation of agricultural interventions, and the development of early warning systems to address climatic and non-climatic vulnerabilities in the agricultural sector, rely on high- quality and disaggregated agricultural data. In the absence of good quality data, inefficient allocation of resources is likely which would lead to a failure in resolving critical development problems (Kelly et al. 1995).

The compilation of official agricultural statistics relies on data collected using administrative records or probability-based field surveys. The advantages of these methods lie in its lower implementation cost, but estimates derived are likely to be biased and prone to large measurement errors. Household and/or agricultural surveys can provide better estimates, when objectively designed and conducted. Both agricultural and population-based census frames are commonly used in developing countries as a basis for designing multistage sample agricultural and household surveys (Grosh and Munoz 1996). However, in some countries, a complete frame is not available if the reference is a census with low coverage, or the existing lists of sampling units change rapidly rendering the list frame out of date (Griffin 2014). Field listing activities may not be accurate if households systematically over or under reporting agricultural holdings.

An alternative to the list frame approach is the area frame approach. In the area frame approach, the final stage sampling units are land areas and the selection probabilities are proportional to their area measures. A multistage stratified approach can then be implemented based on an area frame to select a sample of grids within each stratum of land cover and/or land use, depending on the survey objective (Faulkenberry and Garoui 1991).

To fill in the gap in the existing literature, this study utilizes an area frame approach through the innovative combination of satellite data, GIS methods, and crop cutting to estimate paddy rice area, yield, and production for the 2015 rainy season in selected provinces of three pilot countries—Lao People's Democratic Republic (Lao PDR), Thailand, and Viet Nam[1] —and compares

[1] The pilot provinces include: Savannakhet (Lao PDR), Ang Thong (Thailand), and Thai Binh (Viet Nam).

them to estimates obtained from existing administrative data sources. The pilot provinces were stratified into rice growing areas using satellite data and GIS methods and within each stratum, square meshes were randomly selected to identify plots eligible for crop cutting. Crop cutting was then implemented in randomly selected subplots to obtain unbiased rice yield estimates. This allowed for the calculation of both direct and indirect estimates of total paddy rice area.

2. Data Description and Methodology

An area frame was used for this study and constructed based on the expected likelihood of finding paddy rice area in each grid square mesh. Two sources of rice maps were utilized to implement the stratification process: (i) rice extent maps using 2015 MODIS data produced by the International Rice Research Institute (IRRI)[2] and (ii) land use maps from 2009 produced by the European Space Agency (ESA) under its GLOBCOVER initiative.[3] These two sources allow for identification of land most recently used for growing rice alongside providing information on those areas which are repeatedly used for rice cultivation. The primary sampling unit (PSU) in this study is a 200 m x 200 m square "mesh" [4]which is spatially defined on a digitized satellite image map (Figure 1).

Figure 1: Sample 200m x 200m Mesh

[2] IRRI has been developing remote sensing-based maps of rice systems in Asia as part of its contribution to various projects that need good baseline data on rice (http://irri.org/our-work/research/policy-and-markets/mapping/remote-sensing-derived-rice-maps-and-related-publications).

[3] GlobCover is an ESA initiative which began in 2005 in partnership with the Joint Research Center (JRC of the European Commiss ion), United Nations Environment Programme, Food and Agriculture Organization, and other institutions. The aim of the project was to develop a service capable of delivering global composites and land cover maps using as input observations from a sensor onboard the Environmental Satellite (ENVISAT) mission (http://due.esrin.esa.int/page_globcover.php).

[4] The choice of 200 m x 200 m mesh is based on pixel size of the satellite images used in the study.

The stratification in this study was conducted prior to the selection of meshes to improve statistical efficiency and lower fieldwork costs. The first stratum consisted of meshes that both IRRI and ESA maps (*IRRI+GlobCover* stratum) identified as paddy rice area, considered to be the most likely to contain paddy rice. The second stratum, considered a medium probability stratum[5], consisted of meshes that were only identified as rice by the IRRI area map (*IRRI* stratum) but not by the ESA map.

The third stratum is the low probability stratum, identified as rice by ESA's map (*GlobCover* stratum) but not by IRRI's map. The final stratum consists of all remaining areas where presumably no rice is grown as indicated by both IRRI and ESA maps, henceforth referred to as the *Other* stratum.

In the first sampling stage, a stratified random sample of 120 meshes was selected for each pilot province[6]. The number of selected meshes was higher in the stratum where the expected likelihood of finding rice growing plots is highest (Stratum 1), and lower in areas with low (Stratum 3) or no likelihood (Stratum 4) of finding rice growing plots. A ground-truthing field operation was conducted to verify whether rice was planted in any plots within the boundaries of each sample mesh. Only sample meshes with rice were enumerated for eligibility to be selected for crop cutting in the second sampling stage.

Systematic random sampling was then used to select a sample of four plots per mesh from the list of plots that met the selection criterion. This involved calculating a sampling interval, which was used to systematically select the sample plots from the ordered list, following a random start. The selection of four plots was driven by the need to ensure sufficient sample size within a mesh to capture variability in rice yields across plots and budgetary constraints.

At the third sampling stage, a random point was selected within each sample plot to identify a 2.5 m x 2.5 m crop-cutting[7] subplot. This was followed by the measurement of area planted in rice in sample plots within each sample mesh, which was also used as component for the sample weighting procedure. Two sources of objective measurements for the area of the sample rice plots were used: (i) unmodified tracks that are based on the

[5] the resolution of the IRRI map obtained from MODIS is better and more recent than the ESA map obtained from ENVISAT

[6] The total number of meshes was based on the expected number of rice plots to be found and interviewed in each stratum using d ata from pretests and the available budget for the pilot project.

[7] Crop cutting is a method wherein a small portion of a randomly selected plot, henceforth referred to as a subplot, is harvested, threshed, dried, and weighed to obtain objective yield estimates (Huddleston 1978).

boundaries of the plot recorded by the enumerators using a handheld GPS navigation device, and (ii) modified[8] track data where plot boundaries were digitized on Google Earth Pro. A GIS software named QGIS was used for data processing.

3. Results

Two sets of weights were calculated for the sample subplots based on each source of area measurement—unmodified and modified[9] data. These weights were used to produce alternative direct estimates of the total area planted in rice alongside calculating corresponding SE, 95% confidence intervals and design effects.

Table 3: Standard Error, Coefficient of Variation, and Design Effects of Estimates of Total Area Planted in Rice Paddy in m^2 based on Area of Sample Plots from Unmodified Track Data

Domain/Stratum	Area, m^2	SE	CV	95% confidence interval		DEFF	No. of observations
				Lower	Upper		
Savannakhet	2,700,047,316	509,013,893	0.189	1,685,814,995	3,714,279,638	6.25	136
IRRI+GlobCover	1,092,874,253	112,474,296	0.103	868,764,326	1,316,984,181	0.38	105
IRRI	44,413,015	12,865,940	0.29	18,777,070	70,048,961	0.08	16
GlobCover	804,851,923	203,242,057	0.253	399,883,285	1,209,820,561	1.73	13
Other	757,908,125	452,738,125	0.597	0	1,660,008,531	10.98	2
Ang Thong	292,337,345	40,269,296	0.138	210,408,846	374,265,843	1.69	104
IRRI+GlobCover	274,834,018	39,838,268	0.145	193,782,453	355,885,582	1.62	82
IRRI	2,779,591	976,108	0.351	793,684	4,765,498	0.03	8
GlobCover	14,723,736	5,794,467	0.394	2,934,805	26,512,667	0.71	14
Thai Binh	474,230,049	45,807,978	0.097	382,631,336	565,828,761	4.34	256
IRRI+GlobCover	446,207,926	45,109,146	0.101	356,006,614	536,409,237	3.97	220
IRRI	110,039	19,651	0.179	70,744	149,334	0	8
GlobCover	27,912,084	7,970,911	0.286	11,973,262	43,850,906	0.92	28

[8] Field staff were asked to identify all obstructions and draw the correct boundaries of the plot on the printed map which was factored into the digitization process.

[9] Given the greater quality control involved in measuring the modified track area, the values are considered more accurate than the unmodified track data. Hence, the subplot weights based on the digitized area of rice plots inside the mesh are used for the analysis.

Table 4: Standard Error, Coefficient of Variation, and Design Effects of Estimates of Total Area Planted in Rice Paddy in m^2 based on Area of Sample Plots from Modified Track Data

Domain/Stratum	Area, m^2	SE	CV	95% confidence interval		DEFF	No. of observations
				Lower	Upper		
Savannakhet	2,496,699,709	410,394,222	0.164	1,678,971,372	3,314,428,047	4.91	136
IRRI+GlobCover	1,086,335,396	108,016,428	0.099	871,107,964	1,301,562,827	0.37	105
IRRI	44,287,871	13,014,555	0.294	18,355,804	70,219,939	0.08	16
GlobCover	714,442,692	189,790,467	0.266	336,276,931	1,092,608,453	1.98	13
Other	651,633,750	347,226,250	0.533	0	1,343,497,156	9.35	2
Ang Thong	299,114,441	41,045,835	0.137	215,606,062	382,622,821	1.65	104
IRRI+GlobCover	278,414,050	40,259,423	0.145	196,505,638	360,322,462	1.56	82
IRRI	2,567,025	982,144	0.383	568,838	4,565,213	0.04	8
GlobCover	18,133,366	7,935,667	0.438	1,988,130	34,278,603	0.87	14
Thai Binh	484,750,036	45,200,548	0.093	394,365,955	575,134,118	4.89	256
IRRI+GlobCover	456,083,074	44,435,139	0.097	367,229,522	544,936,626	4.36	220
IRRI	109,143	14,156	0.13	80,837	137,450	0	8
GlobCover	28,557,819	8,282,979	0.29	11,994,980	45,120,658	1.02	28

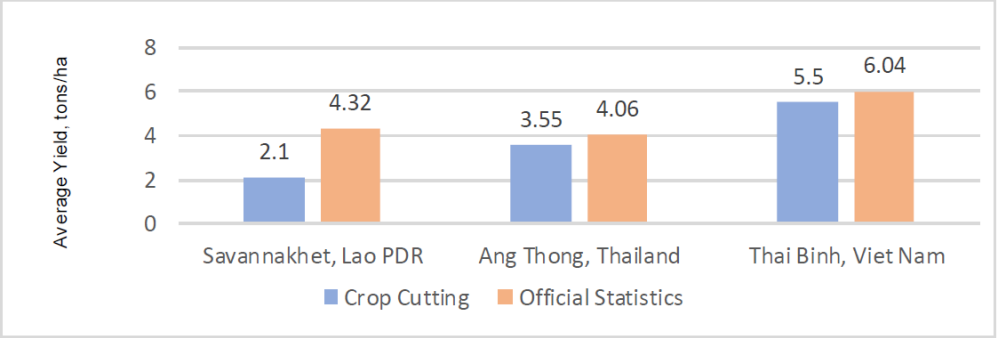
In both Ang Thong and Thai Binh, roughly 93% of the estimate of total rice area is from the *IRRI+GlobCover* stratum. The estimates for this stratum have higher design effects, which are mostly measuring the clustering effects due to the similarity of the plot areas within a mesh. In Savannakhet , only 40% of the estimate of total rice area comes from the *IRRI+GlobCover* stratum while, more than 26% of the weighted area comes from the *Other* stratum, which has a much lower sampling rate and thus, a much higher weight. This indicates a problem with the efficiency of the stratification of the sampling frame, either in discriminating the rice found in this stratum from the satellite images, or because of more recent rice planting activities after the satellite images were generated.

High CV was calculated for all pilot areas which resulted in relatively wide confidence intervals for the estimates for all pilot provinces. One reason for this is the variability in the size of the sample plots selected randomly within each mesh.

Table 5: Estimates of Statistical Parameters for the Unweighted Paddy Area Data by Source

Pilot Area	Source of data	Mean (m^2)	SD	Min value	Max value
Savannakhet	Unmodified track data	5,758.35	5,465.67	52.97	27,488.38
	Digitized track data	5,676.47	5,365.19	7.83	28,260.85
Ang Thong	Unmodified track data	3,444.54	3,442.93	0	17,803.71
	Digitized track data	3,497.89	3,452.85	0	17,086.20
Thai Binh	Unmodified track data	720.02	522.56	39.70	3,190.60
	Digitized track data	729.39	497.79	31.00	2,743.70

Figure 2: Average Yield per Pilot Area in tons/ha



The direct estimation of total production (kg) and the average yield (kg/m^2) of harvested paddy are derived from crop-cutting data for the sample of 2.5 m x 2.5 m (i.e., 6.25 m^2) subplots by multiplying data on harvested paddy in each subplot by the corresponding subplot weights used for the estimation of total area. Table 7 shows relatively high CVs for the direct estimate of the total production of rice, resulting in corresponding wide confidence intervals. The main source of this discrepancy is due to the variability in the estimate of total area planted in rice paddy.

Table 6: Direct estimates of total production of rice paddy in kg

Domain/Stratum	Production, kg	SE	CV	95% confidence interval		DEFF	No. of observations
				Lower	Upper		
Savannakhet	480,210,025	70,689,811	0.147	339,357,502	621,062,547	27.78	136
IRRI+GlobCover	232,891,720	28,133,072	0.121	176,835,350	288,948,089	1.27	105
IRRI	8,612,733	2,300,195	0.267	4,029,494	13,195,972	0.16	16
GlobCover	133,816,763	40,578,604	0.303	52,962,128	214,671,397	4.27	13
Other	104,888,810	50,533,806	0.482	4,198,002	205,579,617	11.01	2
Ang Thong	106,137,971	13,871,631	0.131	77,915,925	134,360,017	15.63	103
IRRI+GlobCover	98,753,252	13,644,497	0.138	70,993,313	126,513,191	10.31	81
IRRI	772,357	319,637	0.414	122,050	1,422,663	0.14	8
GlobCover	6,612,362	2,479,451	0.375	1,567,882	11,656,842	0.78	14
Thai Binh	260,670,296	27,265,098	0.105	206,150,363	315,190,230	72.01	253
IRRI+GlobCover	244,447,928	26,764,357	0.109	190,929,289	297,966,567	29.19	219
IRRI	43,599	2,245	0.051	39,111	48,087	0	8
GlobCover	16,178,769	5,201,422	0.321	5,777,883	26,579,656	1.57	26

Direct estimates of the total area planted in rice from the sample plots have a relatively high CVs because of the variability in the sizes of the sample plots. To produce mesh level estimates, which may improve the precision of the estimates for total area planted in rice, additional information to delineate rice area planted within each mesh were recorded on the paper maps, and subsequently digitized using the ALIS[10] methodology.

[10] The methodology provides for strict digitization guidelines to identify crop area planted.

Figure 3: Estimates of Total Planted Area in Savannakhet from Different Data Sources

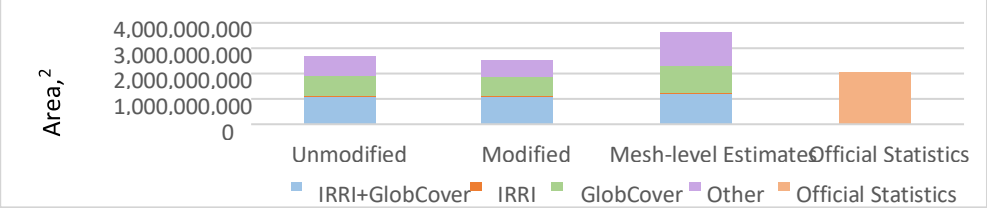


Figure 4: Estimates of Total Planted Area in Ang Thong from Different Data Sources

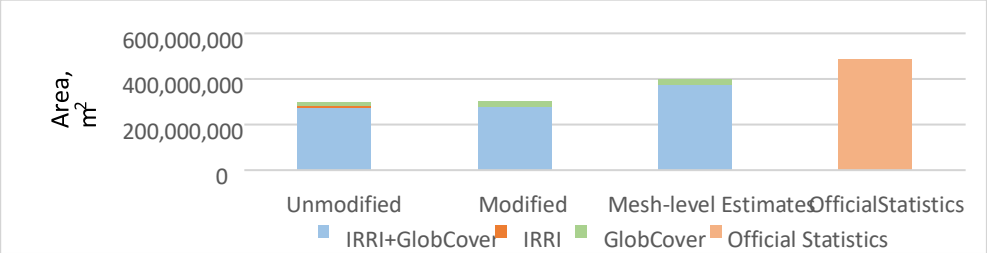


Figure 5: Estimates of Total Planted Area in Thai Binh from Different Data Sources

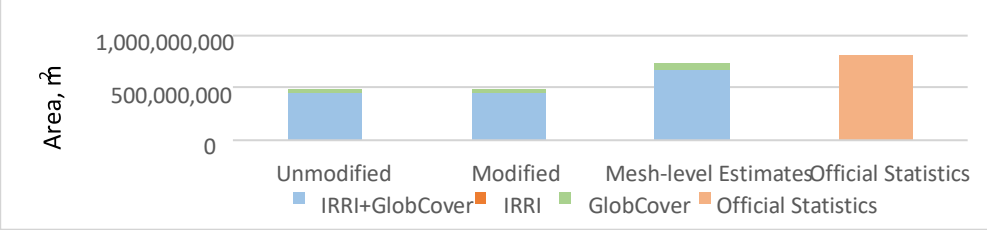


Table 7: Estimate of Total Area Planted in Rice Paddy in m^2, based on Independent Measure of Total Area Planted with Rice in Each Sample Mesh, using Digitized Google Earth Images

Domain	Area, m^2	SE	CV	95% confidence interval		DEFF	No. of observations
				Lower	Upper		
Savannakhet	3,607,317,242	270,231,340	0.075	3,068,869,543	4,145,764,940	1.24	78
IRRI+GlobCover	1,212,448,081	80,751,195	0.067	1,051,547,813	1,373,348,348	0.07	58
IRRI	41,913,872	10,269,821	0.245	21,450,806	62,376,938	0.02	8
GlobCover	1,049,786,538	217,401,053	0.207	616,605,485	1,482,967,592	1.08	10
Other	1,303,168,750	138,331,250	0.106	1,027,537,718	1,578,799,782	0.68	2
Ang Thong	398,383,039	17,855,187	0.045	362,702,295	434,063,784	1.14	66
IRRI+GlobCover	373,212,463	17,069,114	0.046	339,102,562	407,322,365	0.68	50
IRRI	2,639,752	393,107	0.149	1,854,190	3,425,314	0.01	10
GlobCover	22,530,824	5,224,800	0.232	12,089,895	32,971,753	0.21	6
Thai Binh	733,069,437	35,794,877	0.049	661,493,157	804,645,717	1.05	64
IRRI+GlobCover	670,129,990	35,046,872	0.052	600,049,439	740,210,542	0.7	55
IRRI	560,772	252,639	0.451	55,589	1,065,954	0.02	2
GlobCover	62,378,675	7,275,036	0.117	47,831,341	76,926,009	0.07	7

These estimates based on the mesh-level rice area data from the digitized Google Earth images could be considered the most accurate estimates of total

area planted in rice under the rice crop- cutting pilot survey methodology. These estimates can also be used for improving the level of precision of the

Domain	Estimate, kg	SE	CV	95% confidence interval	
				Lower	Upper
Savannakhet	693,823,889	93,955,279	0.135	521,700,609	890,005,302
Ang Thong	141,362,508	14,234,430	0.101	114,827,845	170,626,811
Thai Binh	395,857,496	28,887,778	0.073	337,361,510	450,601,602

estimates of total rice paddy production through ratio estimation.

Table 8: Ratio Estimate of Total Rice Paddy Production in the Pilot Areas

The measures of precision and design effects in the three provinces determined in this study can be useful for determining the sample size that would be needed for nationally representative surveys for measuring the total area and production of rice. In this case, it will be necessary to determine the scope of the survey in terms of the geographic domains to be covered. The sample size will be determined based on a target level of precision for each geographic domain covered by the survey.

4. Conclusion

This study explores the use of an area frame multi-stage stratified sampling methodology to collect paddy rice area and production data in three major rice-producing pilot areas: Savannakhet, Lao PDR; Ang Thong, Thailand; and Thai Binh, Viet Nam, comparing three approaches: (i) a direct estimate obtained through plot measurement using a GPS device, (ii) an alternative direct estimate obtained through digitization of farmer identified plot boundaries on a high-resolution Google Earth image, and (iii) a ratio estimate of total production of rice paddy involving the calculation of the total area planted in paddy rice based on independent mesh-level measures from the digitized Google Earth map. Yield estimates were calculated using crop-cutting techniques.

Results suggest that the direct estimates of the total rice paddy area and production from the sample plots have relatively high CVs and wide confidence intervals. We also found some inconsistencies in the stratification results. There are two possible explanations for the inconsistencies between satellitebased land cover classification and what was found during the fieldwork: (i) the power of discrimination in the satellite imagery and stratification might not be sufficient or (ii) field teams might not have accurately reported the status of all meshes, thereby systematically excluding some rice-growing meshes from the survey. This indicates that it will be necessary to improve the land use stratification of the frame by using higher resolution satellite images and a greater power of discrimination in the models used for defining the strata.

Mesh variability is also another important issue. The variability in the percentage of area planted in rice inside the meshes increases the CVs of the estimates of total area planted and the corresponding ratio estimates of total rice production. One alternative that can be explored is the possibility of using a different source of satellite data with a stronger discriminatory power for stratifying the meshes. Future work could also test the interviewer effort hypothesis to explain whether a mesh was visited and correctly enumerated by using a logistic regression framework with distance to main road, terrain, slope, enumerator fixed effects, and other covariates as explanatory factors.

The deviation between official statistics could be due to the presence of non-sampling errors, subjective intervention, and political leadership at the local government levels involving subsequent revisions in the administrative data collection method. In Lao PDR, there is almost a doubling of yield estimates from official data compared to crop cutting results, which warrants further investigation into the existing administrative data collection methods.

Despite these challenges, the use of remote sensing and GIS techniques to obtain rice area and production estimates with relatively high precision is a major reason for the benefit of this methodology compared to the existing administrative data collection system for which measures of precision are not publicly available. With the aid of handheld devices with inbuilt GPS functionalities, the field teams could navigate to the selected meshes, identify plot owners, conduct area measurements, and implement crop cutting. The European Space Agency recently launched the Sentinel-2 satellite which can provide images at 10 m spatial resolution every 5 days for no charge. As satellite data gets cheaper and better, there is a higher likelihood for developing the methodology to adopt area frames.

References

1. Asian Development Bank. 2016. *Results of the Methodological Studies for Agricultural and Rural Statistics.* Manila.
2. Faulkenberry, G. David, and Abderrazak Garoui. 1991. "Estimating a Population Total Using an Area Frame." *Journal of the American Statistical Association* 86(414): 445–449.
3. Grosh, Margaret E., and Juan Munoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Survey (English)." Living Standards Measurement Study Working Paper 126. http://documents.worldbank.org/curated/en/363321467990016291/Amanual-for-planning-and-implementing-the-living-standards-measurement-study-survey.
4. Griffin, Richard A. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30*(1): 177–189.* https://doi.org/10.2478/jos-2014-0012.
5. Google Earth. http://www.earth.google.com.
6. Huddleston, Harold F. 1978.Sampling Techniques for Measuring and Forecasting Crop Yields. U.S. Department of Agriculture. http://ageconsearch.umn.edu/record/142840/files/escs09.pdf.
7. Kelly, Valerie A., Jane Hopkins, Thomas Reardon, and Eric W. Crawford. 1995. Improving the measurement and analysis of African agricultural productivity: Promoting complementarities between micro and macro data. MSU International Development Paper No. 16. Michigan: Michigan State University. Retrieved from http://ageconsearch.umn.edu/bitstream/54055/2/idp16.pdf.

Enhancing the efficiency of the Proxy Mean Test Formula (PMTF) in targeting the poorest of the poor households

Adnan Dawood Khaleel Badran, Hanan Ali Mohamed Al Marzouqi, Eid Mohamed Al Qubaisi, Aysha Ali Al-Hosani
Statistics Centre-Abu Dhabi, Abu Dhabi, United Arab Emirates

Abstract

All targeting methods have the same goal to correctly and efficiently identify which households are eligible for benefit or which are not. There are six targeting mechanisms (Tools) available worldwide for Social Safety Net Programs (SSNP). The first is the Means Testing: this tool uses income as a source of family welfare and is used in developed countries where income is verified through special agencies, such as Tax Departments, Social Security…etc. The second is the Categorical Targeting: where eligibility is determined by age, gender, or some other demographic characteristics such as: Disability, aged members living alone, widowed. The third method is the Geographic Targeting: eligibility for benefits is determined, at least partly, by location of residence. The geographic method could be used as first stage in combination with other methods, the smaller the unit used the more accurate (sub-districts vs. districts or districts vs. provinces) and it is more viable for community goods and services. The fourth method is the Community-Based Targeting: A community leader or group of community members decides who in the community should receive benefits, like use existing local actor (teacher, nurse, and clergyman) or civic committee to identify beneficiaries and amounts, local actor may have best information, but personal relationships may lead to inequitable distribution. The fifth is the Self-Targeting: the good, or service are available for all and who ever think that he is poor, can benefit and get this good or service. The sixth method is the Proxy Means Testing Formula (PMTF): PMTF aims to predict household expenditure based on a number of easily observable characteristics (independent variables/ characteristics). Weights are determined for each characteristic that are found to be related to the household expenditure level; no need to measure the household income or expenditure.

The PMTF is a statistical method used to predict the income/ expenditure of a household based on observable characteristics that correlate with, but are easier to measure, than income/ expenditure. Measuring these variables from the households who are nominated for benefits is more credible than the measurement of income or expenditure. The suggested variables/ observable characteristics are related to the family welfare level and these are related to: geographic location of the household (geographic location of the household's

dwelling), dwelling characteristics (such as the number of bedrooms, availability of some durable goods), demographic characteristics (such as household size, gender of the head of the household), education characteristics for the household members, health characteristics, and the labor force characteristics for the household members. These suggested independent variables are used in constructing the PMTF to predict the dependent variable; per capita income/ expenditure. The predicted income/ expenditure (per capita) from the PMTF is used to compare with the household poverty line. The household is classified as poor when the predicted per capita income from the PMTF model is less than the poverty line; otherwise, the household is classified as non-poor. The weakness of the PMTF is always overestimated for the poorest segment of the households (our desired target); especially when the national poverty incidence is low.

The aim of this paper is to enhance the efficiency of the PMTF through building more than one formula to target the poor households. The main idea is to select three cutoff points on the natural logarithm of the annual per capita income/ expenditure. These cutoff points should be critical as much as we can, to separate the population into four segments of households: the first segment is the High Welfare Level (HWL) which don't include any poor household, the second is the Middle Welfare Level (MWL) which could include poor households, the third segment is the Lower Middle Welfare Level (LMWL) which also could include poor households, and the fourth is the Bottom Welfare Level (BWL) which includes most of the poor households. Three PMTFs are required to classify the population on these four categories. The households which belongs to the HWL will be ignored because they are not including any poor household. For the remaining three categories of households, we need to build another three specific PMTFs, formula for each category. The range of the first PMTF is lying between the first cutoff point and the second cutoff point and will represent the MWL. The range for the second PMTF is lying between the second cutoff point and the third cutoff point and will represent the LMWL. The range of the third PMTF is lying between the third cutoff point and the origin point and will represent the BWL. Figure (1) and Figure (2) explain the idea:

Figure (1) High Welfare Level and the Low Welfare Level on the Natural logarithm (Ln) annual per capita expenditure

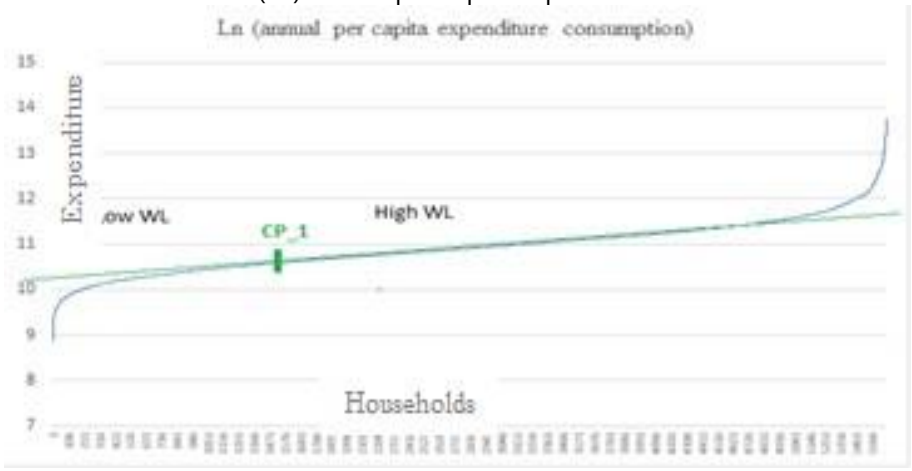


Figure (2) Middle WL, Lower Middle WL, and the Bottom WL and the Cutoff points

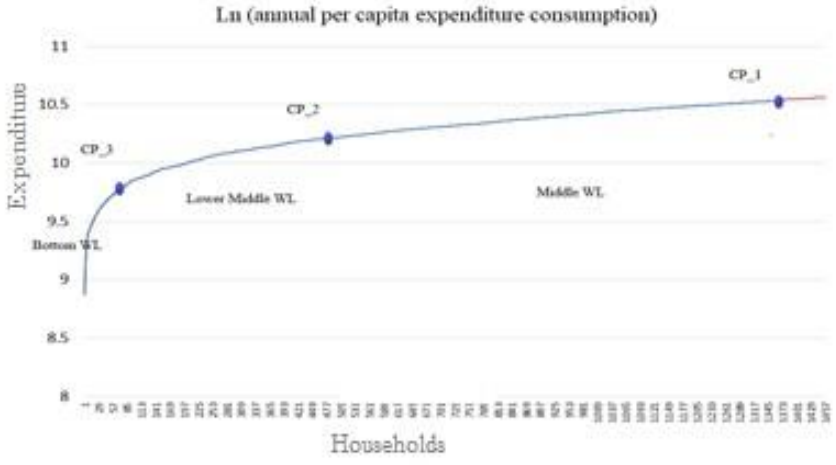


Figure (1) shows the natural logarithm of the annual per capita expenditure consumption and the trend line which represents the linear regression on the annual per capita function. It's clear that the line is higher than the lowest annual per capita expenditure consumption. That means the linear regression formula will overestimate these households, and will be excluded from benefit program. The suggested technique will enhance the efficiency for the PMTF and will target these household at the bottom of expenditure scale.

Keywords
Lower Welfare Level (LWL); High Welfare Level (HWL); Middle Welfare Level (MWL); Lower Middle Welfare Level (LMWL); Bottom Welfare Level (BWL);

1. Introduction

From a poverty perspective, as we know, the Proxy Mean Test Formula (PMTF) is used to build a linear regression to estimate the average annual per

capita expenditure to be used in determining whether the household is poor or not poor. As known, the results of the PMTF is overestimated for the poorest segment of households, whom their average per capita expenditure is very low. Because the estimation is a linear, it will always be above the lowest households expenditure, whom are our target.

When building the PMTF for targeting the poor households, it is used to use part of the population in order to enhance the efficiency of the formula. The households whom their expenditure is high will be ignored during the process of building the formula. The most common approach in constructing the PMTF for targeting the poor households is to play with the covered size of the population during the running process of the linear regression. Some of researchers include 50 percent of the population; some includes 30 percent of the population and so on. In fact, poverty incidence and poverty line play a key role in determining the size of population that should be covered in the running process. An accurate linear regression should be derived to predict the annual per capita expenditure, especially when we are targeting the poorest of the poor (the poorest of the abject poor).

As known, there are two specific indicators playing a major role in accepting or rejecting the efficiency of the formula. The first indicator is the leakage rate and is defined as: the proportion of households who are in reality non-poor and became poor by the formula (errors of inclusion), the second indicator is the under-coverage rate and is defined as: the proportion of households who are in reality poor and became not poor by the formula (errors of exclusion). The predicted poverty incidence, is a third indicator that should be taken into consideration through the process of constructing the formula; the predicted poverty incidence should be around the actual poverty incidence. A common approach to evaluate the under-coverage and leakage rates of a PMTF is a two-by-two table. Consider an actual case where there are 100 households and a poverty line that implies that 20 of these are classified as poor. And consider a PMTF that predicted 30 households are poor. Of these, 15 are actual poor and 15 are non-poor. Both the 15 poor households and the 65 non-poor households are treated as successful targeting. The 5 predicted non- poor households are treated as “errors of exclusion, while the 15 actual non-poor households are seen as “errors of inclusion. While trying to reduce errors of inclusion, it will also raise errors of exclusion. Similarly, raising the poverty line in order to reduce under-coverage will also tend to increase leakage.

Table (1) Leakage and Under-coverage Rate calculations

Predicted Welfare Level By the Formula	Actual Welfare Level		
	Poor	Non-Poor	Total
Non-Poor	5 Under-Coverage (Exclusion error) (Under-coverage rate=0.25)	65 Successful Targeting	70
Poor	15 Successful Targeting	15 Leakage (Inclusion error) (Leakage rate=0.50)	30
Total	20	80	100

There are no such predetermined rates for the leakage and under-coverage telling whether to accept the formula or to reject it, more reducing for these rates more efficiency of the formula. In the process of constructing the regression formula for targeting the poor, actions are taken to reduce these rates, but when reducing the leakage rate we could found that the under coverage increased. Alternatively, when trying to reduce the under-coverage rate we found that the leakage rate increased.

2. Methodology

This enhanced methodology of the PMTF is based on dividing the population into four categories of households; the first is the High Welfare Level (HWL) which don't have poor households, the second is the Middle Welfare Level (MWL) which could include poor households, the third segment is the Lower Middle Welfare Level (LMWL) and could include poor households, and the fourth is the Bottom Welfare Level (BWL) in which most of it are poor households. Determining these four segments is mainly based on selecting three cutoff points on the natural logarithm of annual per capita expenditure. Each cutoff point represent a step on the scale of the annual per capita expenditure consumption. To select these cutoff points we need knowledge of the poverty line and poverty incidence. The detailed methodology is explained through the following process:

Based on the poverty line and the poverty incidence in the society, we select a cutoff point (call it CP_1) on the natural logarithm curve for the annual per capita expenditure consumption. This CP_1 should be approximately ten per cent above the poverty line. This cutoff point divide the population into two welfare levels, High Welfare Level (HWL) which represent the non-poor households, and the Lower Welfare Level (LWL) which represent the rest of the

population (households).

For classifying the households whether they belong to HWL or to the LWL, we can build the first linear regression formula based on the suggested independent variables depending on the whole population, in order to predict whether the household belongs to HWL or LWL. Households with predicted annual per capita expenditure above the CP_1 are non-poor and classified as HWL. No more work is required with this category of households because they are not poor.

Based on the households belonging to the LWL and a trend line on the natural logarithm curve a second cutoff point (CP_2) should be selected carefully, and then the population is divided into two segments; the Middle Welfare Level (MWL) and the Rest of LWL. Next, we build the second linear regression formula based on the suggested independent variables for the LWL households. This formula is to predict whether the household belongs to the MWL or to the Rest of LWL. Households for whom the predicted annual per capita expenditure is above the CP_2 are classified with MWL. These households are lying between CP_1 and CP_2. A third regression formula should be built for these households which lies in the MWL, this formula is for predicting the annual per capita expenditure for these households belonging to MWL.

In order to divide the households belonging to the Rest LWL into two final segments; Lower Middle Welfare Level (LMWL) and the Bottom Welfare Level (BWL), a third cutoff point (CP_3) should be selected based on the abject poverty line and the trend line on the natural logarithm curve. The fourth PMTF should be build based on the households belonging to the Rest of LWL. Households with a predicted annual per capita expenditure above the CP_3 are classified as LMWL and the other is classified to the BWL.

A fifth specific PMTF should be built for LMWL households, based on those households lying between CP_2 and CP_3 to predict the annual per capita expenditure. Lastly, in order to predict the average annual per capita expenditure for the households lying in the BWL (below CP_3) the sixth linear regression formula through the origin; based on the suggested independent variables for the BWL, is built.

At the end of construction process for the formulas of the PMTF, we got six formulas: three formulas for classifying the households between the four categories: high welfare category, middle welfare category, lower welfare category, and the bottom welfare category. And we have another three formulas for predicting the annual per capita expenditure consumption; this predicted expenditure will be used as a final prediction for the household to decide whether it is poor or not.

3. Results

We can describe the whole process of the Enhanced PMTF in the following four steps:

i. Applying PMTF1; is the predicted expenditure above CP_1, if yes then the household belongs to HWL, then household is not poor, if no continue.
ii. Applying PMTF2, Is the predicted expenditure above CP_2, if yes then the household belongs to MWL; apply PMTF3 to predict the annual per capita expenditure and decide whether the household is poor or not, if no continue (below CP_2).
iii. Applying PMTF4, Is the predicted expenditure above CP_3, if yes then the household belongs to LMWL; apply PMTF5 to predict the annual per capita expenditure and decide whether the household is poor or not, if no continue (below CP_3).
iv. Applying PMTF6 for the BWL to predict the annual per capita expenditure and decide whether the household is poor or not.

This new mechanism on using the PMTF is enhancing the efficiency of targeting the poor households, where the leakage rate and the under-coverage rate are decreased significantly. Especially the decrease in the under-coverage rate, because of the sixth PMTF which covers the poorest of poor is passing through the origin. Hereafter, a table explains how the under-coverage rate and the Leakage rate are calculated.

4. Discussion and Conclusion

Using the PMTF is one of the best from the six targeting methods available and most accurate method. The PMTF method does not force us to ask about income or expenditure, especially in "poor countries" where no official administrative data on households/ individuals income. The enhancement on the efficiency of the PMTF proposed in this paper; by applying more than one formula for targeting the poorest of the poor households, increase the importance and the credibility of the PMTF in targeting poor households, and guarantee that the assistance goes to those who deserve it.

The suggested number of cutoff points and the number of equations; which serves in increasing the efficiency of the PMTF, are changeable based on poverty line, poverty incidence in the society, and the available number of cases required to run the linear regression are available. In this paper we have suggested three equations for classifying the households in four categories; HWL, MWL, LMWL, and the BWL through three cutoff points on the scale of annual per capita income/ expenditure. The number of cutoff points can be decreased to two points, then; three equations to classify the households in three categories and get two final equations for the final PMTF.

References

1. David Coady, Margaret Grosh, John Hoddinott. (2004) The International Bank for Reconstruction Targeting of Transfers in Developing Countries: Review of Lessons and Experience. http://siteresources.worldbank.org/SAFETYNETSANDTRANSFERS/Resources/2819451138140795625/Targeting_En.pdf
2. Dennis S. Mapa Manuel Leonard F. Albis (October 1-2, 2013). NEW PROXY MEANS TEST (PMT) MODELS: IMPROVING TARGETING OF THE POOR FOR SOCIAL PROTECTION, http://nap.psa.gov.ph/ncs/12thncs/papers/INVITED/Special%20Session/Special%20Session_%20New%20Proxy%20Means%20Test%20(PMT)%20Models-%20Improving%20Targeting%20of%20the%20Poor%20for%20Social%20Protection.pdf
3. John Weiss. 2004 Asian Development Bank Institute. ADB Publishing 3/04. Poverty Targeting in Asia: Experiences from India, Indonesia, the Philippines, People's Republic of China and Thailand, Retrieved from https://www.adb.org/sites/default/files/publication/157277/adbi-rpb9.pdf
4. DEBBIE BUDLENDER. (St.Lucia, 2014) CONSIDERATIONS IN USING PROXY MEANS TESTS IN EASTERN CARIBBEAN STATES
5. Larry Dershem, Ph.D. (December 2013) Using a Proxy Means TesT for TargeTing, Retrieved from https://www.researchgate.net/publication/259451996_Using_a_Proxy_Means_Test_for_Targ eting_in_a_Conditional_Cash_Transfer_Program
6. - Sha'aban, R.A., " Poverty Alleviation in Jordan-Lessons for the future ", The World Bank, Middle East and North Africa Region. S. R. Searle, (1971) " Linear Models.
7. http://worldbank.com/poverty/mission/up1.htm.
8. http://www.worldbank.org/poverty/mission/up2.htm.

Parametric Weibull Time-Varying Covariate Model for HIV-TB Mortality

Mohd Asrul Affendi Abdullah[1], Oyebayo Ridwan Olaniran[2], Siti Afiqah Muhammad Jamil[1]

[1] Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia

[2] Department of Statistics, Faculty of Physical Sciences, University of Ilorin

Abstract

Parametric Weibull survival model has been applied to several failure time distribution of many diseases including the co-infection of Human Immunodeficiency Virus HIV and Tuberculosis (TB). However, covariate(s) in the Weibull survival regression may depend on time. A typical example in HIV-TB co-infection is the occurrence of TB infection at varying time in HIV patients. This modelling situation violates the standard assumption of proportional hazard models like Cox or ordinary Weibull regression that do not incorporate the time-varying effect. Simulating time-varying covariate model poses a serious problem in survival analysis because the covariate that needs to be generated to obtain the hazard or survival function depends on time. In this paper, we present a simulation strategy for generating a parametric Weibull time-varying covariate model. We also present an estimation technique for the model using the maximum likelihood method. The validity of the simulation scheme as well as the estimation method was observed using bias and mean square error criterion. Comparison between the estimation method with standard Cox regression and Weibull regression model under fixed and time-varying covariate assumption was also achieved. Appreciable supremacy was observed for the proposed method over the competing methods.

Keywords

Simulation; Weibull distribution; Time-varying covariate; HIV-TB co-infection

1. Introduction

The occurrence of a particular event such as the time of death, time of relapse, time of recovery, is commonly associated with survival analysis (Collett, 2015). The usual trend in survival analysis is to observe the distribution so that an appropriate method could be perfectly applied to the event of study. Generally, when the event distributions are known in advance, the parametric method can be applied. Otherwise, non-parametric or semi-parametric methods are often suitable for the analysis. Besides, in modelling the lifetime data, sometimes, the semi-parametric method could be more accurate depending on the situation of the data (Leffondré et al., 2013).

Many authors had applied parametric survival procedure in AIDS studies, among others are Elfaki et al. (2012) and Singh & Totawattage (2013), used Weibull distribution to model time to event in HIV/AID studies, Kiani et al. (2012) used Gompertz distribution. On the other hand, Lopez-Gatell et al. (2007) studied the effect of tuberculosis on the progression of the HIV-1 disease. The study focused on the effects of time-varying incident TB on time to AIDS-related, and all-cause mortality with the changes in CD4 cell count among HIV-1 infected women exposed to highly active antiretroviral therapy (HAART), non-exposed and combined. The marginal structural model is used with considering time-varying confounding. It was further observed that the confounding effect of TB onset for exposed HAART is on the high side. Some semi-parametric models including the Cox proportional hazard model have been developed to capture time-varying covariate effect(s). Tseng et al. (2014) considered a Cox-like model called extended hazard to model fixed and time-varying covariate model. Therneau and Lumley (2016) used the counting process strategy to build a Cox model for the fixed and time-varying covariate. Within the parametric context, Sparling et al. (2006) developed parametric time-varying covariate models for interval-censored data. However, it has been observed that parametric methods are more accurate for modelling of complex data structures (Aalen et al., 2008, Jackson, 2016). An example of the complex data structure is time-varying covariate which may be model as a parameter of the assumed distribution. In this paper, we modelled the fixed and time-varying covariate effect of Weibull distribution using its scale parameter.

2. Materials and Methods

Weibull Time-Varying Covariate Model for HIV-TB Mortality: Rodrıguez (2010) described four modelling approaches for parametric survival models namely; parametric families, proportional hazard, accelerated failure time and proportional odds. However, we decided to use the parametric families approach because of its simplicity and flexibility. Now, suppose T is an event time that follows a Weibull distribution with shape and scale parameters define as(a,b), its density function, hazard function and survival function are (Alharpy and Ibrahim (2013));

$$f(t|a,b) = \left(\frac{a}{b}\right)\left(\frac{t}{b}\right)^{a-1} exp\left[-\left(\frac{t}{b}\right)^{a}\right],\ t > 0, a, b > 0 \tag{1}$$

$$h(t|a,b) = \left(\frac{a}{b}\right)\left(\frac{t}{b}\right)^{a-1}, t > 0, a, b > 0 \tag{2}$$

$$S(t|a,b) = exp\left[-\left(\frac{t}{b}\right)^{a}\right],\ t > 0, a, b > 0. \tag{3}$$

If we consider the parametric family approach, the covariate associated with

the event time T can be modelled via the scale parameter b as,

$$b = exp[x'\beta] \tag{4}$$

$$logb = x'\beta$$

The form of parameterization in (4) is when the fixed time covariate or proportional assumption is satisfied. But when time-varying covariate z(t) exist, (4) can be modified as;

$$b = exp[x'\beta + \gamma z(t)] \tag{5}$$

Where β and γ represent the parameters of the fixed and time varying covariate.

For the HIV-TB mortality model, $z(t)$ is a binary vector representing TB infection at varying interval of time before the end of study. The matrix x represents the fixed covariate such as gender, age, marital status etc. substituting (5) in (1), we can obtain the density function for Weibull with fixed and time varying covariate effects as;

$$f(t|a,x,z(t)) = \left(\frac{a}{exp[x'\beta+\gamma z(t)]}\right)\left(\frac{t}{exp[x'\beta+\gamma z(t)]}\right)^{a-1} exp\left[-\left(\frac{t}{exp[x'\beta+\gamma z(t)]}\right)^{a}\right] \tag{6}$$

Now assuming z(t) to be a piece-wise function as in the case of TB infection occurring with HIV infection, z(t) can be define as;

$$z(t) = \begin{cases} 0, & t < t_c \\ 1, & t \geq t_c \end{cases}$$

Where t_c is the time at which the covariate z(t) changes. This implies that f(t | a,x,z(t)) will also be a piece-wise function. Therefore f(t | a,x,z(t)) can be define as;

$$f(t|a,x,z(t)) = \begin{cases} f(t|a,x), & t < t_c \\ f(t|a,x,\gamma z(t)), & t \geq t_c \end{cases}$$

$$f\left(t\middle|a,x,z(t)\right) = \begin{cases} \left(\frac{a}{exp[x'\beta]}\right)\left(\frac{t}{exp[x'\beta]}\right)^{a-1} exp\left[-\left(\frac{t}{exp[x'\beta]}\right)^{a}\right], & t < t_c \\ \left(\frac{a}{exp[x'\beta+\gamma z(t)]}\right)\left(\frac{t}{exp[x'\beta+\gamma z(t)]}\right)^{a-1} exp\left[-\left(\frac{t}{exp[x'\beta+\gamma z(t)]}\right)^{a}\right], & t \geq t_c \end{cases}$$

Consequently, the hazard function h(t | a,x,z(t))associated with T can be obtained as;

$$h(t|a,x,z(t)) = \begin{cases} \left(\frac{a}{exp[x'\beta]}\right)\left(\frac{t}{exp[x'\beta]}\right)^{a-1}, & t < t_c \\ \left(\frac{a}{exp[x'\beta+\gamma z(t)]}\right)\left(\frac{t}{exp[x'\beta+\gamma z(t)]}\right)^{a-1}, & t \geq t_c \end{cases}$$

The cumulative hazard function H(t | a,x,z(t))associated with T can be obtained as;

$$H(t|a,x,z(t)) = \begin{cases} \int_0^t \left(\frac{a}{exp[x'\beta]}\right)\left(\frac{u}{exp[x'\beta]}\right)^{a-1} du, & t < t_c \\ \int_0^t \left(\frac{a}{exp[x'\beta+\gamma z(u)]}\right)\left(\frac{u}{exp[x'\beta+\gamma z(u)]}\right)^{a-1} du, & t \geq t_c \end{cases}$$

Thus;

$$H(t|a,x,z(t)) = \begin{cases} \left(\frac{t}{exp[x'\beta]}\right)^a, & t < t_c \\ \left(\frac{t_c}{exp[x'\beta]}\right)^a + \left(\frac{t}{exp[x'\beta+\gamma]}\right)^a - \left(\frac{t_c}{exp[x'\beta+\gamma]}\right)^a, & t \geq t_c \end{cases}$$

The survival function S(t | a,x,z(t)) follows as;

$$S(t|a,x,z(t)) = \exp[-H(t|a,x,z(t))]$$

$$S(t|a,x,z(t)) = \begin{cases} exp\left[-\left(\frac{t}{exp[x'\beta]}\right)^a\right], & t < t_c \\ exp\left[-\left[\left(\frac{t_c}{exp[x'\beta]}\right)^a + \left(\frac{t}{exp[x'\beta+\gamma]}\right)^a - \left(\frac{t_c}{exp[x'\beta+\gamma]}\right)^a\right]\right], & t \geq t_c \end{cases}$$

Parameter Estimation for Weibull Time-Varying Covariate Model: Following Lessafre and Lawson (2013), Jamil et al. (2017), Olaniran and Yahya (2017), Olaniran and Abdullah (2017), Olaniran and Abdullah (2018a), Olaniran and Abdullah (2018b) among others the likelihood of a parametric survival model with right censored times can be define as;

$$L(\theta) = \prod_{uce} f(t_{uce}|\theta) \prod_{ce} S(t_{ce}|\theta) \quad (7)$$

Where *uce* denote uncensored and *ce* denotes right censored. The likelihood in (7) can be simplified if a censoring indicator s_i that takes 0 for censored and 1 for uncensored is assumed. Thus,

$$L(\theta) = \prod_{i=1}^{n} h(t_i|\theta)^{s_i} S(t_i|\theta) \quad (8)$$

For the case of time-dependent covariate, we define c_i to be the time dependent indicator as a piece-wise function;

$$c_i = \begin{cases} 0, & if\ the\ value\ of\ covariates\ is\ not\ updated \\ 1, & if\ the\ value\ of\ covariates\ is\ updated \end{cases}$$

The corresponding log-likelihood is;

$$l(a,\beta,\gamma) = \sum_{i=1}^{n}(1-c_i)\left[s_i log\left(\left(\frac{a}{exp[x_i'\beta]}\right)\left(\frac{t_i}{exp[x_i'\beta]}\right)^{a-1}\right) - \left[\left(\frac{t_i}{exp[x_i'\beta]}\right)^{a}\right]\right] + \sum_{i=1}^{n} c_i\left[s_i log\left(\left(\frac{a}{exp[x_i'\beta+\gamma]}\right)\left(\frac{t_i}{exp[x_i'\beta+\gamma]}\right)^{a-1}\right) + \left[\left(\frac{t_{ci}}{exp[x_i'\beta]}\right)^{a} + \left(\frac{t_i}{exp[x_i'\beta+\gamma]}\right)^{a} - \left(\frac{t_{ci}}{exp[x_i'\beta+\gamma]}\right)^{a}\right]\right]$$

The model parameters a,β,γ can be obtained using the Newton-Raphson algorithm to find the solution of the likelihood equations given above. However, we have used the SPLUS function *nlminb* to obtain the results.

Simulation of Parametric Weibull Time-Varying Covariate Model: Suppose *W* is a random variable with cumulative distribution function *F*, it suffices that F(*w*) follows uniform distribution, *U~UNIF*[0,1]. Similarly, the survival function 1-F(*w*) follows *U~UNIF*[0,1]. Therefore, the survival function in the case of parametric Weibull time covariate model can be define as;

$$= \begin{cases} exp\left[-\left(\frac{t}{exp[x'\beta]}\right)^{a}\right], & t < t_c \\ exp\left[-\left[\left(\frac{t_c}{exp[x'\beta]}\right)^{a} + \left(\frac{t}{exp[x'\beta+\gamma]}\right)^{a} - \left(\frac{t_c}{exp[x'\beta+\gamma]}\right)^{a}\right]\right], & t \geq t_c \end{cases}$$

Thus, by inverting the above piece-wise Weibull function, the survival time *T* can be obtained as;

$$T = \begin{cases} exp[x'\beta][-\log(U)]^{\frac{1}{a}}, & U < exp\left[-\left(\frac{t_c}{exp[x'\beta]}\right)^{a}\right] \\ exp[x'\beta+\gamma]\left[-\log(U) - \left(\frac{t_c}{exp[x'\beta]}\right)^{a} + \left(\frac{t_c}{exp[x'\beta+\gamma]}\right)^{a}\right]^{\frac{1}{a}}, & U \geq exp\left[-\left(\frac{t_c}{exp[x'\beta]}\right)^{a}\right] \end{cases}.$$

Simulation Studies: To illustrate the simulation strategy and estimation method, we used the following parameters; α=1, β=2 and *γ*=1. The covariate *x_i~Binomial*(1,0.5), t_{ci}*~Weibull*(*shape*=1,*scale*=0.04). Censoring rates 20% and 40% were used to check the effect of censoring rate on parameter estimates. Also, sample sizes *n*=100, and *n*=200 were used to study the effect of sample size on parameter estimates. Performance metrics used to assess the estimating methods are standard error (SE), bias and Mean square error (MSE).

3. Result and Discussion

The estimates of the proposed method obtained were compared with estimates using Cox proportional hazard implemented in SPLUS via coxph, Weibull fixed covariate regression via survreg, and Weibull time-varying covariate regression via flexsurvreg. The results based on performance metrics are presented in the table 1 The proposed method is PWTVC (Proposed Weibull time-varying covariate).

Table 1: Simulation results for average bias (bias(θˆ)), average standard error (SE(θˆ)) and average mean square error (MSE(θˆ)) based on 1000 replications for sample size n=100 and censoring rate 20%.

Metrics	Parameter	Fixed covariate model		Time varying covariate model		
		Coxph FC	Survreg FC	Coxph TVC	Flexsurvreg	PWTVC
$SE(\hat{\theta})$	β	0.2634	0.2751	0.2634	0.2519	0.1878
	γ	0.9863	0.1616	0.9863	0.2976	0.1160
	β,γ	0.7219	0.2256	0.7219	0.2757	0.1561
$bias(\hat{\theta})$	β	-0.8891	-0.9374	-0.8891	-0.9237	-1.3363
	γ	2.0374	0.1657	2.0374	1.5003	0.0237
	β,γ	0.5741	-0.3858	0.5741	0.2883	-0.6563
$MSE(\hat{\theta})$	β	0.8599	0.9544	0.8599	0.9166	1.8209
	γ	5.1237	0.0536	5.1237	2.3395	0.0140
	β,γ	0.8507	0.1998	0.8507	0.1591	0.4551

Table 1 shows the results for moderate sample size 100 and low censoring rate 20%. For all the performance metrics used, the results of the proposed method are better than the competing methods except in the estimation of the fixed covariate effect β. Specifically, the proposed method is more stable in terms of low average standard error, consistent in terms of low average bias as well as efficient in terms of low mean square error. However, the interesting results are more attributable to the time varying effect parameter γ which is our main focus.

4. Conclusion

In this paper, we have presented a simulation strategy for generating HIV-TB survival time with time-varying covariate for Weibull distribution. We also developed the corresponding likelihood-based estimator for estimating the parameters of parametric Weibull time-varying covariate model. The results from the simulation studies affirm the adequacy of the simulation strategy as well as the estimation method regarding consistency and efficiency. We also stressed the use of parametric distribution when the underlying distribution is known in advance.

Funding

This work was supported by Universiti Tun Hussein Onn, Malaysia [grant numbers Vot, U374, U607].

References

1. Aalen O, Borgan O, Gjessing H (2008). Survival and Event History Analysis: A Process Point of View. Springer-Verlag.
2. Alharpy, A.M. & Ibrahim, N.A. (2013). Parametric tests for partly interval-censored failure time data under Weibull distribution via multiple imputation.Journal of Applied Sciences, 13(4), 621.
3. Collett, D. (2015). Modelling survival data in medical research. CRC press.
4. Elfaki, F.A.M., Azram, M. & Usman, M. (2012). Parametric cox's model for partly interval-censored data with application to aids studies. International Journal of Applied Physics and Mathematics, 2(5), 352.
5. Jackson, C. H. (2016). flexsurv: a platform for parametric survival modelling in R. *Journal of Statistical Software*, *70*(8), 1-33.
6. Jamil, S. A. M., Abdullah, M. A. A., Kek, S. L., Olaniran, O. R., & Amran, S. E. (2017). Simulation of parametric model towards the fixed covariate of right censored lung cancer data.
7. In *Journal of Physics: Conference Series* (Vol. 890, No. 1, p. 012172). IOP Publishing.
8. Kiani, K., Arasan, J. & Midi, H. (2012). Interval estimations for parameters of gompertz model with time-dependent covariate and right censored data. Sains Malaysiana, 41(4), 471–480.
9. Leffondré, K., Touraine, C., Helmer, C. & Joly, P. (2013). Interval-censored time-to-event and competing risk with death: Is the illness-death model more accurate than the cox model? International journal of epidemiology, 42(4),1177–1186.
10. Lesaffre, E., & Lawson, A. B. (2013). Bayesian Biostatistics. *Wiley, West Sussex, UK*, 534.
11. Lopez-Gatell, H., Cole, S., Hessol, N., French, A., Greenblatt, R., Landesman, S., Preston-Martin, S. & Anastos, K. (2007). Effect of tuberculosis on the survival of women infected with human immunodeficiency virus. American journal of epidemiology, 165(10), 1134–1142.
12. Rodrıguez, G. (2010). *Parametric Survival Models*. Technical report, Princeton, NJ: Princeton University.
13. Olaniran, O. R., & Yahya, W. B. (2017). Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. Journal of Modern Applied Statistical Methods, 16(2), 618-638.
14. Olaniran, O. R., & Abdullah, M. A. A. (2017). Gene Selection for Colon Cancer Classification using Bayesian Model Averaging of Linear and

Quadratic Discriminants, Journal of Science and Technology-Penerbit UTHM, 9(3).
15. Olaniran, O. R. & Abdullah, M, A. A., (2018a). BayesRandomForest: An R implementation of Bayesian Random Forest for Regression Analysis of High-dimensional Data, Romanian Statistical Review, 66(1), 95-102.
16. Olaniran, O. R., & Abdullah, M. A. A. (2018b). Bayesian Analysis of Extended Cox Model with Time-Varying Covariates using Bootstrap Prior. Journal of Modern Applied Statistical Methods, *Accepted. In press.*
17. Singh, R.S. & Totawattage, D.P. (2013). The statistical analysis of interval-censored failure time data with applications.
18. Sparling, Y.H., Younes, N., Lachin, J.M. & Bautista, O.M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. Biostatistics, 7(4), 599–614.
19. Therneau, T. M., & Lumley, T. (2016). Package 'survival'.
20. Tseng, Y.K, Hsu K.N & Yang Y.F (2014) A semiparametric extended hazard regression model with time-dependent covariates, Journal of Nonparametric Statistics, 26:1, 115-128.

Silver tsunami in Johor

Kamaruzaman Mohamed, Noor Haninah Hasri, Amerudin Abdul Ghani, Norfazilany Ahmad, Thanusha, P.T, Shalini,M, Nor Hidayah Halil
Department of Statistics Malaysia

Abstract
The Silver Tsunami (also known as The Grey Tsunami or Gray Tsunami) is a metaphor used to describe population aging. According to Ngram Viewer, variants of the Silver Tsunami metaphor (for example, *age wave, grey hoard, rising tide, grey* or *gray tsunami*) first occurred in reference to population aging in the 1980s. It can be determined where the population of older people are 7 per cent and above of the total population. This article will explore the possibility that Johor becoming an aging state in 2020. By using descriptive analysis on population data of Johor 2010 till 2020, other related variable also collected to understand more about this Silver Tsunami issue in Johor.

Keywords
Johor, Population, Aging.

1. Introduction

The population growth continue in some fast-growing districts and the phenomenon will have profound impacts on various dimensions of society and this aging trend will be intensified in the coming decades. The impact and wave of Silver Tsunami that will affect the state of Johore are believed to be seen directly. It is due to increasing longevity and declining fertility, most developed areas have seen an increase in their elderly proportion (Weil, 1997).

Recently, however, the aging problem is becoming more and more serious, not only in the countryside, but also in the urban areas to which young people are supposedly migrating. Why is the population aging in the urban areas? The economic factors such as job opportunities, fields of economic activities and to attained academic qualification/further studies were identified as most influenced factors that shift migration of population from Johor to other location.

The growing number of elderly will provides great challenges and huge impact to economy, health and social development. In Malaysia, aging population can be seen through the different between males and females, urban and rural area and also ethnic groups. Since Malaysia is divided into 14 states and many districts for each state, there are very limited studies about the aging population for every state and districts. So, this study aims is to

investigate the pattern of aging population in Johor state from 2010 to projection 2020.

2. Methodology

The source of data will be used in this study is secondary data. The data from Department of Statistics Malaysia and Department of Social Welfare obtained:

i. Actual Population and Housing Census 2010
ii. Projection Data from 2011 to 2020

The data should be analyzed only using descriptive statistics. It focused on the percentage of the aging population.

Formula percentage of elderly population: $\frac{Total\ Population\ of\ Elderly}{Total\ Population} \times 100$

Formula percentage of elderly by gender: $\frac{Total\ Population\ of\ Elderly\ by\ Gender}{Total\ Population} \times 100$

3. Result

a. Population Pyramid 2010 until 2020

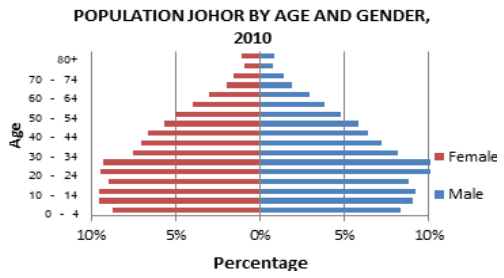


Figure 3.1A:Population Johor by Age and Gender 2010

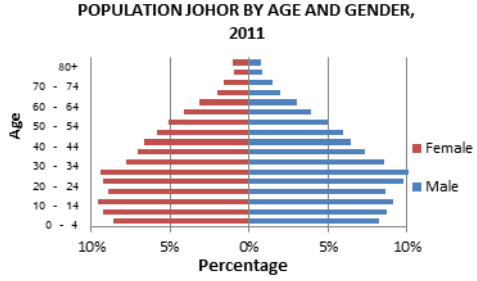


Figure 3.1B:Population Johor by Age and Gender 2011

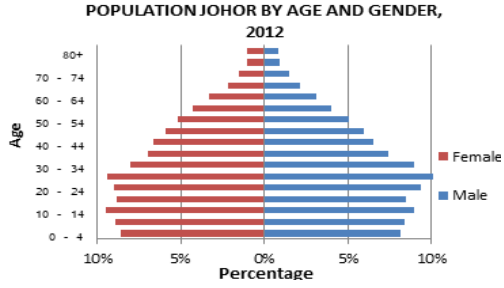


Figure 3.1C:Population Johor by Age and Gender 2012

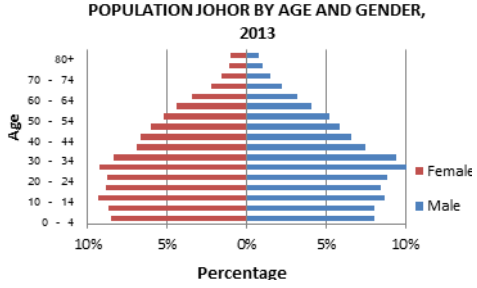


Figure 3.1D:Population Johor by Age and Gender 2013

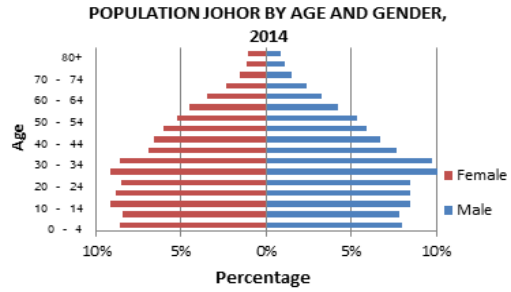


Figure 3.1E:Population Johor by Age and Gender 2014

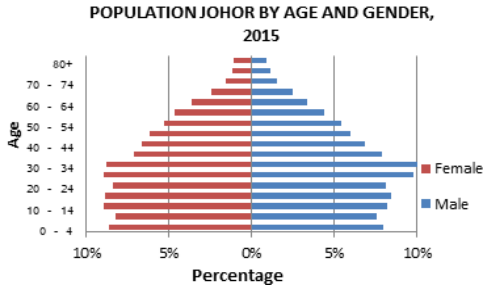


Figure 3.1F:Population Johor by Age and Gender 2015

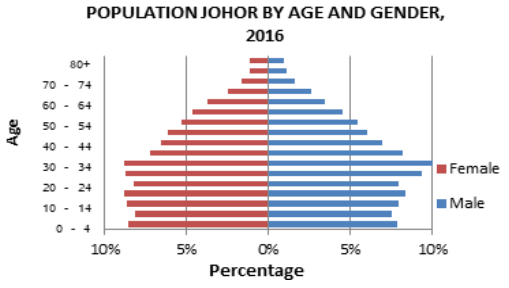


Figure 3.1G:Population Johor by Age and Gender 2016

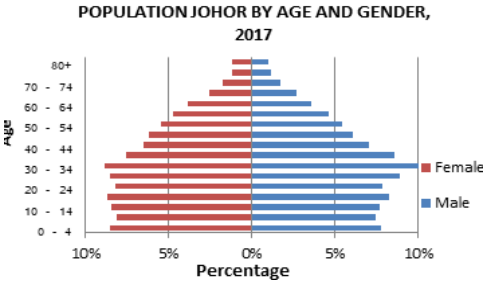


Figure 3.1H:Population Johor by Age and Gender 2017

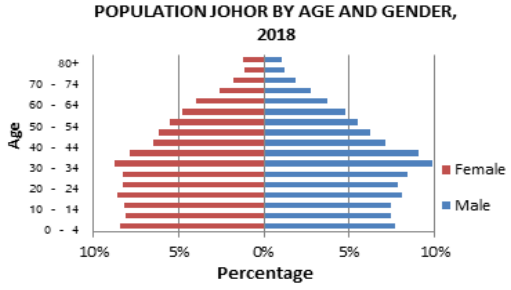


Figure 3.1I:Population Johor by Age and Gender 2018

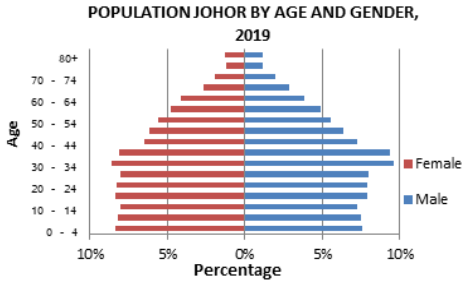


Figure 3.1J:Population Johor by Age and Gender 2019

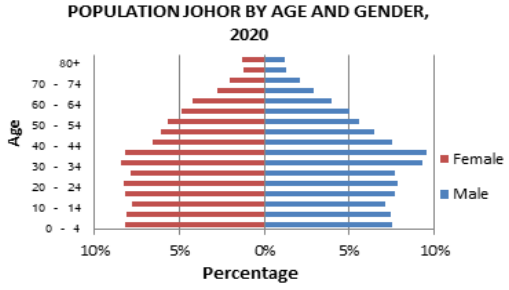


Figure 3.1K:Population Johor by Age and Gender 2020

The population pyramids 2010 till 2020 shows the change of shape from the pyramid to the shape of the bell. This explains the growing "aging wave" situation in Johor.

b. Aging Population by Gender & Districts, Johor 2010-2020

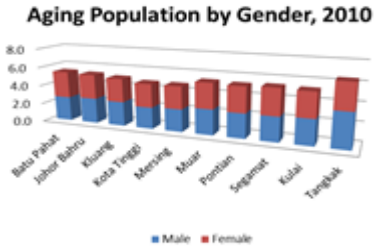


Figure 3.2A:Population Johor Gender & Districts,2010

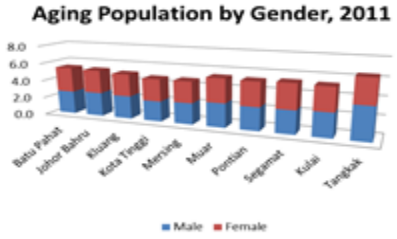


Figure 3.2B:Population Johor Gender & Districts,2011

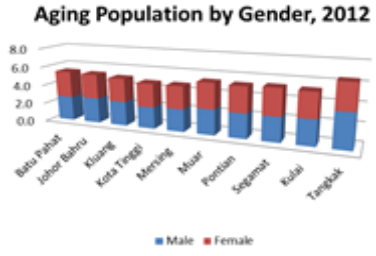


Figure 3.2C:Population Johor Gender & Districts,2012

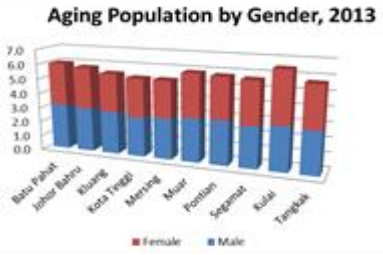


Figure 3.2D:Population Johor Gender & Districts,2013

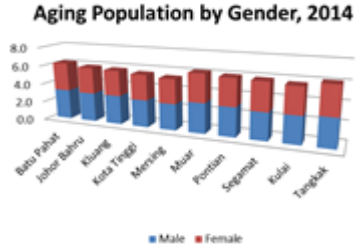


Figure 3.2E:Population Johor Gender & Districts,2014

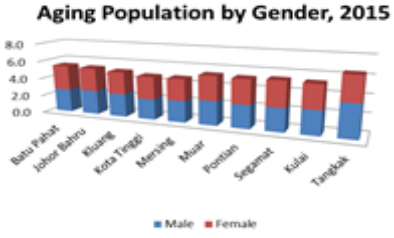


Figure 3.2F:Population Johor Gender & Districts,2015

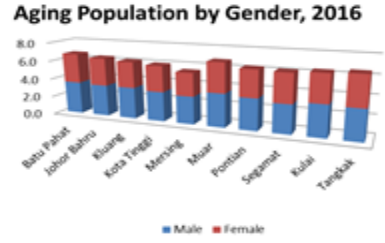


Figure 3.2G:Population Johor Gender & Districts,2016

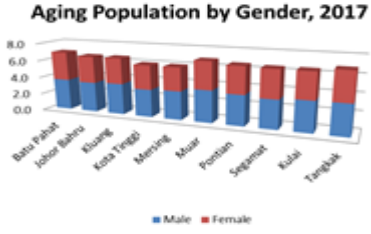


Figure 3.2H:Population Johor Gender & Districts,2017

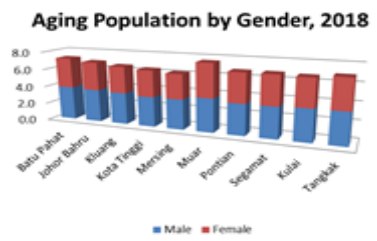


Figure 3.2I:Population Johor Gender & Districts,2018

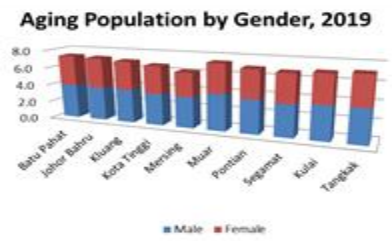


Figure 3.2J:Population Johor Gender & Districts,2019

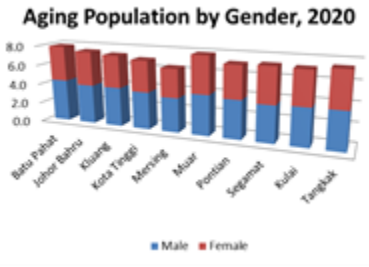
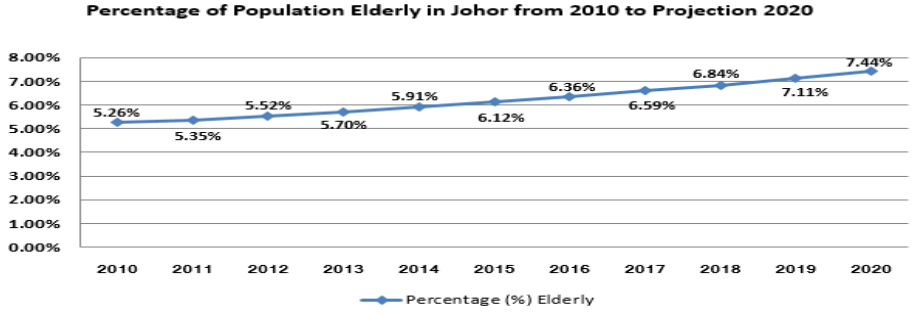


Figure 3.2K:Population Johor Gender & Districts,2020

c. Comparison Population of Elderly from 2010 to Projection 2020

The population of elderly in Johor shows that there are no specific patterns among them. But, there are continuously increasing in elderly group from year to year. According to the table its shows that in 2011 the percentage of elderly is slightly increase by 0.09. Meanwhile bigger impact shown on 2017 to 2018 which is 0.33 per cent. Silver Tsunami in Johor state dominate most of the districts on 2019 with 7.11 % of elderly.



Age Group	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
65 - 69	65.4	69.1	74.1	79.7	85.2	90.1	94.7	98.8	102.8	106.9	112.4
70 - 74	50.9	52.4	53.1	53.4	54.5	57.0	60.6	65.4	70.8	76.2	81.7
75 - 79	28.7	30.8	33.8	37.1	39.9	41.7	43.2	44.0	44.6	46.0	48.8
80+	31.9	31.2	31.7	32.4	33.7	35.7	37.8	40.4	43.4	46.3	49.3
Total of elderly	176.9	183.5	192.7	202.6	213.3	224.5	236.3	248.6	261.6	275.4	292.2
Total Population ('000)	3,362.9	3,428.0	3,490.9	3,551.7	3,610.3	3,665.5	3,717.2	3,770.4	3,822.7	3,874.5	3,926.5
Percentage (%) Elderly	5.26%	5.35%	5.52%	5.70%	5.91%	6.12%	6.36%	6.59%	6.84%	7.11%	7.44%

d. Population of Elderly for Each District from 2010 to Projection 2020

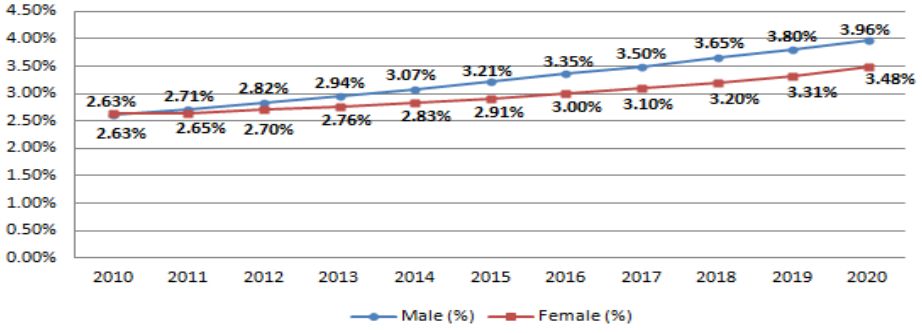
According to this table, Mersing and Kota Tinggi will not facing with the Silver Tsunami situation until 2020.

District	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Batu Pahat	5.4	5.2	6.1	6.0	6.2	6.6	6.6	6.9	7.1	7.3	7.9
Johor Bahru	5.3	5.4	5.5	5.7	5.9	6.1	6.3	6.6	6.9	7.2	7.4
Kluang	5.1	5.3	5.4	5.5	5.7	5.9	6.2	6.5	6.7	7.0	7.3
Kota Tinggi	4.9	4.9	5.0	5.2	5.6	5.6	5.9	6.1	6.4	6.7	6.9
Mersing	4.9	4.9	4.7	6.0	5.3	5.2	5.6	5.9	6.1	6.3	6.5
Muar	5.5	5.6	5.7	6.0	6.1	6.4	6.8	6.9	7.0	7.4	7.9
Pontian	5.3	5.4	5.7	5.8	6.0	6.1	6.3	6.6	6.8	7.0	7.4
Segamat	5.4	5.5	5.4	5.7	5.9	6.2	6.2	6.6	6.8	6.9	7.3
Kulai	5.2	5.3	5.2	5.5	5.7	5.9	6.3	6.4	6.7	7.0	7.3
Tangkak	5.5	5.5	5.7	5.8	6.2	6.2	6.6	6.9	6.9	7.1	7.5

e. The Percentage of Elderly Between Gender, in Johor, 2010-2020

The percentage of both gender elderly increase from 2010 to 2019 with male's are believe to be higher than females.

Percentage of Population Elderly in Johor from 2010 to Projection 2020



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Male (%)	88.3	92.8	98.6	104.5	111.0	117.7	124.7	131.9	139.4	147.3	155.6
	2.63%	2.71%	2.82%	2.94%	3.07%	3.21%	3.35%	3.50%	3.65%	3.80%	3.96%
Female (%)	88.5	90.7	94.3	98.0	102.3	106.7	111.6	116.8	122.2	128.1	136.5
	2.63%	2.65%	2.70%	2.76%	2.83%	2.91%	3.00%	3.10%	3.20%	3.31%	3.48%

4. Discussion and Conclusion

Johor will become an aging state in 2019 and the numbers expected to be increase year by year after. This Silver Tsunami will be affected on labour force, economic performances especially in productivity and certain activities, and social indicators such as health & welfare facilities. This condition may occur due to the decreasing of birth rate and long-life expectancy and an impact of migration injected by job opportunities offered in Johor.

Male's elderly are believe to be higher than females. For future study, it is recommended to investigate the aging population between the ethnic group in Johor. In addition, it also recommended to investigate the aging population in small area such as mukim in each district.

References

1. *Japan's demographic time bomb is getting more dire, and it's a bad omen for the country*- Jeremy Berke, Business Insider US
2. *The issue of Japan's Aging Population*- Dallin Jack, University of Chicago Law School
3. .*Impacts of Population Aging in Modern Japan and Possible Solutions for the Future*- Ritgerð til BA-prófs í Japönsku máli og menningu, Eggert Örn Sigurðsson
4. United Nations, Department of Economic and Social Affairs, Population Division (2017). *World Population Ageing2017* (ST/ESA/SER.A/408). Retrieved from http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Report.pdf
5. United Nations Population Fund (2012). *Ageing in the Twenty-First Century: A Celebration and A Challenge.* Retrieved from https://www.unfpa.org/sites/default/files/pub-pdf/Ageing%20report.pdf
6. United Nations, Department of Economic and Social Affairs, Population Division (2017). *World Population Prospects 2017 - Data Booklet* (ST/ESA/SER.A/401).
7. Samad, I.A & Mansor, N. (2013). Population Ageing and Social Protection in Malaysia. *Malaysia Journal of Economic Studies 50(2),* 139-156. Retrived from https://mjes.um.edu.my/article/view/2873
8. Tey, N.P., Siray, S., Kamaruzzaman, S.B., Chin, A.V, Tan, M.P., Sinnapan, G.S., & Muller, A.M. (2015). Aging in multi-ethnic Malaysia. *The Gerontological Society of America 56(4),* 603- 309. http://doi.org/10.1093/geront/gnv153.
9. Alfian, H. (2017). 9.6 Million Senior Citizens Expected In M'sia By 2050, Why This Is Worrying. Retrieved from http://www.malaysiandigest.com/features/705341-9-6million-senior-citizens-expected-in-m-sia-by-2050-why-this-is worrying.html?
10. Department of Statistics Malaysia (2015). Population Distribution and Basic Demographic Characteristic Report 2010. *Population & Demographic.*
11. Department of Statistics Malaysia (2017). Population and Demographic Ageing. *Population & Demography Division BPPD Newsletter.*
12. Karim, H.A. (1997). The Elderly in Malaysia: Demographic Trends. *Med J Malaysia 52(3).*

Efficiency analysis by combination of parametric and non- parametric approach: Evidence from Bursa Malaysia

Md Zobaer Hasan[1], Omar Sharif[2], Chang Yun Fah[3], Mahboobeh Zangeneh Sirdari[3]

[1] School of Science, Monash University Malaysia, Bandar Sunway, Malaysia

[2] Department of General Educational Development, Daffodil International University, Dhaka, Bangladesh

[3] Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Bandar Sungai Long, Selangor, Malaysia

Abstract

This study estimates the technical efficiency scores of the financial companies listed in Bursa Malaysia using the non-parametric approach (DEA), parametric approach (SFA) and the proposed approach, CDS (combination of DEA and SFA) over the period 2007-2016. In addition, regression analysis between efficiency and profit risk are utilized to find the most efficient approach among the three approaches: DEA, SFA and CDS. The result suggests that the most efficient model is CDS which has a significant positive correlation with profit risk. Based on the results of CDS, this study postulate that the most efficient company is ACSM (Aeon Credit Service M Bhd) with the efficiency score 0.9819 and the least efficient company is MAY (Malayan Banking Bhd) with the efficiency score 0.7693 in Bursa Malaysia. The findings of this research provide a benchmark for future studies about the comparison of different financial companies in Bursa Malaysia.

Keywords

Bursa Malaysia; Efficiency; DEA; SFA; CDS.

1. Introduction

Nowadays, it is very difficult to identify the most efficient company by observing only the stock price. Many techniques are applied by investors to optimize their return and minimize the risk of their investment (Saad et al., 2011). Particularly in policy making, the efficiency is important because it permits an efficient allotment of capital in different productive sectors in an economy (Hubbard, 2008).

There are many methods in the frontier analysis to evaluate performance such as parametric and non-parametric, stochastic method (Fenyves et al., 2015). The present article introduces a non-parametric method, DEA (Data Envelopment Analysis), a parametric approach SFA (Stochastic Frontier Analysis) and the combination of DEA and SFA (CDS).

SFA is become most frequently used procedure because it segregate statistical noise from the effect of inefficiency (Kumbhakar & Lovell, 2003). In spite of this, SFA speculates a distinct probability distribution for the efficiency level. The DEA skips this sorts of specification error and it does not need a prior production function for efficiency (Dong et al. 2014). In DEA, all deviations from the frontier are measured as inefficiency and it does not allow for random errors in the optimization which is its main drawback. Therefore, if any noise exists, this may exaggerate the common inefficiency. Consequently, two methods (DEA and SFA) have their advantages as well as drawbacks (Huang & Wang 2002). Many researchers (Casu et al. (2004), Delis and Papanikolaou (2009), Weill (2004)) found that the consistency of efficiency derived from DEA and SFA is not significant. For this reason, this study will also concentrate on finding the combination of the DEA and SFA efficiency scores which will be a new experiment in literature perspective. However, Fernandes et al. (2018) and Altunbas et al. (2007) found that there is a strong connection between efficiency and profit risk, because inefficient financial firm tend to take less risk by investing and hold more capital. Fernandes et al. (2018) found that profit risk has a positive effect on the efficiency of peripheral European domestic banks.

This study will be a new idea for the estimation of financial company's efficiency by using the combination of DEA and SFA in respect to developing counties like Malaysia. The study provides a unique setting to calculate financial efficiency matric and find the effect of efficiency on profit risk by using regression analysis. Moreover, these findings could provide useful and important signal in case of decision making for management.

2. Methodology

DEA-MPI

The best way to introduce DEA is via the ratio form. For each DMU (decision making unit) needs to obtain a measure of the ratio of all outputs over all inputs, such as $\theta = \frac{p_i y_{it}}{q_i x_{it}}$. Where, p_i is an Mx1 vector of output weight for i^{th} firm and q_i is a Kx1 vector of input weight of i^{th} firm. To select optimal weight we specify the mathematical programming problem:

$$Max_{\theta,\lambda}\theta$$

$$\left.\begin{aligned} &St, \\ &\sum_{j=1}^{n}\lambda_j x_{ij} \le x_{it}\ i=1,2,....m \\ &\sum_{j=1}^{n}\lambda_j y_{ij} \ge \theta y_{rt}\ r=1,2,....s \\ &\lambda \ge 0 \end{aligned}\right\} \quad (1)$$

Where, for each DMU s is output observation, m is input observation, r *is* s^{th} output, i is m^{th} input, y_{rt} is r^{th} output for time period t, x_{it} is i^{th} input for time period t, n is DMU observation, j is n^{th} DMU, λ is no-negative scalar, x_{ij} is m^{th} input for nth DMU, y_{ij} is s^{th} output for n^{th} DMU, θ is a scalar representing the value of efficiency score for each DMU. Similar method is applied for n^{th} DMUs. To measure the technical efficiency, the software DEAP version 2.1 is used.

SFA

The production theory is proposed by Cobb and Douglas (1928) and named "Cobb Douglass production theory". He develops the production theory by using of labour, capital, production, value, and wages for the manufacturing firms. In order to measure statistical noise, Aigner et al. (1977) added symmetric error term to the deterministic frontier. The model expressed as:

$$Y_{it} = X_{it}\,\beta + (V_{it} - U_{it}), i = 1,2, \ldots..N, t = 1, \ldots, T \quad (2)$$

where, Y_{it} is (the logarithm of) the production of the i^{th} firm in the t^{th} time period; X_{it} is a k×1 vector of (transformations of the) input quantities of the i^{th} firm in the t^{th} time period; V_{it} are random variables which are assumed to be iid $N\left(0, \sigma_i^2\right)$ and β is an vector of unknown parameters.

$$U_{it} = U_i e^{-\eta(t-T)} \quad (3)$$

where U_i are the inefficiency level of the i^{th} producer at time T and η is an unknown parameter. The term TE_{it} is express as technical efficiency for the i^{th} firm in the t^{th} time period define by using stochastic frontier model (2) as follows (Battese & Coelli, 1988):

$$TE_{it} = e^{-U_{it}} \quad (4)$$

Here, U_{it} is the stipulations of the inefficiency model in equation (3). The maximum-likelihood estimates are used to measure the parameters of the stochastic frontier model. The software package FRONTIER 4.1 of Coelli (1996) were used in order to carry out the SFA.

Empirical form of Stochastic Frontier Model

The Cobb-Douglas stochastic frontier production model's functional form is defined as:

$$\ln(\mathrm{ROE}_{it}) = \beta_0 + \beta_1 \ln(\mathrm{TV}_{it}) + \beta_2 \ln(\mathrm{DPS}_{it}) + \beta_3 \ln(\mathrm{MC}_{it}) + \beta_4 \ln(\mathrm{P}\,B_{it}) + \beta_5 \ln(\mathrm{FL}_{it}) + (V_{it} - U_{it}) \quad (5)$$

where, the subscripts t and i represents the t^{th} year and i^{th} company of the observations, and i = 1,2,….26; t = 1,2,…10. The five input variables are total volume (TV), dividend per share (DPS), market capital (MC), price to book ratio (PB) and financial leverage (FL). The output variable is return on equity ROE. "ln" represents the natural logarithm.

Combination of DEA and SFA (CDS)

The average of DEA and SFA efficiency scores are called the combination of DEA and SFA (CDS).

$$CDS = \frac{Efficiency\ score\ of\ DEA\ + Efficiency\ score\ of\ SFA}{2}$$

Linear Regression

In this study, the linear regression model is utilized to investigate the impact of profit risk on efficiency score (derived from DEA, SFA and CDS) in the financial sector of Bursa Malaysia. So, the three models take the following forms:

$$Ef(\mathrm{DEA})_i = \beta_0 + \beta_1 Pr_i + \varepsilon_i \quad (6)$$

$$Ef(\mathrm{SFA})_i = \beta_0 + \beta_1 Pr_i + \varepsilon_i \quad (7)$$

$$Ef(\mathrm{CDS})_i = \beta_0 + \beta_1 Pr_i + \varepsilon_i \quad (8)$$

where $Ef(\mathrm{DEA})_i$, $Ef(SFA)_i$, $Ef(\mathrm{CDS})_i$ are the average technical efficiency scores of the companies i derived from DEA, SFA and CDS respectively; Pr_i is the average profit risk of i^{th} company; β_0 is constant and represents the slope parameter; ε_i represents error term.

Data collection and Input and Output Variables

There are 30 listed financial companies in Bursa Malaysia. This study will concentrate on yearly balance data of 26 listed companies (Table 1). The sample is panel data which covers 26 financial companies listed in Bursa Malaysia over the period of 2007 to 2016. There are total 260 observations. Data are collected from Bloomberg. The input and output variables are selected based on Ismail et al. (2012) and others major studies on the efficiency of the financial sector.

Table 1. List of 26 Financial Companies

Company Name	Bursa Malaysia Short Name	Bloomberg Ticker Number (MK Equity)
Malayan Banking Bhd	MAYBANK	MAY
Public Bank Bhd	PBBANK	PBK
CIMB Group Holdings Bhd	CIMB	CIMB
Hong Leong Bank Bhd	HLBANK	HLBK
RHB Bank Bhd	RHBBANK	RHBBANK
Hong Leong Financial Group	HLFG	HLFG
AMMB Holdings Bhd	AMBANK	AMM
BIMB Holdings Bhd	BIMB	BIMB
Affin Holdings Bhd	AFFIN	AHB
LPI Capital Bhd	LPI	LPI
Syarikat Takaful Malaysia	TAKAFUL	STMB
Allianz Malaysia Bhd	ALLIANZ	ALLZ
MNRB Holdings Bhd	MNRB	MNRB
Manulife Holdings Bhd	MANULFE	MHBS
Pacific & Orient Bhd	P&O	PO
Malaysia Building Society	MBSB	MBS
Bursa Malaysia Bhd	BURSA	BURSA
Aeon Credit Service (M) Bhd	AEONCR	ACSM

INSAS Bhd	INSAS	INS
RCE Capital Bhd	RCECAP	RCE
Apex Equity Holdings Bhd	APEX	APX
Johan Holdings Bhd	JOHAN	JOH
ECM Libra Financial Group Bhd	ECM	ECML
Hong Leong Capital Bhd	HLCAP	HLG
TA Enterprise Bhd	TA	TAE
MAA Group Bhd	MAA	MAA
Source: Bloomberg terminal and Bursa Malaysia		

3. Result

Efficiency derived from DEA

It is seen that the average technical efficiency of financial companies listed in Bursa Malaysia was 0.8999 that means companies were less than 10% inefficient to use their existing resources. Moreover, Siew et al. (2018) found the average efficiency score was 0.5865 in the financial companies listed in Malaysia. The results of this study also depict that MAY bank was the least efficient company (0.703) and BIMB bank (0.8455) was the most efficient bank. Among the banks in Malaysia, Sufian et al. (2016) found that RHB was the most efficient bank (0.937) and the least efficient bank was WAH TAT bank (0.288). However, in this study the RHB bank's efficiency score was 0.739.

Efficiency derived from SFA

The average technical efficiency derived from SFA was 0.8809 which means that the financial companies listed in Bursa Malaysia were 12% efficiency behind to get maximum outputs from given inputs. The ACSM seemed to be more efficient in controlling efficiency, as the efficiency score stands at 0.9637. But, JOH was least efficient company as its efficiency score was 0.5857. Hasan et al. (2012) applied the SFA approach for finding the efficiency of the domestic banks listed in Bursa Malaysia over the period 2005–2010. He found that PBK (0.918) was the least efficient bank and RHBBANK (0.986) was the most efficient bank.

Combination of DEA and SFA (CDS)

The average technical efficiency derived by CDS was 0.8904, that means financial companies listed in Bursa Malaysia were 11% efficiency behind to get maximum outputs from given inputs. The ACSM seemed to be more efficient in controlling efficiency, as the efficiency score stands at 0.9819. Whereas, MAY was least efficient as its efficiency score 0.7693. Average efficiency of ALLZ, INS and HLG were same that was around 0.96.

Comparison of DEA, SFA and CDS Efficiency Scores

The empirical findings of efficiency scores are presented in figure 1. From the figure, it is clear that the DEA average efficiency score (0.8999) was greater than the SFA average efficiency score (0.8809). Moreover, CDS average

efficiency was 0.8904 that is greater than SFA. However, such types of differences are not surprising because SFA allows DMUs to depart from the frontier due to inefficiency as well as statistical noise. But, DEA method cannot measure statistical noise. These results coincide with the results of Sufian et al. (2016), Ismail (2005), Isik and Hasan (2002). This study examined that there was a lesser difference among efficiency scores of financial companies estimated by DEA, SFA and CDS (SFA scores < CDS scores < DEA scores). The study suggests that the three models tend to have limited continuity in selecting the most efficient and least efficient financial companies in terms of efficiency score.

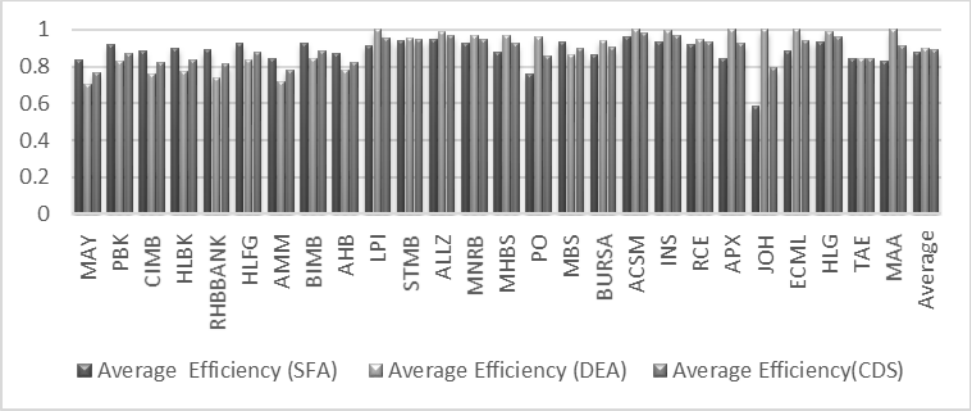


Figure 1: Efficiency derived from DEA, SFA and CDS

Regression Analysis between Efficiency (derived from DEA, SFA and CDS) and Profit Risk

From table 2, it is found that the linear regression relationship between efficiency and profit risk was statistically significant at the 5% level of significance by DEA (since the p-value was less than 0.05) and at 1% level of significance by CDS (since the p-value was less than 0.01). This depicts that the profit risk had positively affected the efficiency of the financial company listed in Bursa Malaysia. That means, more profitable financial company or less leverage company was higher efficient and would face a lesser cost of going insolvent over the period 2007 to 2016. In a study, Fernandes et al. (2018) applied the DEA method and also found that the profit risk positively affects the efficiency in European peripheral domestic banks. They found the coefficient score was 0.216. However, in this study the relation between SFA and profit risk was insignificant because its *p-value* was more than 0.05. Furthermore, its coefficient value was lowest (0.273) among the three methods. The coefficient value of CDS was 0.54 and that was the highest among the three methods. The result postulates that 1% increase in efficiency can increase the profit risk 0.54 %. Finally, from the regression results of three models, it can be concluded that the best way to measure efficiency is CDS.

Table 2: Regression Analysis between Efficiency (derived from DEA, SFA and CDS) and Profit Risk

Model	Constant (β_0)	Coefficients (β_1)	S.E	*p*-value
DEA	-1.765	0.461*	4.317	0.018
CDS	-10.591	0.54*	6.687	0.004
SFA	0.307	0.273@	6.386	0.177

* 5% significant, @ insignificant

4. Discussion and Conclusion

The study concentrates on three methods, SFA, DEA, and combination of DEA and SFA (CDS) on a sample of financial companies that are listed in Bursa Malaysia for finding most efficient method. The result shows that CDS has the most significant relationship with profit risk. However, all models present the unique conclusion that ACSM is the most efficient company. Additionally, this study finds most efficient method is CDS. All the companies' efficiency scores lies between 0.9819 and 0.7693. Considering no consistency on different efficiency scores across the different methods, this study will help to investigate the efficiency of the financial sector and other sectors of Bursa Malaysia.

References

1. Aigner, D., Lovell, C. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. Journal of econometrics, 6(1), 21-37.
2. Altunbas, Y., Carbo, S., Gardener, E. P., & Molyneux, P. (2007). Examining the relationships between capital, risk and efficiency in European banking. European Financial Management, 13(1), 49-70.
3. Battese, G. E., & Coelli, T. J. (1988). Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. Journal of econometrics, 38(3), 387399.
4. Cobb, C. and Douglas, P. H. (1928). A theory of production. American Economic Review, Vol. 18 No.1, 139–165.
5. Coelli, T., Rao, D.S.P. and Battese, G.E. (1998), An Introduction to Efficiency Analysis, Kluwer Academic Publishers, Boston.
6. Coelli, T. (1996). A guide to DEAP version 2.1: a data envelopment analysis (computer) program. Centre for Efficiency and Productivity Analysis, University of New England, Australia.
7. Coelli, T. J. (1996). A guide to FRONTIER version 4.1: a computer program for stochastic frontier production and cost function estimation (Vol. 7, pp. 1-33). CEPA Working papers.

8. Delis, M.D., Papanikolaou, N.I., (2009). Determinants of bank efficiency: evidence from a semi-parametric methodology. Managerial Finance 35 (3), 260–275.
9. Dong, Y., Hamilton, R., & Tippett, M. (2014). Cost efficiency of the Chinese banking sector: a comparison of stochastic frontier analysis and data envelopment analysis. Economic Modelling, 36, 298-308.
10. Fernandes, F. D. S., Stasinakis, C., & Bardarova, V. (2018). Two-stage DEA-Truncated Regression: Application in banking efficiency and financial development. Expert Systems with Applications, 96, 284-301.
11. Hasan, M. Z., Kamil, A. A., Mustafa, A., & Baten, M. A. (2012). A Cobb Douglas stochastic frontier model on measuring domestic bank efficiency in Malaysia. PLOS one, 7(8), e42215.
12. Huang, T. H., & Wang, M. H. (2002). Comparison of economic efficiency estimation methods: Parametric and non–parametric techniques. The Manchester School, 70(5), 682-709.
13. Hubbard, R. G. (2008). Money, the financial system, and the economy (Sixth Edition).US: Pearson Education.
14. Ismail, M. K. A., Rahman, N. M. N. A., Salamudin, N., & Kamaruddin, B. H. (2012). DEA portfolio selection in Malaysian stock market. In Innovation Management and Technology Research (ICIMTR), 2012 International Conference on (pp. 739-743). IEEE.
15. Ismail, M., (2005). A study of efficiency and competitive Behaviour of commercial Banks in Malaysia (Doctoral thesis). Retrieved from UMI Dissertation Publishing, UMI Number U584012.
16. Siew, L. W., Fai, L. K., & Hoe, L. W. (2018, April). Investigation on the Efficiency of Financial Companies in Malaysia with Data Envelopment Analysis Model. In Journal of Physics: Conference Series (Vol. 995, No. 1, p. 012021). IOP Publishing.
17. Weill, L. (2004). Measuring cost efficiency in European banking: A comparison of frontier techniques. Journal of Productivity Analysis, 21(2), 133-152.

The quality challenges In the administrative data

Mohammed Al Rifai, Maitha Mohammed Aljunaibi, Dunya Husain Al Khlaifi, Faisal Saeed Al Shamsi
Statistics Centre - Abu Dhabi, UAE

Abstract
Statistical data and indicators are considered basic decision-support pillars, for both government and the private sector, in addition to their role in research centres and universities. The provision of statistical data and indicators that meet international statistical standards and best practices - characterized by diversity, convenience and consistency - have become a major challenge for official statistical agencies. Beside the censuses and sample survey data, many national statistical agencies and international statistical organizations now recommend using administrative data collected by their countries' official institutions, wherever possible, as sources for the construction of statistical indicators. As a result, administrative data has become a key source of official statistics. This paper discusses the quality challenges in the administrative data, and the role of statistical agencies and official institutions in producing administrative data to achieve integration in data quality. Further, it discusses recommended guidelines that statistical agencies should adopt and institutions should use to ensure the required level of quality in administrative data.

Keywords
Data Quality; administrative data; official statistics

1. Introduction

As a result from the increasing demand for a wide variety of information and statistical data to serve decision-makers, planners, and researchers, statistical data has become a key requirement for building economic, social, demographic, and other indicators, to evaluate the performance of the governments and institutions, and prepare programs of economic and social development. Therefore, there is a need to focus on:

- Building official indicators based on statistical information issued by a strong, independent and reliable official statistical institution. This is an essential element for evaluating various achievements of the Millennium Development Goals.
- Classifying all official statistics according to the best international standards, to increase their comparability with developed countries.

- Building data sets (about individuals, households and establishments), consistent with each other, for trends analysis over periods of time, and to meet the confidentiality standards of raw data protection.
- Utilize the potential of all existing administrative data sources, with an emphasis on reducing the data collection burden on respondents.

In order to meet this increasing demand for official statistical data, official national statistical agencies have focused on:

- Building national statistical strategies and systems.
- Building partnerships with other national institutions in order to achieve integration in the statistical work at the national level.

Traditionally, statistical institutions have developed censuses and sample survey methodologies in order to build the required statistical databases that could keep pace with the needs of users. However, official statistical agencies face major challenges in maintaining these sources of statistical data, for example:

- The relatively high financial cost of implementing censuses and sample surveys.
- Providing the data collected through census and sample surveys to meet users' needs in a timely manner.
- Reducing the burden on the various types of respondents, and maintaining response rates, which affect the accuracy and efficiency of statistical data.

2. Methodology

As a way of effectively utilizing the available resources and, at the same time, relieving the response burden for the various groups of residents and businesses who provide data for official statistics, agencies are making more use of administrative data collected by other official national agencies. Censuses and sample surveys will be used mostly by those countries, where administrative data is unavailable or is unreliable.

For administrative data to be useful for official statistics, it must be collected so that it can provide statistical indicators with the following characteristics:

- Outputs at the level of administrative regions and small geographic areas.
- Outputs at the level of particular population units, such as individuals, families, and business establishments.
- Better coverage than a sample survey to reduce the sample error rates caused by survey non-response.
- Production costs that is lower in time and budget than the costs of censuses and sample surveys.

- Databases that can be integrated with sample surveys, and used in the imputation of missing values for sample surveys.
- Construct new sampling frames or update the available frames, for the design and selection survey samples.
- Regular time series for the proposed statistical variables.

3. Result

The implementation of the concept of quality is not limited to censuses and sample surveys, but also is a comprehensive concept with standards and conditions that apply to all official statistical databases, including administrative data. In fact, the process of achieving quality in administrative data may be more complicated and difficult than for sample surveys, because administrative databases are not usually constructed for the purpose of providing statistical indicators, but only to serve the administrative objective of the data-producing institutions.

Harmonization of administrative data between official statistics agencies' requirements and the requirements of the producing institutions is a big challenge. This challenge varies for different countries. Many countries have been working on harmonizing their administrative databases with their official statistics outputs for some time and have successful models, while other countries are still working through the challenges of conceptual harmonization, to develop and improve their administrative databases. The following are common challenges face the administrative data:

Relevance: Relevance is the degree to which administrative data meet the official statistics needs. A comprehensive evaluation could be implemented to see whether all statistics that are needed are produced and the extent to which concepts (definitions, classifications etc.) reflect user needs. It is important to note that, while the administrative data will be relevant for the users of the data-producing entity, the data may not be relevant for official statistics indicators.

Inclusion and Coverage: In some situations, administrative data bases may exclude variables that fall within the requirements of official statistics. Data-producing institutions are focused on collecting and producing data that serves their own purposes. This is one of the biggest challenges for official statistical agencies' use of administrative data. Coverage of statistical units can also be an obstacle. Statistical agencies require extensive coverage of the populations being measured in order to output statistical indicators for small geographic areas and administrative regions.

Accuracy: The accuracy of data is one of the key data quality dimensions. It is well known that the accuracy of the statistical indicator is inversely proportional to the amount of statistical error in its estimated value,

while the error is the amount of the difference between the actual and estimated value. In sample surveys, sample error can be measured statistically and controlled through the adoption of probability sampling methods. Non-sample error can be minimized by following optimal procedures throughout the various stages of survey implementation.

For administrative data, it may be difficult to optimize the organizational and technical procedures for constructing the databases in order to minimize qualitative measurement error because these procedures are controlled by the data-producing agency rather than the statistical agency. Thus, it may be difficult to accurately evaluate the quality of indicators sourced from administrative data.

Consistency of methodologies, statistical concepts and classifications: Official statistics agencies depend on the concepts, definitions, manuals, and statistical classifications recommended by the international statistical organizations for many categories of economic, social and demographic data. Using the administrative data as a source for official statistics requires consistent concepts, definitions, manuals, classifications, and appropriate methodologies to achieve integration between data from statistical surveys and administrative data, which maximize the benefit from the entire database in optimum way.

Timeliness: Lack of timeliness in the delivery of the data to the national statistical agencies is a very common problem. Some delays are inevitable whereas others are because of inefficiencies in processes, lack of priority given to statistical needs or insufficient resources.

The large size of administrative files can mean significant processing costs with a subsequent impact on timeliness.

Comparability: Administrative statistics should be consistent internally, over time and comparable between region and geographic areas. On other hand it should allow to combine and make joint use of related data from different sources.

Interpretability: The interpretability of statistical data reflects the availability of supplementary information and metadata. The provision of metadata is essential to enable public evaluation of the appropriateness of the data for various uses. Metadata also provides explanations of the procedures and methods of treatment that have been implemented in the construction of the data. A major challenge for statistical agencies that use administrative data as their source for statistical indicators is ensuring the adequacy of metadata about the administrative databases.

4. Conclusion

There are many challenges for both statistical agencies and data-producing institutions, when governments utilize administrative data for official statistics. The biggest challenge is creating the climate of co-operation between all the institutions involved to ensure the quality of the final outputs. Statistics agencies are responsible for ensuring the statistical aspects of the methodologies and technical processes follow international quality standards. In addition, they must work within their countries' legal environments to co-ordinate access to the administrative databases. The data-producing institutions are responsible for implementing the technical recommendations and guidelines, which will achieve the common interest for all parties. This will involve developing the statistical capacity among staff in the area of administrative data.

The recommended procedures and guidelines, which should be adopted by the statistical agencies and applied by the data-producing institutions, can be classified into two parts. The first is recommended technical guidelines, and the second is organizational guidelines.

A. Procedures and technical guidelines:

Statistical agencies must conduct comprehensive evaluations of the administrative databases to assess the following criteria:

- **Statistical coverage:** Identify whether the coverage of the data meets the required scope of statistical outputs, particularly for outputs for small geographic areas, such as administrative regions. This assessment needs to identify the gaps in the statistical coverage that limit the statistical usage to the data.
- **Data accuracy and consistency:** Various statistical methods can be used, including, the comparison of indicators resulting from these data with available indicators from other sources; or linking the variables in the database with each other to ensure that the data is homogenous.
- **Data collection methodologies:** The statistical agencies should study the methodologies, concepts, manuals and classifications used in the construction of administrative databases, to understand how consistent the collection of the administrative data is with optimum statistical methodologies; and whether the collection meets international statistical standards. Such assessments will result in statistical agencies providing technical support and techniques that fit the requirements of official statistics. Statistical agencies will need to justify and clarify the importance of international classifications and standards and highlight the benefits to both parties. The data-producing institutions need to co-operate in the implementation of the proposed methodologies and make adjustments wherever required.

- **Periodicity and regularity of time series:** Time series administrative data will need evaluations for regularity and periodicity; gaps will need to be identified; and technical solutions found for integrating administrative data time series with sample survey time series. Solutions will also need to be developed for maintaining continuous updates.
- **Continuous quality review:** Programs of follow-up data quality assessment procedures will need to be developed by statistics agencies, in coordination with the institutions producing data.

B. Organizational procedures:

- **Raising role awareness**: Statistical agencies will need to raise the awareness of the data-producing institutions about the importance of their roles in providing quality data as input to official statistics at the national level.
- **Obtaining commitment:** Data-producing institutions will need to be fully committed to producing quality data as input for official statistics in order to commit resources to implementing the technical procedures necessary to meet international standards.
- **Harmonization of technical and policy environments**: the IT systems, legal environment and institution policies will need to be harmonized in order to achieve the appropriate application of technical procedures which enable the administrative data flow.
- **Increasing electronic communication**: Electronic data-sharing facilities will need to be built to enable the statistical agencies to utilize the administrative data provided by other institutions. Building these facilities and systems may require the initiation of new technical projects, or the expansion of e-government initiatives.
- **Building statistical knowledge:** Human resources programs and policies will need to be developed for staff to share and exchange experiences between the statistical agencies and the data-producing institutions, in order to build the required knowledge about statistical methodologies and application of statistical methods.
- **Statistical co-ordination:** Statistical agencies will need to assign statisticians as co-coordinators in the data-producing institutions, to provide support and guidance in the application of methodologies and statistical methods.

References

1. Hand, David. J (2018). Statistical challenges of administrative and transaction data, In J. R. Statist. Soc.A 2018.
2. Lee, Hyunshik (2015). The Use of Administrative Data for Statistical Purposes, In Westat paper at Statistics Korea Symposium 2015 October 15, 2015.
3. McLennan, David (2018). Data quality issues in administrative data, ADRN Publication - February 2018.
4. Pronab Sen. (2009) Challenges of Using Administrative Data for Statistical Purposes: India Country, Paper https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2426
5. Statistics Canada (2015). The use of administrative data at Statistics Canada.
6. United Nations Fundamental Principles of Official Statistics (2015). Implementation guidelines.

The statistical confidence crisis in sport sciences: Is it all "shoddy statistics"?

Marijke Welvaert[1,2]
[1] Research Institute for Sport and Exercise, University of Canberra, Australia
[2] Australian Institute of Sport, Canberra, Australia

Abstract
After decades of gaining popularity within sport science, Hopkins' magnitude-based inference has been discredited as a statistical method. While its rationale, encouraging interpretation of magnitudes of effects rather than solely relying on p-values, is applaudable, the implementation has fundamental flaws and did not survive rigorous statistical review. The questions remains what is the impact of this polarised debate in sport science. A frequency analysis of recent sport science literature demonstrates that the majority of studies use traditional statistical inference. The analysis also shows that statistical practice within sport science could be updated to be more aligned with more modern statistical methods. A continued focus on interpreting evidence beyond p-values using validated techniques should bring this statistical confidence crisis to an end.

Keywords
Frequentist; Magnitude based inference; Effect size; Confidence intervals; Statistical practice

1. Introduction

Sport science is a multidisciplinary field that studies how the human body works during exercise and how sport and physical activity promotes health and performance. Recently, the field attracted the attention of the popular media following a second statistical review (Sanaini, 2018) of a common analysis method used in the field, namely "Magnitude-based Inference" (MBI; Hopkins et al. 2009). In 2015, Welsh and Knight were the first to publish a statistical review and both reviews agreed that while the rationale behind MBI should be encouraged (i.e. more focus on the magnitude of effects, rather than a significance driven interpretation), the implementation has fundamental flaws.

Despite world-renowned statisticians speaking out against MBI, proponents of the method persist in its validness and still encourage sport scientists to use it under a different name (i.e. reference Bayesian inference with a dispersed uniform prior) to avoid critique during the publication process (Hopkins & Batterham, 2018). On the other hand, high-ranked journals in the field (e.g. Medicine and Science in Sports and Exercise (MSSE),

the flagship journal of the American College of Sports Medicine) have changed their policy and no longer accept submissions using MBI stating it is not an acceptable method of statistical analysis (including labelling it as Bayesian). The result is a polarised debate that is predominantly held outside of peer-reviewed publications.

MBI gained popularity because it was providing the ultimate solution to two problems that are claimed to be common for sport scientists: (1) sport science studies have limited sample sizes because of practical limits in the availability of its population of interest (e.g. elite athletes), and (2) sport science studies are typically investigating small effects. As a net result, within a frequentist null hypothesis significance testing (NHST) analysis framework, this results in non-significant findings due to underpowered studies. MBI completely rejects using p-values as an inference method, and promotes interpreting magnitude of effects using effect sizes and confidence intervals. The latter statisticians will agree with, but the problem arises with the implementation of the method, in which probabilistic statements are asigned to confidence intervals (see Sanaini, 2018 and Welsh & Knight, 2015 for a thorough statistical review).

Going back to the original issues that MBI was aiming to address, the question remains whether small sample sizes are indeed as common as was suggested. Secondly, given that effect sizes and confidence intervals have been well researched within the field of statistics, why is there the impression that there is a need to create a "new" statistical method. This paper aims to provide insight in recent statistical practices in sport science and a rough estimation of the impact of MBI within the field. Based on those findings, suggestions for further statistical education within sport sciences with the goal of elevating statistical practice in the field are formulated.

2. Methodology

The statistical methods sections of all papers published in MSSE in 2018 were analysed. The journal published 329 papers in total, of which 291 were research articles. A further 36 papers were excluded because they did not report human subject data (e.g. animal studies, mechanistic modelling, meta-analyses, etc.). The remaining 255 articles were assessed on the following: study design (within-and/or between-subject; statistical analysis method (or family in case multiple techniques were used); statistical school (frequentist, bayesian or other); reports effect sizes (Yes-No); reports confidence intervals (Yes-No); and sample size.

Additionally, a keyword frequency analysis of the literature was performed using the Web of Science (WoS) database to provide an estimation of popularity of selected statistical methods within sport science. The search included all indexed publications within the research field from 1991 to 2018.

Number of hits for each search term per year were recorded. The search terms were: RM-ANOVA OR repeated measures ANOVA, linear mixed model, t test, magnitude-based inference, and bayesian.

3. Result

MSSE analysis

The journal publishes articles in 5 categories: Applied Sciences, Basic Sciences, Clinical Sciences, Epidemiology and Special Communications. Three quarters of the publications reported a within-subject design and the majority of publications (92.9%) utilised a frequentist inference approach. 3.5% reported MBI hybrid results (i.e. MBI alongside frequentist p values). 2.4% of the articles used a machine learning approach. A few single studies used standalone MBI, graphical analysis or confidence interval analysis.

The reported sample size was grouped in 6 categories: 10, 11-20, 21-100, 101-500, 501-2000, >2000 to differentiate between small sample studies and larger samples. Figure 1 shows the number of studies in each group per category. Overall, 36% of publications reports a sample size between 20 and 100 and 30% have a sample size larger than 100 participants. Of the remaining studies, 10% reports a sample size smaller than 10 and 24% have a sample size between 11 and 20 subjects. Figure 1 illustrates that those smaller samples are more common within the Applied Sciences.

The majority of studies reports either effect sizes (16.1%), confidence intervals (20.4%) or both (16.5%). However, 42.7% of publications reports neither. Figure 2 illustrates the use of effect sizes and confidence intervals in those studies with up to 20 subjects in the sample.

34.5% of the studies utilised repeated measures ANOVA, 16.5% reported a linear mixed model analysis, 14.5% reported a General Linear Model analysis and 9.0% reported using t tests for their statistical analysis. Other less frequently reported methods were Generalized Linear Model (4.3%), Cox Proportional Hazard model (4.3%), Magnitude-based inference (3.1%), Structural Equation Modelling (2.7%), non-parametric tests (1.6%), Random Forests (1.2%), MANOVA (1.2%) and correlations (1.2%).

Web of Science analysis

Figure 3 illustrated the trends of frequency of the selected keywords across publication years within the field of Sport Science. "t-tests" is the most frequently used with repeated measures ANOVA resulting in the second most hits. Bayesian, Linear mixed model and Magnitude-based inference demonstrate similar reporting rates.

4. Discussion and Conclusion

Magnitude-based inference became popular in Sport Science as a remedy towards underpowered studies and small studies. Recent statistical reviews

(Welsh & Knight, 2015; Sanaini, 2018) invalidated the method for not being founded in statistical principles and as a result Sport Science has been victimised for using "shoddy statistics". The data presented in this paper support a more nuanced picture. The Web of Science frequency analysis provides some evidence that MBI shows up in only a small proportion of publications, and definitely far less than more traditional frequentist methods. It should be noted though that the database search might not necessarily correlate with the actual usage statistics of the method as there were only 75 hits in the WoS search while the original MBI paper (Hopkins et al., 2009) has over 2,000 citations thus far.

Sample sizes within Sport Sciences as captured by the 2018 MSSE publications vary a lot. Interestingly though, when we combine the reported sample size with the usage of inference beyond p-values, using either effect sizes or confidence intervals or both, 60% of studies reported at least one of those. However, it were especially studies with a vary small sample (<8) that did not include any magnitude information for their inference. It could be argued that these studies in particular would benefit to supplement their analysis with effect sizes and/or confidence intervals.

Given the majority of studies utilising a within-subject design, it is not surprising that RM-ANOVA is a frequently chosen analysis technique. However, in the context of smaller samples, more complex designs and dealing with missing data, the linear mixed model (and by extension Bayesian hierarchical modelling) would provide a more powerful solution.

Statistical education within Sport Science should maintain a focus on the importance of magnitude-based inference as a concept but using statistically validated tools. By utilising effect sizes and confidence intervals, which are standardly available in statistical software packages, sport scientists can still draw conclusions from their data that are informative beyond statistical significance. In addition, promoting more modern methods that handle missing data would further enhance the ability to design powerful studies within the constraints of the population of interest.

References

1. Hopkins, W.G. and Batterham, A.M. (2018). The vindication of magnitude-based inference. Sportscience, 22, p. 19-29.
2. Hopkins, W.G., Marshall, S.W., Batterham, A.M., & Harlin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine and Science in Sports and Exercise, 41(1), p. 3-12.
3. Sanaini, K.L. (2018). The problem with "magnitude-based inference". Medicine and Science in Sports and Exercise, 50(10), p. 2166-2176.

4. Welsh, A.H. and Knight, E.J. (2015). "Magnitude-based inference": a statistical review. Medicine and Science in Sports and Exercise, 47(4), p. 874-884.

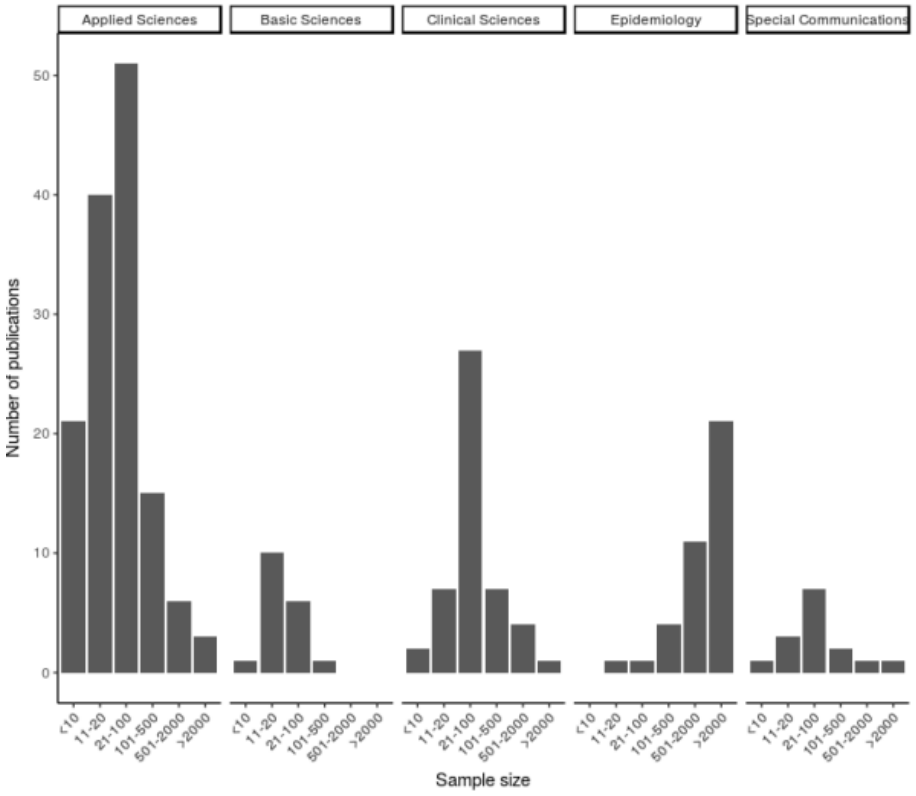


Figure 1: Number of research articles published in MSSE 2018 by journal category and sample size

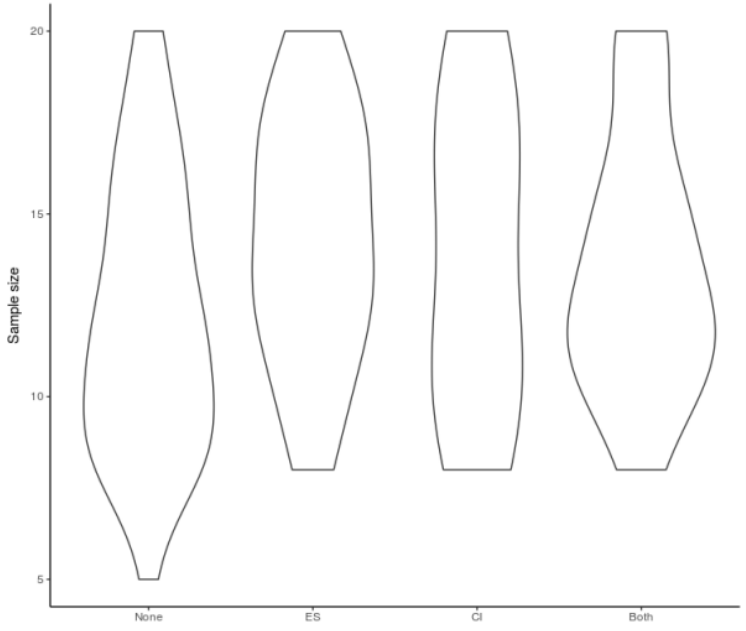


Figure 2: Violin plots illustrating the distribution of MSSE 2018 publications reporting effects sizes (ES) and/or confidence intervals (CI) by sample size

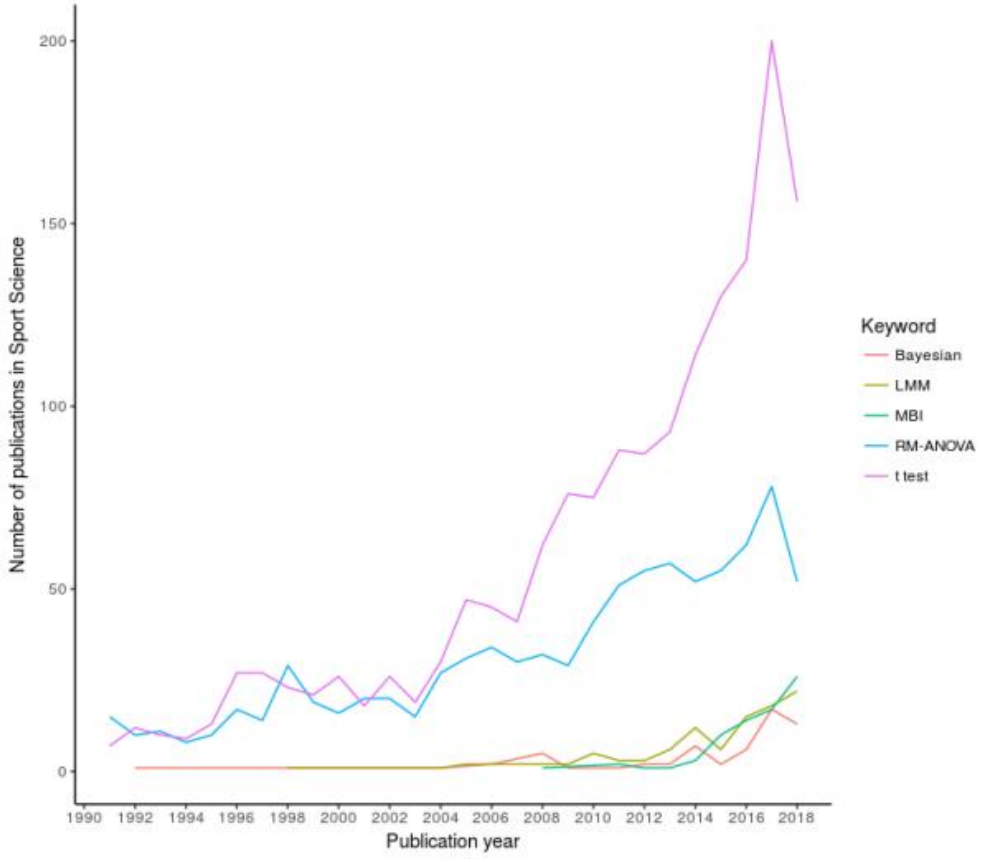


Figure 3: Publications trends in Sport Sciences based on a keyword search using the Web of Science database

Construction of forward looking distributions using limited historical data and scenario assessments

Mentje Gericke, Helgard Raubenheimer, PJ (Riaan) de Jongh
Centre for Business Mathematics and Informatics (BMI), North-West University

Abstract

Financial institutions are concerned about various forms of risk. The management of these institutions have to demonstrate to shareholders and regulators that they manage these risks in a pro-active way. Often the main risks are caused by losses that occur due to defaults on loan payments or by operations failing. In an attempt to quantify these risks, the estimation of extreme quantiles of loss distributions is of interest. Since financial companies have limited historical data available and need to provide a forward-looking view, they often use scenario assessments by experts to augment their historical data. This paper gives an exposition of a particular statistical approach that may be used to combine historical data and scenario assessments in order to estimate extreme quantiles.

Keywords

Loss distribution approach; scenario information; operational risk; economic capital; quantile estimation

1. Introduction

All financial losses need to be carefully managed and provided for by financial institutions. For example, banks are required by regulatory authorities to set aside capital to absorb unexpected losses. In addition, they also calculate economic capital, being the amount that a bank estimates it may need in order to remain solvent at a given confidence level and time horizon. The focus of this paper will be on operational risk in banks.

Financial institutions are more interested in the aggregate loss that may occur over one year in the future, than the individual losses in a particular area or business line. Popular modelling methods involve the construction of annual aggregate loss distributions using the so-called loss distribution approach (LDA). The constructed distribution may be used to answer questions like 'What aggregate loss level will be exceeded once in c years?' or 'If we want to guard ourselves against a one in a thousand year aggregate loss, how much capital should we hold next year?' The aggregate loss distribution and its quantiles provide answers to the above questions and therefore the distribution should be modelled as accurately as possible. It is often the extreme quantiles of this distribution that is of interest, for instance, the

regulator require that when dealing with operational losses, a bank should hold capital that will protect them against a one-in-a-thousand year aggregate loss. To determine the capital, the 99.9% Value-at-Risk (VaR) of the distribution is calculated. One would have hoped that at least a thousand years of historical data is available, but in reality only ten years of data may be available. For this reason, scenario assessments by experts are used to augment the historical data and to provide a forward looking view.

The proposed approach discussed in this paper has also been discussed and studied elsewhere (see de Jongh et al. (2015)), specifically in the context of operational risk and economic capital estimation.

2. Methodology

2.1 Approximating VaR

Let the random variable N denotes the annual number of loss events and assume that N is distributed according to a Poisson distribution with parameter lambda, i.e. $N \sim Po(\lambda)$. One could use other frequency distributions like the negative binomial distribution, but the Poisson is the most popular in practice (see e.g. Embrechts and Hofert (2011)). Furthermore, assume that the random variables $X_1, \dots, X_N$ denote the loss severities of individual loss events. Further, assume that these loss severities are independently and identically distributed according to a severity distribution , i.e. $X_1, \dots, X_N \sim iid\ T$. Then the annual aggregate loss is $A = \sum_{n=1}^{N} X_n$ and the distribution of A is a compound Poisson distribution that depends on λ and T and denoted by $CoP(T, \lambda)$. Of course, in practice we do not know T and λ and have to estimate it. First we have to decide on a model for T, for example a class of distributions $F(x, \theta)$. θ and λ have to be estimated using statistical estimates.

The compound Poisson distribution $Co(T, \lambda)$ and its VaR are difficult to calculate analytically so that in practice, Monte Carlo (MC) simulation is often used. This is done by generating N according to the assumed frequency distribution and then by generating $X_1, \dots, X_N$ independent and identically distributed according to the true severity distribution T and calculating $A = \sum_{n=1}^{N} X_n$. The previous process is repeated I times independently to obtain $A_i, i = 1,2, \dots, I$ and then the 99.9% VaR is approximated by $A_{([0.999*I]+1)}$ where A_i denotes the i-th order statistic and $[k]$ the largest integer contained in k. Note that three input items are required to perform it, namely the number of repetitions I and the frequency and loss severity distributions. The number of repetitions determines the accuracy of the approximation and the larger it is, the higher its accuracy.

In principle infinitely many repetitions are required to get the exact true VaR. The large number of simulation repetitions involved in the MC approaches above motivates the use of other numerical methods such as Panjer recursion, methods based on fast Fourier transforms (see e.g. Panjer,

(2006)) and the single loss approximation (SLA) method (see e.g. Böcker and Klüppelberg (2005)). For a detailed comparison of numerical approximation methods, the interested reader is referred to de Jongh et al. (2016).

2.2 Scenario modelling

The Basel Accord (see BCBS 196, (2011)) suggests the use of scenario assessments to improve severity distribution estimation. BCBS refers to three types of scenarios namely the individual scenario approach, the interval approach and the percentile approach. In the remainder of the paper we discuss a percentile approach suggested by de Jongh et al. (2015), which we believe is the most practical of the existing approaches available in the literature.

As discussed in de Jongh et al. (2015), we advocate the use of the so-called one-in-c year scenario approach. In the one-in-c years scenario approach, the experts are asked to answer the question: 'What loss level q_c is expected to be exceeded once every c years?'. Popular choices for c vary between 5 and 100 and often 3 values for c are used. As an example, one bank used c = 7, 20 and 100 and motivated the first choice as the number of years of historical data available to them. In this case the largest loss in the historical data may serve as a guide for choosing q_7 since this loss level has been reached once in 7 years. If the experts judge that the future will be better than the past, they may want to provide a lower assessment for q_7 than the largest loss experienced so far. If they foresee deterioration they may judge that a higher assessment is more appropriate. The other choices of c are selected in order to obtain a scenario spread within the range that one can expect reasonable improvement in accuracy from the experts' inputs. Of course the choice of c = 100 may be questionable because judgments on a one-inhundred years loss level are likely to fall outside many of the experts' experience. In the banking environment, they may take additional guidance from external data of similar banks which in effect amplifies the number of years for which historical data are available.

If the annual loss frequency is $Poi(\lambda)$ distributed and the true underlying severity distribution is T, and if the experts are of oracle quality in the sense of actually knowing λ and T, then the assessments provided should be

$$q_c = T^{-1}\left(1 - \frac{1}{c\lambda}\right). \quad (1)$$

To see this, let N_c denote the number of loss events experienced in c years and let M_c denote the number of these that are actually greater than q_c. Then $N_c \sim Poi(c\lambda)$ and the conditional distribution of M_c given N_c is binomial with parameters N_c and $1 - p_c = \mathrm{P}(X \geq q_c) = 1 - T(q_c)$ with $X \sim T$ and $p_c = T(q_c)$. Therefore $EM_c = E[E(M_c \mid N_c)] = E[N_c(1 - p_c)] = c\lambda(1 - T(q_c))$. Requiring that EM_c=1, yields (1) and $p_c = 1 - \frac{1}{c\lambda}$.

As illustration of the complexity of the experts' task, take λ = 50 then $q_7 = T^{-1}(0.99714)$, $q_{20} = T^{-1}(0.999)$ and $q_{100} = T^{-1}(0.9998)$ which implies that the quantiles that have to be estimated are very extreme.

2.3 Estimating VaR

In the previous section we discussed a way of modelling the true severity distribution T. The estimation of the 99.9% VaR of the aggregate loss distribution is of interest and de Jongh et al. (2015) discuss three approaches to estimate it, namely the naïve approach, the generalized Pareto distribution (GPD) approach and Venter's approach. The naïve approach make use of historical data only, whereas the GPD approach (which is based on a mixed model formulation) and Venter's approach make use of both historical data and scenario assessments. Below we focus our discussion on the GPD and Venter approaches and in Section 3 we demonstrate that, as far as estimating VaR is concerned, that Venter's approach is preferred.

2.4 The GPD approach

Select a number b and let q_b be the corresponding quantile given by (1), i.e., $q_b = T^{-1}(1 - \frac{1}{b\lambda})$. We use q_b as a threshold that splice T in such a way that the interval below q_b is the expected part and the interval above q_b the unexpected part of the severity distribution. Define two distribution functions

$$T_e(x) = T(x)/T(q_b) \text{ for } x \leq q_b \text{ and}$$

$$T_u(x) = [T(x) - T(q_b)]/[1 - T(q_b)] \text{ for } x > q_b,$$

i.e. $Te\ (x)$ is the conditional distribution function of a random loss $X \sim T$ given that $X \leq q_b$ and $T_u(x)$ is the conditional distribution function given that $X > q_b$. Note that we then have the identity

$$T(x) = \mathrm{T}(q_b)T_e(x) + [1 - T(q_b)]T_u(x) \text{ for all } x. \qquad (2)$$

This identity represents $T(x)$ as a mixture of the two conditional distributions. Instead of modelling (x) with a class of distributions $F(x, \theta)$ we may now consider modelling $T_e(x)$ with $F_e(x, \theta)$ and $T_u(x)$, with $F_u(x, \theta)$. Borrowing from extreme value theory, a popular choice for $F_u(x, \theta)$ could be the GPD, while a host of choices are available for $F_e(x, \theta)$, the obvious being the empirical distribution.

Suppose we have available a years of historical loss data $x_1, x_2, \ldots, x_k$ and scenario assessments $\tilde{q}_7, \tilde{q}_{20}$ and $\tilde{q}_{100}$. Then the annual frequency λ can be estimated as $\hat{\lambda} = K / a$. Next b and the threshold q_b must be specified. One possibility is to take b as the smallest of the scenario c-year multiples and to estimate q_b as the corresponding smallest of the scenario assessments $\tilde{q}_b$ provided by the scenario makers, in this case $\tilde{q}_7$. $T_e(x)$ can be estimated by fitting a parametric family $F_e(x, \theta)$ (such as the Burr) to the data $x_1, x_2, \ldots, x_k$ or by calculating the empirical distribution and then conditioning it to the interval $(0, \tilde{q}_b]$. We denote it by $\widetilde{F_e}(x)$. For the sake of future notational

consistency, we put tildes on all estimates of distribution functions which involve use of the scenario assessments.

Next, $F_u(x)$ can be modelled by the $GPD(x;\ \sigma, \xi, q_b)$ distribution. For ease of explanation, suppose we have actual scenario assessments $\tilde{q}_7, \tilde{q}_{20}$ and $\tilde{q}_{100}$ and thus take $b = 7$ and estimate q_b by $\tilde{q}_7$. Substituting these scenario assessments into $F_u(q_c) = 1 - \frac{b}{c}$; with $b = 7$, $c = 20$, 100 yields two equations

$$F_u(\tilde{q}_{20}) = \text{GDP}(\tilde{q}_{20}; \sigma, \xi, \tilde{q}_7) = 0.65 \text{ and } F_u(\tilde{q}_{100}) = \text{GDP}(\tilde{q}_{100}; \sigma, \xi, \tilde{q}_7) = 0.93 \quad (3)$$

that can be solved to obtain estimates of the parameters σ and ξ in the GPD that are based on the scenario assessments. Some algebra shows that a solution exists only if $\frac{\tilde{q}_{100} - \tilde{q}_7}{\tilde{q}_{20} - \tilde{q}_7} > 2.533$. This fact should be borne in mind when the experts do their assessments.

With more than three scenario assessments, fitting techniques can be based on (3) which links the quantiles of the GPD to the scenario assessments. An example would be to minimize $\sum_c |\text{GDP}(\tilde{q}_c; \sigma, \xi, \tilde{q}_7) - (1 - b/c)|$ .Other possibilities include a weighted version of the sum of deviations in this expression or deviation measures comparing the GPD quantiles directly to the q_c assessments. Whichever route we follow, we denote the final estimate of $F_u(x)$ by $\tilde{F}_u(x)$ All these ingredients can now be substituted into (2) to yield the estimate $\tilde{F}(x)$ of $T(x)$, namely

$$\hat{\lambda}\tilde{F}(x) = (\hat{\lambda} - \frac{1}{7})\,\tilde{F}_e(x) + \frac{1}{7}\,\tilde{F}_u(x). \quad (4)$$

When using the GPD 1-in-c years integration approach to model the severity distribution in the LDA, we realised that the 99.9% VaR of the aggregate distribution is almost exclusively determined by the scenario assessments and their reliability greatly affects the reliability of the VaR estimate. The fitted distribution has little effect on the VaR estimates. However, if the assessments are supplied by experts and not oracles, this may render undesirable results. The challenge is therefore to find a way of integrating the historical data and scenario assessments such that both sets of information are adequately utilised in the process.

2.5 Venter's approach

A retired colleague, Hennie Venter suggested that, given the quantiles q_7, q_{20}, q_{100}, one may write the distribution function T as the following identity:

$$T(x) = p_7 T_e(x) + [p_{20} - p_7]T_{u_1}(x) + [p_{100} - p_{20}]T_{u_2}(x) + [1 - p_{100}]T_{u_3}(x) \quad \text{for all } x, \quad (5)$$

where:

$$T_e(x) = T(x)/T(q_7) \text{ for } x \le q_7,$$
$$T_{u_1}(x) = [T(x) - T(q_7)]/[T(q_{20}) - T(q_7)] \text{ for } q_7 < x \le q_{20},$$
$$T_{u_2}(x) = [T(x) - T(q_{20})]/[T(q_{100}) - T(q_{20})] \text{ for } q_{20} < x \le q_{100}, \text{ and}$$
$$T_{u_3}(x) = [T(x) - T(q_{100})]/[1 - T(q_{100})] \text{ for } q_{20} < x \le q_{100}.$$

As was the case in (4), this is clearly a mixed distribution where $T_e(x)$ is the conditional distribution function of a random loss $X \sim T$ given that $X \le$

q_7 , $T_{u1}(x)$ the conditional distribution function given that $q_7 < X \leq q_{20}$, $T_{u2}(x)$ the conditional distribution function given that $q_{20} < X \leq q_{100}$ and $T_{u3}(x)$ is the conditional distribution function given that $X > q_{100}$. Equivalently, we can write (5) as follows:

$$T(x) = \begin{cases} R(7)T(x) & for\ x \leq q_7 \\ p_7 + R(7,20)[T(x) - T(q_7)] & for\ q_7 < x \leq q_{20} \\ p_{20} + R(20,100)[T(x) - T(q_{20})] & for\ q_{20} < x \leq q_{100} \\ p_{100} + R(100)[T(x) - T(q_{100})] & for\ q_{100} < x < \infty. \end{cases} \quad (6)$$

where:

$$R(7) = p_7/T(q_7), \quad R(7,20) = [p_{20} - p_7]/[T(q_{20}) - T(q_7)],$$
$$R(20,100) = [p_{100} - p_{20}]/[T(q_{100}) - T(q_{20})], \text{and } R(100) = [1 - p_{100}]/[1 - T(q_{100})].$$

Again $T(q_c) = p_c = 1 - \frac{1}{c\lambda}$ and it should be clear that the expressions on the right reduces to $T(x)$ and all the R ratios are equal to 1. Also, the definition of $T(x)$ could easily be extended for more quantiles. Given the previous discussion we can model (x) by $F(x, \theta)$ and estimate it by $F(x, \hat{\theta})$ using the historical data and maximum likelihood, and estimate the annual frequency by $\hat{\lambda} = K/a$. Given scenario assessments $\tilde{q}_7$, $\tilde{q}_{20}$ and $\tilde{q}_{100}$, then $T(q_c)$) can be estimated by $F(\tilde{q}_c, \hat{\theta})$ and b_c by $\hat{p}_c = 1 - \frac{1}{c\hat{\lambda}}$. The estimated R ratios are then

$$\tilde{R}(7) = \frac{\hat{p}_7}{F(\tilde{q}_7; \hat{\theta})}, \quad \tilde{R}(7,20) = \frac{\hat{p}_{20} - \hat{p}_7}{F(\tilde{q}_{20}; \hat{\theta}) - F(\tilde{q}_7; \hat{\theta})},$$
$$\tilde{R}(20,100) = \frac{\hat{p}_{100} - \hat{p}_{20}}{F(\tilde{q}_{100}; \hat{\theta}) - F(\tilde{q}_{20}; \hat{\theta})} \text{ and } \tilde{R}(100) = \frac{1 - \hat{p}_{100}}{1 - F(\tilde{q}_{100}; \hat{\theta})}. \quad (7)$$

Notice that if our estimates were actually exactly equal to what they are estimating, these ratios would all be equal to 1. With the formulation in (6) the true severity distribution function T may now be estimated by $\tilde{H}$ as follows (see de Jongh et al. 2015):

$$\tilde{H}(x) = \begin{cases} \tilde{R}(7)F(x; \hat{\theta}) & for\ x \leq \tilde{q}_7 \\ \hat{p}_7 + \tilde{R}(7,20)[F(x; \hat{\theta}) - F(\tilde{q}_7; \hat{\theta})] & for\ \tilde{q}_7 < x \leq \tilde{q}_{20} \\ \hat{p}_{20} + \tilde{R}(20,100)[F(x; \hat{\theta}) - F(\tilde{q}_{20}; \hat{\theta})] & for\ \tilde{q}_{20} < x \leq \tilde{q}_{100} \\ \hat{p}_{100} + \tilde{R}(100)[F(x; \hat{\theta}) - F(\tilde{q}_{100}; \hat{\theta})] & for\ \tilde{q}_{100} < x < \infty. \end{cases} \quad (8)$$

Also note that $\tilde{H}(\tilde{q}_7) = \hat{p}_7$, $\tilde{H}(\tilde{q}_{20})$ and $\tilde{H}(\tilde{q}_{100}) = \hat{p}_{100}$, i.e. the equivalents of $T(q_c) = p_c$ hold for the scenario assessments when estimates are substituted for the true unknowns. Hence at the estimation level the scenario assessments are consistent with the probability requirements expressed. Thus this new estimated severity distribution estimate $\tilde{H}$ 'believes' the scenario quantile information, but follows the distribution fitted on the historical data to the left of, within and to the right of the scenario intervals. The ratios $\tilde{R}(7)$, $\tilde{R}(7,20)$, $\tilde{R}(20,100)$ and $\tilde{R}(100)$ in (7) can be viewed as measures of agreement between the historical data and the scenario assessments and could be useful for assessing their validities and qualities.

3. Result: GPD and Venter model comparison

We conducted a simulation study to investigate the effect of the two approaches by perturbing the quantiles of the true underlying severity distributions. We assumed a different extreme value index (EVI) for the true underlying severity distributions and then perturbed the quantiles in the following way. For each simulation run, we chose three perturbation factors u_7, u_{20}, u_{20} and u_{100} independently and uniformly distributed over the interval $[1 - \epsilon, 1 + \epsilon]$ and then tentatively took $\tilde{q}_7 = u_7 q_7, \tilde{q}_{20} = u_{20} q_{20}$ and $\tilde{q}_{100} = u_{100} q_{100}$ but truncated these so that the final values are increasing, i.e. $\tilde{q}_7 \leq \tilde{q}_{20} \leq \tilde{q}_{100}$. Here the fraction ϵ expresses the size or extent of the possible deviations (or mistakes) inherent in the scenario assessments. If $\epsilon = 0$ then the assessments are completely correct (within the simulation context) and the scenario makers are in effect oracles. We chose the values 0, 0.1, 0.2, 0.3 and 0.4 for this purpose in the results below. Choosing the perturbation factors to be uniformly distributed over the interval $[1 - \epsilon, 1 + \epsilon]$ implies that on average they have the value 1, i.e. the scenario assessments are about unbiased.

We assumed a Burr distribution as our true underlying severity distribution. For each combination of parameters of the assumed true underlying Poisson frequency and Burr severity distributions and for each choice of the perturbation size parameter ϵ, the following steps were followed:

(a) The VaR approximation algorithm in Section 2.1 was used to determine the 99.9% VaR. Note that the value obtained here approximately equals the true 99.9% VaR; the only approximation involved is that it is based on 1 million repetitions rather than infinitely many. We refer to this value as the approximately true (AT) VaR.

(b) We generated a data set of historical losses, i.e. generate $K \sim Poi(7\lambda)$ and then generated $x_1, x_2, \ldots, x_K \sim iid$ Burr Type XII with choice of parameters. Here the family $F(x, \theta)$ is chosen as the Burr Type XII but it is refitted to the generated historical data to estimate the parameters as required.

(c) We added to the historical losses three scenarios $\tilde{q}_7, \tilde{q}_{20}, \tilde{q}_{100}$ generated by the quantile perturbation scheme explained above. We then estimated the 99.9% VaR using the GPD approach.

(d) We used the historical losses and the three scenarios of item (c), calculated the severity distribution estimate $\tilde{H}$ and applied Venter's approach to estimate the 99.9% VaR.

(e) We repeated items (a)-(d) 1000 times and then summarised and compared the resulting VaR estimates.

Because we are generally dealing with positively skewed data here, we shall use the median as the principal summary measure. Denote the median of the 1000 AT values by MedAT. Then we constructed 90% VaR bands for the

1000 repeated GPD and Venter VaR estimates, i.e. $\left[\frac{VaR_{(51)}}{Median(VaR_1,\ldots,VaR_{1000})} - 1, \frac{VaR_{(951)}}{Median(VaR_1,\ldots,VaR_{1000})} - 1\right]$. The results are given in Figure 1 overleaf. Note that light grey represents the GPD band and dark grey the Venter band, while the overlap between the two bands are even more dark.

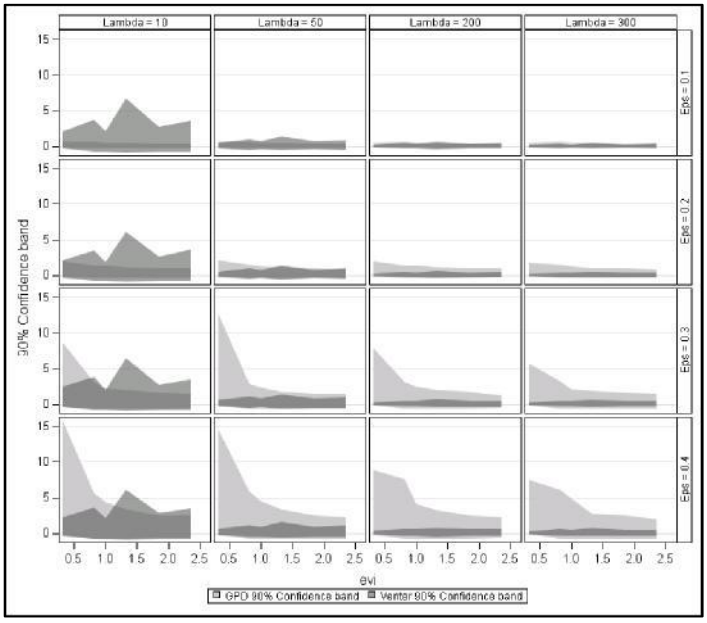


Figure 1: VaR bands for different Burr parameter sets and frequency combinations.

From Figure 1 we make the following observations:

- For small frequencies ($\lambda \leq 10$) the GPD approach outperforms the Venter approach, except for short tailed severity distributions and higher quantile perturbations.
- When the annual frequency is high ($\lambda \geq 50$) and for moderate to high quantile perturbations ($\epsilon \geq 0.2$) the Venter approach is superior, and more so for higher λ and ϵ.
- Even for small quantile perturbations ($\epsilon = 0.1$) and high annual frequencies ($\lambda \geq 50$) the Venter approach performs reasonable when compared to the GPD. The above information suggest that provided enough loss data is available, the Venter approach is the best choice.

4. Discussion and Conclusion

In this paper we motivated the use of Venter's approach whereby the severity distribution may be estimated using historical data and experts' scenario assessments jointly. The way in which historical data and scenario assessments are integrated incorporates measures of agreement between these data sources, which can be used to evaluate the quality of both.

References

1. BCBS 196. (2011). "Operational Risk: Supervisory guidelines for the advanced measurement approaches; Issued June 2011." Basel Committee on Banking Supervision, Bank of International Settlements.
2. Böcker, K., and Klüppelberg, C. (2005). "Operational VaR: a closed-form approximation." Risk 18(12), 90–93.
3. de Jongh, PJ, de Wet, T, Panman, K and Raubenheimer, H. (2016). "A simulation comparison of quantile approximation techniques for compound distributions in Operational Risk." Journal of Operational Risk, 11(1):23-48.
4. de Jongh, PJ, de Wet, T, Raubenheimer, H and Venter, JH. (2015). "Combining scenario and historical data in the loss distribution approach: A new procedure that incorporates measures of agreement between scenarios and historical data." Journal of Operational Risk, 10(1):1- 31.
5. Embrechts, P. and Hofert, M. (2011). "Practices and issues in operational risk modelling under Basel II." Lithuanian Mathematical Journal, 51(2):180–193
6. Panjer, H. (2006). Operational Risk: Modeling Analytics. Wiley, Chichester.

Construction of a survival tree based on concordance probability

Asanao Shimokawa, Etsuo Miyaoka
Tokyo University of Science, Tokyo, Japan

Abstract

Survival tree is one of the popular analysis method for time-to-event data in the field of medical research. It is well known that setting of the splitting criterion to construct the tree model is especially important in the analysis. Various authors have proposed several criterions. For example, Log-rank test statistics, exponential log-likelihood loss, and residual-based methods are used. In this study, we consider the concordance probability-based splitting criterions for constructing a survival tree. Concordance probability is one of the measure for prediction accuracy of the survival model. We propose the new method to construct the tree model that maximizes prediction accuracy based on the classification and regression tree algorithm. We study the performance of the splitting ability of the criterion based on concordance probabilities, and compare the survival trees constructed by proposed method and conventional methods through simulations.

Keywords

CART; C-index; Prediction accuracy; Tree structure

1. Introduction

In the medical research, analysis of time-to-event data is an important subject. In order to handle a regression problem that includes censored data based on covariates, the Cox proportional hazard model Cox (1972) has been most widely used. In addition to the simpleness of inference, this semi-parametric model has an advantage in that it can easily describe the covariate effects. However, this model requires proportional hazard assumptions, and certain assumptions about the relationship between covariates and response variables. Moreover, when this model includes many covariates, interpretation is difficult. In this study, we deal with survival trees, which involve constructing a tree structure model based on covariates. Because the proposed method uses a hierarchical structure, the relationship between covariates and hazards can be determined easily. Moreover, it is easy to incorporate a new patient into the model.

One of the most widely used method for constructing a survival tree is the classification and regression tree (CART) algorithm, proposed by Breiman et al. (1984), that is composed of three steps: splitting, pruning, and selection. Samples are recursively dichotomized in the splitting step, and a maximum

size tree is constructed. Various authors have proposed several criterions, and essentially these criterions divided into two types. One is the minimization of the risk within the node, and the other is the maximization of the degree of separation between nodes. For example, Log-rank test statistics is widely used (Leblanc and Crowley (1993)). The maximum size tree obtained by the splitting step suffers from an overfitting problem. To handle this problem, a set of nested subtrees is produced from the maximum size tree in the pruning step. In the selection step, the optimal size tree is selected by cross-validation or bootstrap method.

In this study, we consider the concordance probability-based splitting criterions for constructing a survival tree. The area under the curve of the receiver operating characteristic curves is widely used to evaluate the prediction accuracy of the model for binary outcome, and it is relevant to Kendall's tau and Mann-Whitney U test statistics. In survival data case, this idea is inherited by concordance probability and it is used to evaluate the prediction accuracy of the model. We use the four measures which evaluate the concordance probabilities as the splitting criterions: Harrell's C (Harrell et al(1996)), Uno's approach (Uno et al. (2011)), Begg's approach (Begg et al. (2000)), and Korn and Simon's approach (Korn and Simon (1990)).

In the Schmid et al. (2016), it has been proposed that Harrell's C is used as the splitting criterion to construct a random forest. In their research, maximum size trees are constructed using Harrell's C from bootstrap samples, and then the trees are aggregated to construct a forest. In this research, we propose the pruning and selection methods to construct a tree model based on the concordant measures. We study the performance of the splitting abilities of the criterions based on these measures, and compare the survival trees constructed by these criterions and conventional criterions through simulations.

The remainder of this paper is organized as follows. In Section 2, we introduce the method to construct a survival tree based on the measures for concordance probabilities. In Section 3, the results of the simulation studies are described. Finally, in Section 4, we present the conclusions of this paper.

2. Methodology

1. Concordance probability

Let U_i and D_i be the true failure and censoring time for subject i, respectively. Then, we can observe the time $X_i = \min(U_i, D_i)$. Let $\delta_i = I(X_i = U_i)$ be the event indicator for i, which is 1 if the observation experience an event and 0 if the observation is censored. Let $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{ip})$ denote p dimensional covariate vector for i. Then, an observed sample is represented by $\mathcal{L} = \{(x_i, \delta_i, z_i); i = 1, \cdots, n\}$.

Let $H_i = g(\boldsymbol{Z}_i)$ be the estimated risk score for the subject with Z_i, and we assume the high value of H_i represents the high risk of the event. There is a lot of measures to quantify how well the score H_i predicts the distribution of the failure time, like explained variation-based measures and discrimination-based measures. The C-index is one of the representative measure among them. Let $S_i(\mathrm{u})$ be the conditional survival probability for i:

$$S_i(u) = \Pr(U_i > u | H_i = \eta_i).$$

For two independent subjects (U_i,H_i) and (U_j,H_j), if the following inequalities is satisfied, the pair is said to be concordant at u:

$$U_i < U_j \text{ and } \hat{S}_i(u) < \hat{S}_j(u)$$

or

$$U_i < U_j \text{ and } \hat{S}_i(u) < \hat{S}_j(u)\ .$$

In a same meaning, the pair is said to be concordant if the following is satisfied:

$$U_i < U_j \text{ and } H_i > H_j$$

or

$$U_i > U_j \text{ and } H_i < H_j.$$

Based on the concordant pairs, the concordance probability is defined as follows:

$$C = \Pr(\mathrm{H}_i > \mathrm{H}_j | U_i < U_j, U_i < \tau),$$

where τ is the max time point to evaluate C. When the event time may be censored, the estimation of C is not simple. Harrell et al. (1996) proposed the measure that is derived as Kendall's tau for censored survival data. The measure uses only usable pairs in the observed sample:

$$\hat{C}_H = \frac{\sum_{i,j} \delta_i I\big(x_i < x_j, x_i \le \tau\big) I(\eta_i > \eta_j)}{\sum_{i,j} \delta_i I\big(x_i < x_j, x_i \le \tau\big)},$$

where $\sum_{i,j} = \sum_i \sum_{j \neq i}$. This measure becomes the consistent estimator if there are no censoring. However, it may depend on the censoring distribution. Therefore, it is no clear what the effect of censoring.

As the modification of the censoring bias of Harrell's C-index, Uno et al. (2011) proposed a new measure:

$$\hat{C}_U = \frac{\sum_{i,j} \delta_i \big\{\hat{G}(x_i)\big\}^{-2} I\big(x_i < x_j, x_i \le \tau\big) I(\eta_i > \eta_j)}{\sum_{i,j} \delta_i \big\{\hat{G}(x_i)\big\}^{-2} I\big(x_i < x_j, x_i \le \tau\big)},$$

where $\hat{G}(.)$ is the Kaplan-Meier estimator for the censoring distribution $G(d) = \Pr(D > d)$.

As the other modification of the bias of Harrell's C-index, Begg et al. (2000) proposed the measure that uses the all pairs including the unusable pairs. In their approach, the concordance values for unusable pairs are replaced by estimates of conditional expectation of it:

$$\hat{C}_B = \frac{2}{n(n-1)} \sum_{i,j} I(\eta_i > \eta_j) p_{ij},$$

where p_{ij} is defined as follows:

If $\delta_i = \delta_j = 1$, then

$$P_{ij} = \begin{cases} 0, & x_i > x_j \\ 1, & x_i < x_j \end{cases}.$$

If $\delta_i = 0, \delta_j = 1$, then

$$p_{ij} = \widehat{\Pr}(U_i < x_j | U_i > x_i) = \begin{cases} 0, & x_i > x_j \\ 1 - \dfrac{\hat{S}_l(x_j)}{\hat{S}_l(x_i)}, & x_i < x_j \end{cases}.$$

If $\delta_i = 1, \delta_j = 0$, then

$$P_{ij} = \widehat{\Pr}(U_j > x_i | U_j > x_j) = \begin{cases} \dfrac{\hat{S}_J(x_i)}{\hat{S}_J(x_j)}, & x_i > x_j \\ 1, & x_i < x_j \end{cases}.$$

If $\delta_i = \delta_j = 0$, then

$$P_{ij} = \widehat{\Pr}(U_i < U_j | U_i > x_{j,} U_j > x_j) = \begin{cases} \dfrac{1}{2}\dfrac{\hat{S}_J(x_i)}{\hat{S}_J(x_j)}, & x_i > x_j \\ 1 - \dfrac{\hat{S}_l(x_j)}{2\hat{S}_l(x_i)}, & x_i < x_j \end{cases}$$

For p_{ij} in the case of $\delta_i = \delta_j = 0$, it is assumed that if the subject with the shorter censored value of U lives as long as the time in paired subject, the remaining conditional probability of concordance is 1/2.

Korn and Simon (1990) proposed the measure based on the rank correlation between observed and predicted survival times:

$$\hat{C}_K = \frac{2}{n(n-1)} \sum_{i,j} I(\eta_i > \eta_j) p_{ij},$$

where

$$p_{ij} = \widehat{\Pr}(U_i < U_j) = \sum_{x_j^* \le \tau} [1 - \hat{S}_i(x_j^{*-})][\hat{S}_j(x_j^{*-}) - \hat{S}_j(x_j^*)] + [1 - \hat{S}_i(\tau^-)]\hat{S}_j(\tau^-),$$

$x_1^*, x_2^*, \cdots$ are the ascending-ordered event times, and x^-represents just before x.

2.2 Splitting criteria based on measures for concordance probability

We define a tree-structured model as T. The tree-structured model is constructed by the splitting rules of the covariate space and the nodes that are subsets of the resulting spaces. Let t be a node in tree T. If the node does not exist in the bottom layer of the tree, we call it an internal node. Each internal node has a splitting rule to separate that node. Although there are several splitting methods for obtaining the splitting rules, the most popular one is the dichotomize method. The splitting rule of t can be induced by any question of the form ``$\boldsymbol{Z} \in t_L$?'', where t_L is called the child node of t. The counterpart t_R of t_L that is obtained by division of t is also called as the child node of t. That is, the splitting rule divides the internal node t into two child nodes, t_L and t_R, and t is called the parent node of t_L and t_R. The most widely

used splitting rule consists of a single covariate Z_k $(k = 1, \cdots, p)$. If Z_k is a quantitative variable, then the rule becomes ``$Z_k \leq s$?'', where $s$ is a threshold. If $Z_k$ is a categorical variable with the set of possible values $\mathcal{F}_k$, then the rule becomes ``$Z_k \in \mathcal{F}_{kt_L}$?'', where $F_{kt_L} \subset \mathcal{F}_k$.

The CART algorithm for constructing the tree-structured model comprises splitting, pruning, and selection. In the splitting step, covariates space are recursively divided based on the optimal splitting rules and the maximum-size tree T_0 is constructed. To determine the optimal splitting rule of a node t into t_L and t_R, we evaluate all the possible splitting rules for t. In order to build the model with measures for concordance probability, we assume the following to dichotomize the node t: t_L has higher risk than t_R, we evaluate the concordance probabilities from $x = 0$ to $\tau = \max\{x_i\,;\, \delta_i = 1, i \in t\}$, and the contribution of the pair (i,j), where $\eta_i = \eta_j$ to the estimate of C is 0.5. Under these assumptions, the splitting criterions based on the measures $\widehat{C}_H, \widehat{C}_U, \widehat{C}_B$ and $\widehat{C}_K$ are given by as follows:

(i) The criterion based on Harrell's C

$$\hat{C}_H = \frac{\sum_{i\in t_L}\sum_{j\in t_R}\delta_i I(x_i < x_j) + 0.5\{\sum_{i,j\in t_L}\delta_i I(x_i < x_j) + \sum_{i,j\in t_R}\delta_i I(x_i < x_j)\}}{\sum_{i,j\in t}\delta_i I(x_i < x_j)}$$

(ii) The criterion based on Uno's approach

$$\hat{C}_U = \frac{\sum_{i\in t_L}\sum_{j\in t_R}\{\hat{G}_t(x_i)\}^{-2}\delta_i I(x_i < x_j) + 0.5\left\{\sum_{i,j\in t_L}\{\hat{G}_t(x_i)\}^{-2}\delta_i I(x_i < x_j) + \sum_{i,j\in t_R}\{\hat{G}_t(x_i)\}^{-2}\delta_i I(x_i < x_j)\right\}}{\sum_{i,j\in t}\{\hat{G}_t(x_i)\}^{-2}\delta_i I(x_i < x_j)}$$

where $\hat{G}_t(.)$is the Kaplan-Meier estimator for the censoring distribution based on the samples included in t.

(iii) The criterion based on Begg's approach

$$\hat{C}_B = \frac{2}{n_t(n_t-1)}\left\{\sum_{i\in t_L}\sum_{i\in t_R} p_{ij} + \frac{n_{t_L}(n_{t_L}-1) + n_{t_R}(n_{t_R}-1)}{4}\right\},$$

where n_t is the number of samples included in the node t. p_{ij} is defined as follows: If $\delta_i = \delta_j = 1$, then

$$P_{ij} = \begin{cases} 0, & x_i > x_j \\ 1, & x_i < x_j \end{cases}.$$

If $\delta_i = 0, \delta_j = 1$, then

$$p_{ij} = \widehat{\Pr}(U_i < x_j | U_i > x_i) = \begin{cases} 0, & x_i > x_j \\ 1 - \dfrac{\hat{S}_{t_L}(x_j)}{\hat{S}_{t_L}(x_i)}, & x_i < x_j \end{cases}.$$

If $\delta_i = 1$, $\delta_j = 0$, then

$$p_{ij} = \widehat{\Pr}(U_j > x_i | U_j > x_j) = \begin{cases} \dfrac{\hat{S}_{t_R}(x_i)}{\hat{S}_{t_R}(x_j)}, & x_i > x_j \\ 1, & x_i < x_j \end{cases}.$$

If $\delta_i = \delta_j = 0$, then

$$P_{ij} = \widehat{\Pr}\left(U_i < U_j | U_i > x_j, U_j > x_j\right) = \begin{cases} \frac{\hat{S}_{t_R}(x_i)}{2\hat{S}_{t_R}(x_j)}, & x_i > x_j \\ 1 - \frac{\hat{S}_{t_L}(x_j)}{2\hat{S}_{t_L}(x_i)}, & x_i < x_j \end{cases}.$$

$\hat{S}_t(.)$ is the Kaplan-Meier estimator for the failure distribution based on the samples included in t.

(iv) The criterion based on Korn and Simon's approach

$$\hat{C}_K = \frac{2}{n_t(n_t-1)}\left\{\sum_{i\in t_L}\sum_{j\in t_R} p_{ij} + \frac{n_{t_L}\left(n_{t_L}-1\right) + n_{t_R}\left(n_{t_R}-1\right)}{4}\right\},$$

where

$$p_{ij} = \widehat{\Pr}\left(U_i < U_j\right) = \sum_{i\in t_R}\left[1-\hat{S}_{t_L}(x_i^{*-})\right]\left[\hat{S}_{t_R}(x_i^{*-}) - \hat{S}_{t_R}(x_i^{*})\right] + \left[1-\hat{S}_{t_L}(\tau^{-})\right]\hat{S}_{t_R}(\tau).$$

2.3 Pruning and Selection

The maximum-size tree T_0 is obtained by recursively splitting in the splitting step, after which an optimal-size tree is constructed from the T_0 in the pruning and selection steps. In the pruning step, the nested subtrees $T_m \prec T_{m-1} \prec \cdots \prec T_1 \prec T_0$ are obtained by recursively removing the node in the T_0. T_m is the tree which has the root node only. For this purpose, we propose the concordance-complexity measure:

$$C_\alpha(T_h) = \hat{C}(T_h) - \alpha\left|\tilde{T}_h\right|$$

where $\hat{C}(T_h)$ is the concordance measure for the sub-tree $T_h(h = 0,1,\cdots,m)$. For example, if we use the $\hat{C}_H$ for splitting step, then $\hat{C}(T_h)$ is given by

$$\hat{C}(T_h) = \frac{\sum_{(t_a,t_b)\in\tilde{T}_h} \max\left\{\sum_{i\in t_a}\sum_{i\in t_b}\delta_i I(x_i < x_j), \sum_{i\in t_a}\sum_{i\in t_b}\delta_i I(x_i > x_j)\right\} + 0.5\left\{\sum_{t\in\tilde{T}_h}\sum_{i,j\in t}\delta_i I(x_i < x_j)\right\}}{\sum_{i,j}\delta_i I(x_i < x_j)},$$

where $\tilde{T}_h$ represents the set of terminal nodes in T_h, and $\left|\tilde{T}_h\right|$is the number of terminal nodes in T_h. $\sum_{(t_a,t_b)\in\tilde{T}_h}$ represents the all pairs of terminal node in T_h. The optimal tree for an arbitrary α is defined as the subtree that maximizes $C_\alpha(T_h)$. If the value of α is 0, then the optimal subtree is T_0. On the other hand, if α approaches ∞, then a model that is not considered to be a tree structure is selected as the optimal subtree. Therefore, by gradually increasing α from 0, we can obtain a set of optimal subtrees.

In selection step, we select a sub-tree from T_0, T_1,... T_m based on the V-fold cross validation. First, we construct the $T_1^v, T_2^v, \cdots$ based on the $\mathcal{L} - \mathcal{L}^{(v)}$ for each $v\,(v = 1,\cdots,V)$. Then, select the best subtree $T^v(\alpha_k')$ from $T_1^v, T_2^v, \cdots$,which maximize the value of $\widehat{C}(T^v(\alpha_k'))$ for each $\alpha_k' = \sqrt{\alpha_k\alpha_{k+1}}$. $\widehat{C}(T^v(\alpha_k'))$is the concordance measure obtained from test samples $\mathcal{L}^{(v)}$for

the sub-tree T^v . For example, if we use the $\widehat{C}_H$ for splitting step, then $\hat{C}(T_h)$ is given by

$$\hat{C}_S = \frac{\sum_{i,j\in\mathcal{L}^{(v)}} \delta_i I(x_i < x_j) I(\eta_i < \eta_j) + 0.5 \sum_{i,j\in\mathcal{L}^{(v)}} \delta_i I(x_i < x_j) I(\eta_i = \eta_j)}{\sum_{i,j\in\mathcal{L}^{(v)}} \delta_i I(x_i < x_j)},$$

where $\hat{\eta}_i = \sum_{l=1}^{m} \{1 - \hat{S}_t(x_l^*)\}$, $x_1^*, \ldots, x_m^*$ are unique event times in $\mathcal{L} - \mathcal{L}^{(v)}$, and $\widehat{S}_t(.)$ is the KaplanMeier estimate in $t \in \tilde{T}^v(\alpha_k')$, which $i \in \mathcal{L}^{(v)}$ is included. Finally, we select the subtree that satisfy

$$\arg\max_{T_k} \left\{ \frac{1}{V} \sum_{v} \hat{C}\left(T^v(\alpha_k')\right) \right\}.$$

3. Results

We present simple simulation studies to examine the properties of the proposed approach in several situations. As the first simulation, we used the following model to generate the data to compare the performance in terms of splitting criterion. As the quantitative covariate, $z_i = (100 \times i)/300$ for $i = 1, \cdots, 300$. The true splitting rule is given by "$z_i \leq 50$?". Then, in each child node, the exponential models with parameter μ_{t_L} = 1.0 and μ_{t_R} = 0.5 are specified as the distribution of failure times, respectively. The censoring times are generated from uniform distribution as the censoring rate become about 50%. Simulations are repeated 1,000 times. The results of the simulation are shown in Table 1. Table 1 lists the average values and standard deviations of the selected thresholds as the splitting rules. Table 1 lists the average values and standard deviations of the selected thresholds as the splitting rules.

Table 1: The average and standard deviation of the split points.

Criterion	Ave. selected threshold (std.)
$\hat{C}_H$	49.6 (6.1)
$\hat{C}_U$	50.6 (5.8)
$\hat{C}_B$	50.3 (5.4)
$\hat{C}_K$	48.8 (7.1)
Log – rank test	47.0 (11.9)

To compare the performance of a tree obtained by using each criterion, we used the following model to generate data. The four categorical covariates$Z_1, \cdots, Z_4$ are generated from a discrete uniform distribution with {1,2,3,4}. The failure and censoring times are generated from exponential distributions with parameter μt and 0.37, respectively. The model of μ_t is given by

$$\mu_t = \begin{cases} 0.4, & Z_1 \in \{1,2\} \cap Z_2 \in \{1,2\} \\ 0.7, & Z_1 \in \{1,2\} \cap Z_2 \in \{3,4\} \cap Z_3 \in \{1,2\} \\ 1.0, & Z_1 \in \{1,2\} \cap Z_2 \in \{3,4\} \cap Z_3 \in \{3,4\}. \\ 1.3, & Z_1 \in \{3,4\} \cap Z_3 \in \{1,2\} \\ 1.6, & Z_1 \in \{3,4\} \cap Z_3 \in \{3,4\} \end{cases}$$

The Z_4 is nuisance. The sample size was set to 500. On average, approximately 70% of the subjects experienced the event. Table 2 lists the average values and standard deviations of the Harrell's C, integrated Brier scores, and tree sizes.

4. Discussion and Conclusion

In this study, we consider the concordance probability-based splitting criterions for constructing a survival tree. We proposed the new method to construct the tree model that maximizes prediction accuracy based on the CART. We study the performance of the splitting ability of the criterion based on concordance probabilities, and compare the survival trees constructed by proposed method and conventional methods through simulations. From the simulation results, our proposed approach has the advantage to construct the model with high prediction performance.

Table 2: The average and standard deviation of the Harrell's *C*, integrated Brier scores, and tree sizes.

Criterion	Harrell's C (std.)	Integrated Brier Score (std.)	Tree size (std.)
$\hat{C}_H$	0.612 (0.010)	0.542 (0.018)	5.8 (4.6)
$\hat{C}_U$	0.610 (0.013)	0.557 (0.031)	11.9 (10.1)
$\hat{C}_B$	0.602 (0.006)	0.542 (0.009)	4.2 (2.1)
$\hat{C}_K$	0.601 (0.013)	0.584 (0.020)	22.9 (2.3)
Log – rank test	0.604 (0.013)	0.539 (0.009)	2.2 (0.5)

References

1. Cox D R. Regression models and life-tables, Journal of the Royal Statistical Society: Series B, 1972; 34: 187-220.
2. Breiman L, Friedman J H, Olshen R A, and Stone C. Classification and Regression Trees. Wadsworth, California. 1984.
3. Leblanc M, and Crowley J. Survival Trees by Goodness of Split, Journal of the American Statistical Association, 1993; 88: 457-467.
4. Harrell F E, Lee K L, and Mark D B. Tutorial in Biostatistics Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors, Statistics in Medicine, 1996; 15: 361-387.
5. Uno H, Cai T, Pencina M J, D'agostino R B, and Wei L J. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data, Statistics in Medicine, 2011; 30: 1105-1117.
6. Begg C B, Cramer L D, Venkatraman E S, and Rosai J. Comparing tumour staging and grading systems: a case study and a review of the issues, using thymoma as a model, Statistics in Medicine, 2000; 19; 1997-2014.
7. Korn E L, and Simon R. Measures of Explained Variation for Survival Data, Statistics in Medicine, 1990; 9; 487-503.
8. Schmid M, Wright M N, and Ziegler A. On the use of Harrell's C for clinical risk prediction via random survival forests, Expert Systems With Applications, 2016; 63: 450-459.

Validation of Measurement Instrument for Survey Research: A Review of Rating Scales Related Factors

Nurul Hafizah Azizan, Zamalia Mahmud, Adzhar Rambli
Faculty of Computer and Mathematical Science, Universiti Teknologi MARA

Abstract

The use of a questionnaire as a measurement instrument has been recognized for many years before. It is a tool most widely used in a survey by many researchers regardless of field of studies. In fact, the questionnaire is considered a powerful tool to gain information on the phenomenon of interest from respondents. A well-validated questionnaire is an important requirement to guarantee the quality of the data obtained. There are several factors that may influence the quality of the constructed questionnaire. One of them is the choice of rating scales to be used. In this paper, several number of articles was reviewed to examine the effect of rating scales used in a survey on the quality of the measurement instrument. These include validity and reliability. Studies conducted by several researchers have found that the types of rating scale, specifically the number of response categories used in constructing the questionnaire will affect its validity and reliability. From the review, it was found that validity and reliability were relatively poor for four and less point scales. In addition, it also revealed that 5-point and 7-point scales produced better validity and reliability. Meanwhile, for the scale point beyond seven, there was no significant improvement in validity and reliability indexes. In addition, the scales with end points labelled produced much lower reliability indexes compared to the scales options clearly labelled on each category.

Keywords

Measurement instrument; questionnaire; rating scales; validity; reliability

1. Introduction

The quality of a measurement instrument used plays a key role in ensuring the quality of data gained in the survey. In fact, only with a well-developed instrument, the quality of the data obtained can be preserved. Questionnaire is a well-known measurement instrument used by most of the researchers. It is a powerful tool used for collecting data in survey research. When conducting a survey, researchers usually deal with either observed and unobserved types of variables. Observed variables such as gender, income, race and education level are directly measurable. In contrast, the unobserved variables; for instance, life satisfaction, happiness, stress, anxiety, loyalty etc. cannot be directly measured from respondents. This is because these types of variables

are subjective in nature, which is also referred to as latent constructs. Therefore, extra caution should be given when constructing a questionnaire with these types of variables. This is to ensure that the questionnaire clearly depicts the actual meaning of the variables to be investigated.

In practice, these types of constructs will be measured indirectly using several indicators or multiple reflective items with appropriate types of rating scale to represent these components. This enables the respondents to provide their responses based on the scales corresponding to the items. Based on the responses given by respondents, the composite scores of these multiple items will be computed as a single meaningful factor(s) where it can be used in further analyses. Since these latent constructs are subjectively measured, it is crucial for the researchers to ensure that the measurement instrument used for the survey meets the psychometrics requirements of both validity and reliability. It is acknowledged that the development of questionnaire as a measurement instrument for a survey that fulfils requirement of the psychometric properties is not an easy task. It requires several important steps to be followed which takes considerable time and effort from the researchers.

The utility of the measurement instrument will depend on its quality, as it can be useful if certain psychometric and practical requirements are met (Shultz, Whitney, & Zickar, 2013). The quality of measurement instrument can be affected by many factors and this requires careful attention at the beginning of the questionnaire development. As Radhakrishna (2007) remarked, a systematic development of a valid and reliable questionnaire is a must to reduce measurement errors. Designing a questionnaire that produces usable data is not as easy as it might seem. In fact, the development of inappropriate instrument will lead to poor quality of data, misleading conclusions and affect the recommendation to be given (Boynton, 2004), consequently leading to ineffective corrective action taken to deal with the problems in hand. A well-developed questionnaire will produce precise findings. This is because it could avoid or reduce potential errors and biases, which will lead to better quality data. Only with appropriate and well-developed questionnaire, a relevant information can be obtained and better decision can be made.

Therefore, before the questionnaire can be distributed to the respondents, it is advisable to follow three main guidelines to minimise bias in a survey research as highlighted by Sekaran and Bougie (2016). These include principles of wording, principles of measurement and general appearance of the questionnaire. In constructing a questionnaire, any researcher should take particular concern on the measurement to be used. This is because measurement is fundamental in any scientific research. Sekaran and Bougie (2016) mentioned that measurement of the variables has become an integral part of research, where the variables of interest should be measured correctly

to find the answers to the research questions. Thus, the researchers should ensure that they follow the correct procedures and processes required in developing a questionnaire as a research instrument. This includes emphasis on the suitable rating scales to be used. At times, this needs to be refined in order to enhance the quality of the data and utilization of research findings for the purpose of decision making. Usually the quality of the questionnaire to be used for the survey will be assessed through validity and reliability indexes. Validity and reliability are two different concepts. According to Sekaran and Bougie (2016), validity is a test of how well an instrument that is developed measures the particular concept it is intended to measure while reliability is an indication of consistency that the instrument measures the concept it is measuring. This paper highlights the effect of choice of rating scales used in survey research on validity and reliability the measurement instrument.

2. Methodology

In general, for this review, there are three main stages that have been followed which comprise of Searching Process, Screening Process and Systematic Review Process (Fig. 1). This review was conducted within a period from September 2018 to December 2018. A total of 50 articles were extracted to be reviewed. These include those considered from Google Scholar, ScienceDirect and Scopus. The eligibility of the articles to be included in this study was confirmed through screening process. In order to ensure the validity of the journal sources, all articles obtained from Google Scholar were examined through Scimago and WoS online system.

As shown in Figure 1, a total of 50 articles were extracted through databases mentioned above. Out of 50, 25 articles were obtained from Google Scholar, 10 from ScienceDirect and 15 from Scopus databases. All these articles were assessed for their eligibility to be included in this review. Seventeen duplicate articles were excluded, and 8 articles removed as they were not published within the year range of 1990 and 2018. Ten articles were also excluded as they were not indexed either in Scopus and ISI. Thus, out of 50, only 15 articles were eligible and chosen to be reviewed at the final stage.

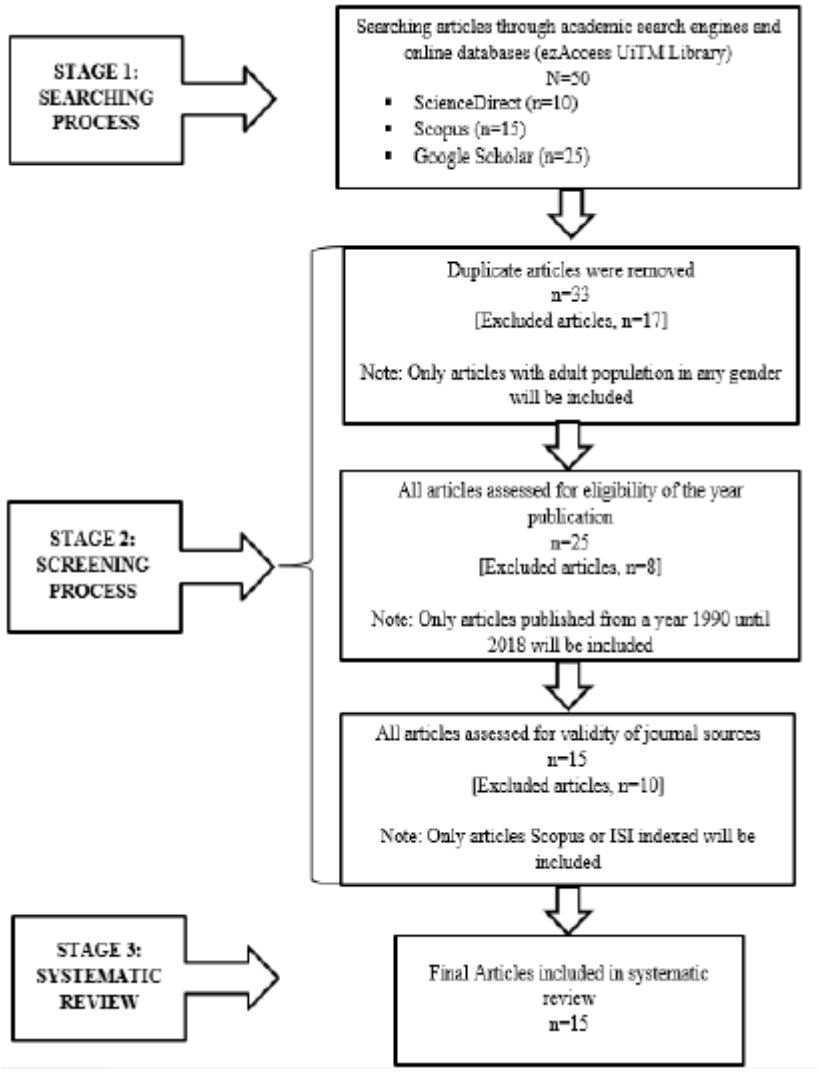


Figure 1: Flow Diagram of Searching Related Articles

3. Result and Discussion

The use of questionnaire as a measurement instrument has been recognized for many decades. It is a powerful tool and the most widely used instrument by many researchers in various field of studies to gain information on phenomenon of interest from respondents. Well-validated questionnaire is an important requirement to guarantee the quality of the obtained data. There are several factors that may influence the quality of the constructed questionnaire. One of them is the choice of rating scales to be used. Therefore, during the questionnaire development, apart from avoiding the principles of wording such as double-barrelled, ambiguous word, recall-dependent, leading and loaded questions (Sekaran & Bougie, 2016), it is also vital to wisely decide on the choice of rating scales to be used.

Survey research can take many forms of rating scale such as Likert scale,

dichotomous scale, category scale, numerical scale, itemized rating scale etc. (Sekaran & Bougie, 2016). Hence, researchers need to carefully choose which rating scale (i.e., types of rating scale, labels of rating scale and number scale point) is suitable that can fulfil the psychometric properties and maximize both validity and reliability of the survey instrument. Studies conducted by several researchers have found that the characteristics of rating scale used such as types of rating scale, labels of rating scale and number of response categories with either odd or even response alternative might influence the quality of measurement instrument (Allahyari, Jafari, & Bagheri, 2016; Cicchetti, Showalter, & Tyrer, 1985; Daher, Ahmad, Winn, & Selamat, 2015; Eutsler & Lang, 2015; Osteras et al., 2008; Preston & Colman, 2000; Revilla, Saris, & Krosnick, 2014).

Gaskell, Wright and O'Muircheartaigh (1995) also remarked that the way in which the respondents respond to the questions may substantially be influenced by the construction of the response scale. In another study by Preston and Colman (2000), they also claimed that reliability, validity and discriminating power can be affected by the format of rating scale used. The findings from this study suggested that with four-point and a smaller number of categories, the reliability, validity and discriminating power performed were relatively poor but significantly higher with more response categories of up to seven-point. This reveals that besides playing a significant role on the quality of response given by the respondents, the choice of suitable rating scales used in the survey can also give a huge impact on the quality of the measurement instrument in overall. However, Maydeu-Olivares et al. (2009) said that the use of optimal number of response alternatives for a survey is still a debate, and no consensus about this issue has been reached. Apart from that, the use of continuous rating scales has also received considerable attention by the researches. One of the mostly used continuous rating scales is Visual Analogue Scale (VAS). The use of a continuous rating scale in a survey has offered some advantages. Since a wide range of scales are being provided, hence the respondents will have the freedom to rate their response. Besides that, the actual experience of the respondents can also be assessed (Hasson & Arnetz, 2005). Thus, the quality of the information gained from the survey could be preserved.

3.1 Number of Response Categories and Its Effect on Validity and Reliability

Numerous studies in the last few decades confirmed that by varying the number of response categories may influence the validity and reliability of a survey instrument. Osteras et al. (2008) conducted a study with randomized design to compare the quality of instrument in terms of internal consistency and discriminant validity with original 4-point scale and new 5-point scale

version of Norwegian Function Assessment Scale (NFAS). NFAS is an instrument that comprised of 39-items used to evaluate the need for rehabilitation, the right to social security benefits and adjustment of work demands among sick-listed persons among two different groups of respondents with no major differences in demographic characteristics. The result from this study suggested that 5-point scale produced a better data quality in both internal consistency and discriminant validity as compared to 4-point original scale. Although odd number of response with 5-point and 7-point rating scales are the most frequently used, previous study suggested that the 7-point scale will maximize the variance, and for the scale point beyond seven, it will not increase the variance (Eutsler & Lang, 2015). However, another study from Revilla, Saris, and Krosnick (2014) reported that 5-point scale produced better data quality rather than 7-point and 11-point scale.

A study conducted by Preston and Colman (2000) on response categories ranging from 2 to 11 found that reliability, validity and discriminating power were relatively poor for four and less point scales, and significantly higher for five to seven point scales. However, the study also revealed that test-retest reliability tends to drop with more than 10 response categories despite internal consistency that does not significantly differ. The result obtained is almost similar with a study carried out by Lozano, García-Cueto, and Muniz (2008). Based on simulated data using Monte Carlo method of 30-items with response alternative ranging between two to nine and four different sample sizes (50, 100, 200 and 500), the results showed that both validity and reliability are better with response alternative between four and seven, and decrease with less than 4-point scale. By using three different samples of 50 participants each, Daher et al. (2015) distributed the Malay Spiritual Well-Being Scale (SWBS) with the original 6-point scale, 3-point and 4-point modified scales to study the impact of rating scales categories on reliability and fit statistics using Rasch model. The results showed that reliability and fit statistics were robust with the original 6-point scale and became worse for both new modified scales (3-point and 4-point). Similar with Osteras et al. (2008), the findings obtained in this study might also be affected by different groups of respondents involved in the study, where the same sample is more preferable to be used to get more accurate result for comparison purposes. Through a comparison between 7-point and 11-point categories of rating scales used for quality of life survey, Alwin (1997) reached the conclusion that questions with more categories are both more reliable and valid.

3.2 Labels and Rating Scale Format and Its Effect on Validity and Reliability

Other than studying the influence of number of response categories (ranged from 3 to 9) on quality of measurement instrument, Weng (2004) also put an

effort by examining the effect of scale format (either anchor labels or end points labels) on the reliability of Likert-type rating scales. As mentioned in previous studies, the findings of this study also revealed that few response categories tended to result in lower test-retest reliability. Besides, scale format significantly affects the performance of measurement instrument, with the scales with end points labelled were likely to produce lower test-retest reliability rather than the scales with all the response options clearly labelled. The use of graphic or visual rating scale also had important implication on survey research especially with children as respondents. Cremeens (2007) yielded important findings as the results suggested that among three types of rating scale (categorical scale, faces and thermometer), categorical scale is more preferable to be used for the items measuring ability factors while faces scale is found to be more effective to be used for items assessing social construct in order to improve measurement instrument reliability.

4. Conclusion

From the review, it is found that the rating scales used has a significant effect on the quality of measurement instrument used for the survey. These include number of response alternatives, format and labels of rating scales used. Apart from that, it is also important to study the effect using a central category such as "neutral" response and to examine the interaction between the number of response categories and items used for future research. In addition, it is suggested to examine whether the type of psychometric model used gives a significant implication on the findings obtained.

References

1. Allahyari, E., Jafari, P., & Bagheri, Z. (2016). A Simulation Study to Assess the Effect of the Number of Response Categories on the Power of Ordinal Logistic Regression for Differential Item Functioning Analysis in Rating Scales. *Computational and Mathematical Methods in Medicine*, *2016*.
2. Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research*, *25*(3), 318–340.
3. Boynton, P. M. (2004). Administering, analysing, and reporting your questionnaire. *British Medical Journal*, *328*(7452), 1372–1375.
4. Cicchetti, D., Showalter, D., & Tyrer, P. (1985). The effect of number of rating scle categories on levels of interrater reliability: A monte carlo investigation. *Applied Psychol Measure*, *9*(1), 31– 36.
5. Cremeens, J., Eiser, C., & Blades, M. (2007). Brief report: Assessing the impact of rating scale type, types of items, and age on the measurement of school-age children's self-reported quality of life. *Journal of Pediatric Psychology*, *32*(2), 132–138.

6. Daher, A. M., Ahmad, S. H., Winn, T., & Selamat, M. I. (2015). Impact of rating scale categories on reliability and fit statistics of the Malay Spiritual well-being Scale using Rasch analysis. *Malaysian Journal of Medical Sciences*, *22*(3), 48–55.
7. Eutsler, J., & Lang, B. (2015). Rating Scales in Accounting Research: The Impact of Scale Points and Labels. *Behavioral Research in Accounting*, *27*(2), 35–51.
8. Gaskell, G., Wright, D. B., & O'Muircheartaigh, C. (1995). Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics*.
9. Hasson, D., & Arnetz, B. B. (2005). Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. *International Electronic Journal of Health Education*, *8*, 178–192.
10. Lozano, L. M., García-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(2), 73.
11. Maydeu-Olivares, A., Kramp, U., Garcia-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods and Instrumentation*, *41*(2), 295–308.
12. Osteras, N., Gulbrandsen, P., Garratt, A., Benth, J. S., Dahl, F. A., Natvig, B., & Brage, S. (2008). A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale. *Health and Quality of Life Outcomes*, *6*, 1–9.
13. Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales:
14. reliability, validity, discriminating power, and respondent preferences. *Journal of Acta Psychologica*, *104*(1), 1–15.
15. Radhakrishna, R. B. (2007). Tips for developing and testing questionnaires/instruments. *Journal of Extension*, *45*(1), 1–4.
16. Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the Number of Categories in AgreeDisagree Scales. *Sociological Methods and Research*, *43*(1), 73–97.
17. Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach* (7th ed.). John Wiley & Sons.
18. Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2013). *Measurement theory in action: Case studies and exercises*. Routledge.
19. Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956–972.

Variation in copy number on the genome of the Brazilian population

Ana C. M. Ciconelle[1], Júlia M. P. Soler[1], Alexandre C. Pereira[2]
[1]Institute of Mathematics and Statistics - University of Sao Paulo – USP, Brazil
[2] Heart Institute - University of Sao Paulo – USP, Brazil

Abstract

Copy number variation (CNV) is an alteration in the number of copies of a DNA segment, unbalancing the diploid state in humans at any given locus on the genome. The CNV region can include from a single nucleotide polymorphism (SNP) to several genes, and such variation can be classified in five states: 0 (deletion of two copies), 1 (deletion of one copy), 2 (normal state), 3 (single copy duplication) and 4 (double copies duplication). Several diseases (such as uric acid, pancreatitis and nervous system disorders) and phenotypes (such as height and cholesterol levels) have been associated to this kind of structural variation, suggesting that inheritance patterns can be involved besides revealing variability across populations. In this study we propose a pipeline for CNVs calling from SNP array data. Further, in collaboration with Heart Institute (USP), this work uses dataset from Baependi Heart Study to characterized the CNVs in the Brazilian population and associate them with height. Genomic and phenotype data consisted of 1,120 related individuals sampled according to family-based design. The results pointed out to CNV regions specific for Brazilian population, but also for similarities with others populations according the length and number of CNVs in samples. In addition, based on trios data (parents and offspring) it was observed that the CNV transmission could not follow the Mendelian laws. Our work also identified a region in the chromosome 9 associated to height, where it carries a duplication with an expected height dropped by approximately 3cm.

Keywords

CNV calling; association studies; height, missing heritability; mixed model

1. Introduction

As described by Lewis (2012), Genome Wide Association Studies (GWAS) aim to associate genetic markers, candidate genes or genome regions with complex traits and diseases, which are likely derived from multiple genes and environment, such as height and diabetes. In addition, discovering the associations between diseases and genetic factors is an important step to understand the pathogenesis of the diseases and to facilitate the process of diagnosis and treatment. The most used genetic variant for GWAS is the single nucleotide polymorphism (SNP), but other variants, as small

insertions/deletions (indels) and copy-number variations (CNVs), are also available.

Several studies are being performed to catalogue the human genetic variants to facilitate GWAS, such as the HapMap Project and 1000 Genomes Project. In both projects, samples are majorly from African, Asian and European populations and they aim to identify genetic variants with frequencies of at least 1% in the studied populations, including not only SNPs, but also structural variants and small insertions/deletions (The International HapMap Project, 2003; 1000 Genomes Project Consortium, 2016). Even though there is a major success in gene discovery, the percentage of variance explained by GWAS loci for many traits is relatively low. Thus, a substantial part of the traits variation is still unexplained. This phenomenon is called missing heritability. One example of trait with a high missing heritability is the height. In Manolio et al. (2010), two of the solutions cited to revealing the missing heritability is to use different types of genetic variants including common and rare variants.

Based on these scenarios, in this work, our focus is on CNV detection since this kind of variant is not as well characterized as SNPs, but it is expected to have an important role on the association with several traits and diseases. Copy number variation occurs when the number of copies of a particular region (one or more loci) of the DNA differs from two in autosomes or one/two in allosomes and can to explain phenotypic variability in humans. The effects of CNVs to human diseases are not yet well known although several diseases have been associated to this kind of polimorphism, such as uric acid (Scharpf et al., 2014).

GWAS are usually based on reference maps which do not take into account the population-specific and rare variants. In addition, Sanna et al. (2011) shows that adding rare variants in association studies doubled the explained heritability of traits. Therefore, identifying different types of variants and including data from specific populations can explain the missing heritability of traits and diseases. This motivates to build genomic reference maps for specific populations, as proposed by the project Genome of Netherlands (Boomsma et al., 2014), which aims to characterize genetic variants from Dutch population, including rare variants.

Considering the unknown influence of CNVs on anthropometric measurements and the lack of studies based on Brazilian population, this work was developed in collaboration with the Laboratory of Genetics and Molecular Cardiology (Heart Institute-USP, Brazil). Using the database from the Bapendi Heart Study described by Egan et al. (2016), we analysed the genotype (SNP data) and phenotype data from 80 families to characterize the CNVs in the Brazilian population and to understand their association with phenotypes, such as height. The main purpose of this project is to present methodologies

to quantify and call CNVs from SNP platforms and to analyse such data considering family based designs and characterize the patterns of the CNVs detected in this population.

2. Methodology

Dataset

Due to multiple waves of immigration, Brazil has a highly admixed population, which can be driven by genetic and environmental influences on several traits. The Baependi Heart Study is being conducted by the Heart Institute since 2005 to develop a longitudinal family-based cohort study for understanding the variation of cardiovascular risk factors within the Brazilian population and disentangle its genetic and environmental components. The data provides information about 105 families (1,666 individuals, 723 male and 943 females) living in the village of Baependi, in the state of Minas Gerais, Brazil. Data from 631 nuclear families were available, with offspring ranging from 1 to 14. The number of generations per family varied from 2 to 4 (54% of the families had 3 generations, and 45% had 2 generations). Only individuals aged 18 years or older were considered eligible for participating in the study. The mean age was 44 years, with a range of 18 to 100 years.

For each participant a questionnaire was used to obtain information regarding family relationships, demographic characteristics, medical history and environmental risk factors. Anthropometric measures, physical and clinical examination and electrocardiogram of the participants were performed by trained medical students. Genomic DNA was extracted by standard procedures. From DNA samples, genotyping with SNP array was made based on Affymetrix Platform 6.0 and 1,120 CEL files were obtained, which stores the intensity values of each probe array for a single sample and several others information. More details are described in Egan et al. (2016).

Overview

The methodology used in this work is summarized by Figure 1, which describes the pre-processing of SNP data, the CNV calling and the CNV analysis. For the pre-processing of SNP data and the CNV calling, the software Affymetrix Power Tools (APT) (Affymetrix, 2017), PennCNV by Wang et al. (2007) and packages from the R environment were used. Using APT, given the CEL files, signal intensity values for probes are normalized through quantile normalization. Then, the median polish is applied to get the final cleaned intensity values for alleles A and B for each SNP. Also, the individual genotype calls is made using the Birdseed algorithm. For each SNP in each sample, the genotype will be coded as 0, 1 and 2 for AA, AB, BB and -1 for missing values, respectively, with its corresponding confidence scores. In addition, a final report will infer the sample sex. PennCNV generates canonical genotype clustering files based on the output files from APT. These files contain cluster

positions of each SNP for each canonical genotype (AA, AB and BB). Then, the calculation of LRR and BAF values for each SNP and each sample are made. All these values are used by a hidden Markov model (HMM) to CNV calling for each sample. Quality control values are also generated. The identified CNV regions are specific for each sample (individual). We excluded the samples that do not pass in the quality control. Then, a new set of minimal regions, defined by the overlap regions across all samples, was built and all minimal regions with a low frequency of CNVs (less than 2%) were removed. The final minimal regions are then ready for the CNV analysis of this work.

The SOLAR package combined with R scripts was used to analyse the CNVs, calculate the heritability of traits and associate CNVs and traits. Heritability corresponds to the intraclass correlation coefficient defined under linear mixed model formulation and considering family-based designs.

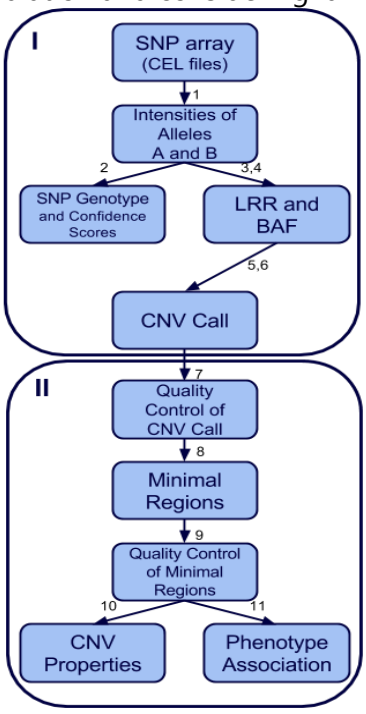


Figure 1: Flowchart of the pipeline. Box I indicates CNV calling and box II indicates CNV analysis. The number indicates which function was used: 1 (apt-probeset-summarize) and 2 (apt-probeset-genotype) are from APT, 3 (generate_affy_geno_cluster.pl), 4 (normalize_affy_geno_cluster.pl), 5 (kcolumn.pl), 6 (detect_cnv.pl) e 7 (filter_cnv.pl) are from PennCNV, 8 (CNTools) is from bioconductor package cntools, 9, 10 and 11 are basic functions from R environment and SOLAR software.

3. Results

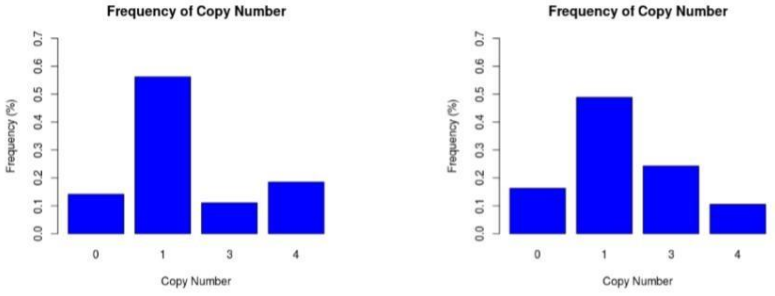
By the end of the CNV calling, each sample has a file describing the identified CNVs as showed in Table 1. From the 1,120 samples 910 were

considered for analysis due the quality control filtering. From the original data, we were able to identify 375,312 CNVs and, after the cleaning procedure, this value dropped to 135,414 CNVs. From these CNVs, we obtained 64,107 minimal regions, in which we considered the overlap of CNVs. Due the low frequency of some CNVs in the samples, after filtering, only 8,794 were considered.

How many CNVs does an individual have? For this question, the number of CNVs we obtained from each sample varies from 17 to 2,921 CNVs. However, we also can observe that a subgroup of 83% of the samples contains less than 100 CNVs, which is expected limit for PennCNV. For this subgroup of samples, the mean number of CNVs per sample is 56.49 (standard deviation equal to 15). For both, the complete samples and the subgroup, the median of 60 and 57 CNVs, respectively, are compatible with similar studies. We also can observe that deletions are more frequent than duplication as show in Figure 2.

Sample	Chr	Start	End	Number	Length	State	CN	First Marker	Last Marker
1	15	22231485	22264715	31	33231	2	1	CN_691574	CN_691602
1	19	59989695	60040503	29	50809	2	1	CN_170378	SNP_A-4271224
1	17	18296117	18373803	21	77687	2	1	CN_749706	CN_751779
1	17	67057139	67076931	22	19793	2	1	CN_744214	CN_744222
1	9	44181813	44569219	23	387407	2	1	CN_1322576	CN_1322482
1	4	64380064	64390853	30	10790	1	0	CN_1052052	CN_1052079

Table 1. Illustration of the file containing the CNVs from sample 1. PennCNV generates a file with this structure for each sample. Each line describes a CNV. Columns "Chr", "Start" and "End" indicates the region of the CNV. "Number" is the number of markers from the Affymetrix 6.0 platform inside the region of the CNV. "Lenght" is the size of CNV in base pairs (bp). "State" corresponds to HMM states and "CN" is the number of copies associated to the state. "First and Last Markers" identify the markers where the CNV starts and ends.



(a) *All samples* (b) *Samples with less than 100 CNVs*

Figure 2. Distribution of CNVs regarding the number of copies. 0 and 1 indicate deletion, while 3 and 4 indicate duplication. (a) contains CNVs from all the samples and (b) contains only CNVs from samples with less than 100 CNVs.

How long are the CNVs identified in the Brazilian population data? According our results, the length of a CNV varies between 3bp to 27,435,314bp (27.5Mb) and follows a log-normal distribution as obtained by Scharpf et al.

(2014). Figure 3 shows histograms of the size of CNVs, indicating that deletions are, in general, shorter than duplications.

Where are the CNVs? The literature shows that chromosomes 19, 22 and Y present the biggest proportions of CNVs. From our dataset, chromosomes 19 and 8 have more regions of CNVs based on the number of base pairs. However, when only CNVs detected in at least 5% of the samples are considered, chromosomes 19 and 9 have the biggest proportions.

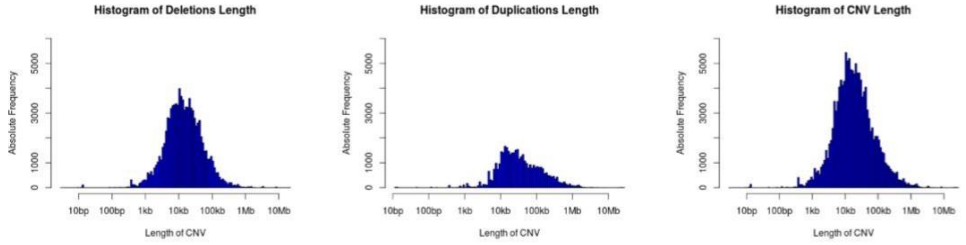


Figure 3. Histograms of CNV length. Figure (a) considers only deletions; Figure (b) considers only duplications and (c) contains all CNVs. The data is presented in exponential scale.

For understanding the distribution of CNVs along the genome, Figure 4 shows the absolute frequency of the detected CNVs along the positions on the chromosome 1 and 6. The region with highest presence of CNVs across samples is in chromosome 1, between positions 72,541,505bp and 72,583,736bp (Figure 4), which, on average, 818 samples from the total of 910 has a deletion or duplication.

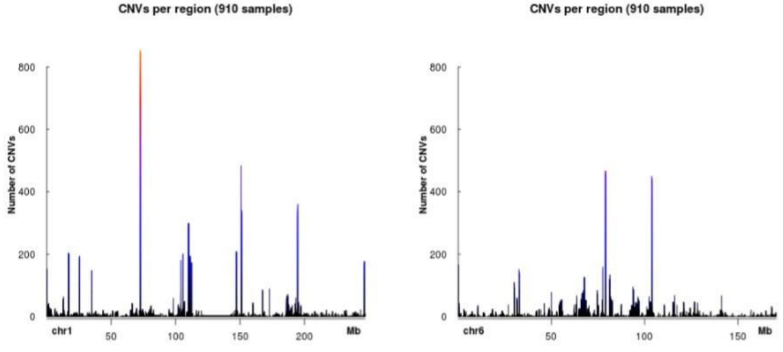


Figure 4. Number of CNVs per region after finding the minimal regions. The x-axis represents the positions of the chromosome by base pairs. The y-axis indicates the number of CNVs detected in the respective position for 910 samples.

Does the CNV follow Mendelian laws? This question is related to the inheritance pattern of the CNVs. At total, 106 trios were analysed from the Baependi families. They include trios with the same parents with different offsprings. As expected, normal parents and normal offspring is the most

common combination of CNV occurrences, in which, on average, 77.45% of the trios are all normal for all 8,794 CNVs. When we consider the case of one parent being normal and another having a deletion, we expected, under the Mendelian law, that the proportions of children with two copies would be similar to the children with one deletion. However, on average, 7.52% of the trios has parents with one normal parent and another with single deletion and the mean frequency of the trios with offspring with deletion is 1.29%, while with normal offspring is 6.06%. It means that, in general, the affected parent transmits preferentially the normal allele instead of the allele with deletion. A similar situation can be found for trios in which one of the parents is normal and the other has one duplication.

Are the CNVs associated to height? Height is a complex phenotype and its heritability is estimated to be around 80%. Due to the missing heritability, we explored this phenotype in association with CNVs. A linear mixed model was applied considering height as response variable and age, sex, ancestry coefficients and CNV as covariates. We adjusted the model in three different ways, in which the CNV was considered as dichotomous, having linear effect and as categorical covariate in five levels. Based on the results, the region from 78,960,219bp to 78,967,224bp in chromosome 9 is the most significant among the CNVs. Figure 5 shows the Manhattan plot for the model with CNV as a discrete variable. Also, the results indicate that the presence of duplication decreases the expected height by approximately 3cm as show in Figure 6.

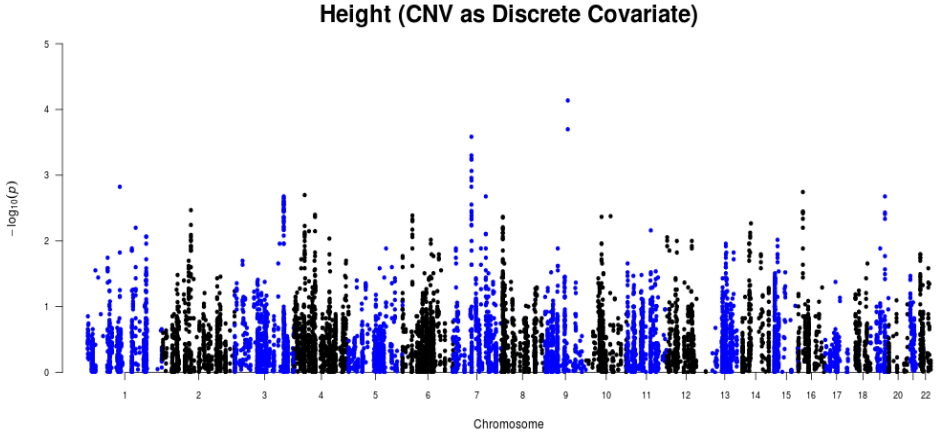
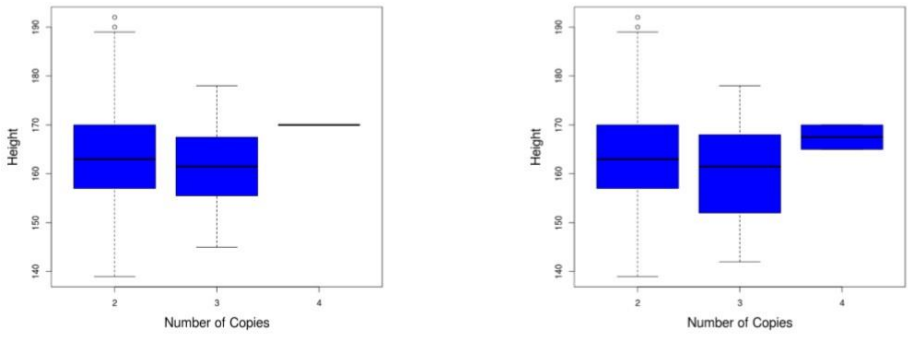


Figure 5. Manhattan plot from the second model. y-axis indicates the –log10 (p-value) of each CNV in association with height. The position used in the x-axis is the center position (in bp) of the CNV.



(a) (b)

Figure 6. Distribution of the height. These distributions are based on the number of copies for the regions from 78,960,219bp to 78,961,487bp **(a)** and from 78,961,488bp to 78,967,224bp **(b)** of chromosome 9.

4. Discussion and Conclusion

Our approach allows us to characterize CNVs occurring on Brazilian population. The CNV database built in this work can be used for association studies with different phenotypes. From this descriptive analysis of the obtained CNVs, we could observe that the distribution of the length and the number of CNVs per sample are similar to other populations as described in the literature, but specific CNV regions were also identified. The minimal regions identified can be considered as genetic markers specific for the Brazilian population. Further work can be performed based on the CNV database obtained and the annotation of the common identified CNVs should be made for better understanding the biological system.

References

1. 1000 Genomes Project Consortium (2016). HHS Public Access. *Nature* 526(**7571**): 68–74.
2. Affymetrix (2017). Affymetrix Power Tools: MANUAL: apt-probesetsummarize (1.20.0).
3. Boomsma et al. (2014). The Genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics* 22(**2**):221–227.
4. Egan et al. (2016). Cohort profile: the Baependi Heart Study—a family-based, highly admixed cohort study in a rural Brazilian town. *BMJ Open* 6(**10**):e011598.
5. Lewis C. M., Knight J. (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols* 2012(**3**):297–306.
6. Manolio T. A. et al. (2010). Finding the missing heritability of complex diseases. *Nature* 461(**7265**):747–753.
7. The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**:789– 796.
8. Sanna et al. (2011). Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genetics* 7(**7**).
9. Scharpf R. B. et al. (2014). Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations. *BMC Genetics* 15(**1**):1–13.
10. Wang et al.(2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* 17(**11**):1665–1674.

Export Competitiveness of Sumatera Coffee

Dewati Werdaningyas, Irwanto Leorisa
Central Bank of Indonesia, Jakarta, Indonesia

Abstract

The Indonesian coffee competitiveness, especially Sumatera, is considered low compared to other competing countries such as Brazil and Vietnam. This paper has two aims, the first is to identify determinants of competitiveness of coffee exports, second is to provide recommendations to improve coffee export performance. This research employs framework and toolkit in Trade Competitive Diagnostic by using Trade Outcome Analyisis and Competitivenes Analysis with Analtycal Hierarchy Process, enhanced by focus group discussion with coffee experts in Province of Lampung, Province of North Sumatera, Province of West Sumatera and Province of Aceh. Based on Trade Outcome Analyisis, the addition of new export destination countries from Indonesia and Vietnam tends to be bigger in number than Brazil. In terms of quality, Indonesia's coffee export products are similar to Brazil and Vietnam and are dominated by Arabica. Based on Analytical Hierarchy Process, the main determinant of competitiveness in developing Sumateran coffee are: firstly is access to the market, such as tariff barriers from export destination countries and non-tariff barriers such as certification standards; secondly is production, to overcome the condition of having low productivity along with the limitation of infrastructure and lack of human resource capabilities; thirdly is the infrastructure for trade promotion and lastly is the macro incentive. We recommend to establish a comprehensive institution, from upstream to downstream of the coffee sector chain, and strengthening the role of state owned enterprises as has been successfully applied in Vietnam. Also, in order to increase productivity of coffee plantation through replanting and rejuvenation is the most effective way. It is necessary to support road infrastructure improvements in coffee plantations as well as the application of post-harvest technology through dome infrastructure such as drying houses.

Keywords

Trade Competitive Diagnostic; Analytical Hierarchy Process; Sumatera; Coffee

1. Introduction

Indonesia's trade balance experienced a declining trend in recent years. The declining performance was mainly caused by commodity exports, both in volume and price. In this regard, an appropriate method is needed to

determine the strategy to increase export competitiveness. One way to do that is by diagnosing the obstacles of increasing competitiveness spatially as it is important to identify the problems of major commodity exports. It will help the process in policy making to increase export competitiveness.

2. The Research Objectives

The discussion to increase export competitiveness is still limited to certain commodities only. According to Farole and Winkler, 2012, export competitiveness in Indonesia's manufactures only addresses manufacturing industries located in Java (apparel, wood furniture, and automotive components). Some other commodities that require further discussion include coffee which is one of Indonesia's main commodities. Accordingly, the research objectives are formulated as follows: 1) to identify competitiveness and critical points of coffee export, and 2) to make recommendations to improve coffee export performance.

3. Methodology

The research used secondary and primary data and made a reference to framework and toolkit in Trade Competitiveness Diagnostic (TCD) (World Bank, 2011). The TCD framework consists of Trade Outcome Analysis (TOA) uses secondary data analysis, Competitiveness Diagnostic uses primary data analysis. TCD provides a framework for analyzing determinants of trade competitiveness across three broad are as:1).Market access focuses on the external trade policy environment that may facilitate or constrain exporters from entering and maintaining competitiveness in markets. 2.Supplyside factors covers a broad range of determinants including governance and macrofiscal, trade, and domestic policies that establish the incentive framework faced by the private sector; as well as the factor inputs that determine competitiveness at the factory or farm gate.3.Trade promotion infrastructure covers the range of interventions by government to address market failures. The tools used in TCD are Analytical Hierarcy Process (AHP) and focus disscusion group of coffee experts, corporations, farmers, association, owner of coffee cafe, traders, exporters and official from regional government. The respondent of AHP questionaire consist of 57 coffee experts in Province of Lampung, Province of North Sumatera, Province of West Sumatera and Province of Aceh.

4. Results

a. The Analysis of Coffee Trade Outcome

i. Export Growth

The development of Indonesian coffee exports has increased especially after 2005 which is in line with the price increase. In terms of volume, coffee

exports increased two times within 10 years (2005 which only reached 690 thousand tons has increased to 1.2 million tons in 2015). In 2017, the share of coffee exports reached 1.09% out of Indonesia's exports. However, the development of Indonesian coffee exports is relatively limited compared to Vietnam and Brazil. The development of coffee exports in Indonesia is supported by the increase in coffee extract and essence exports since 2011, while Vietnam and Brazil are still focusing on exports for non-roasted coffee.

ii. Market Share Analysis

The market share for roasted coffee commodity reached 6%, while market shares for coffee extract and essence as well as non-roasted coffee are 5% and 0.08%, respectively. The development of Indonesian coffee extract and essence has increased since 2010 with the Philippines as the main export destination. The three largest coffee producing countries (Brazil, Vietnam, and Indonesia) dominate the market for the non-roasted coffee market, while the market distribution for roasted coffee is still low. Non-roasted coffee production is dominated by countries in Europe such as Italy, Germany, Switzerland. This happens due to some factors: 1) certain preferences for the mentioned countries, and 2) shorter durability of roasted coffee compared to non-roasted ones. Meanwhile, coffee extract and essence production are from more varied countries such as Germany, Malaysia, the Netherlands, and Indonesia. Some countries producing coffee are Brazil, Colombia, Vietnam, and Indonesia. Indonesia exports coffee to some countries as the United States, Germany, and Japan. Meanwhile, exports from Brazil are aimed at more varied destinations, especially for countries in Europe. In 2015, Indonesia's coffee exports concentrated more in the United States (non-roasted coffee) and the Philippines (coffee extract and essence).

iii. Survival Analysis/Firm Participation

Based on survival analysis, the addition of new export destination countries from Indonesia and Vietnam tends to be bigger in number than Brazil. In 2016, the destination countries for coffee exports from Indonesia and Vietnam increased 10 and 11 countries respectively, higher than Brazil with only 7 countries. Some of Brazil's export destination countries that have not been accessed by Indonesia and Vietnam are countries in Africa and South America.

iv. Quality Sophistication

In terms of quality, Indonesia's coffee export products are similar to Brazil and Vietnam and are dominated by Arabica. The selling price of Indonesian coffee lies between Vietnam and Brazil. Coffee commodities sold by Brazil are dominated by Arabica which has a higher selling price, while Vietnam's coffee is dominated by Robusta which has a lower selling price.

b. Competitiveness Diagnostic

Based on AHP results, priorities that need to be improved to increase coffee exports competitiveness are Market Access and Production. Meanwhile,

Infrastructure for Trade Promotion and Macro Incentive are respectively in the third and fourth most important factors. Of all the alternatives seen through AHP, the priorities which are beneficial in increasing the competitiveness of coffee exports in Sumatera are tariff barriers, access to finance, and technical barriers to trade.

i.Market Access

Market access is defined as things that affect the ability of exporters to access international markets, including trade policies of importing countries that facilitate or hinder exporters from entering or penetrating markets and maintaining their competitiveness in the market. Increased market access can provide benefits such as larger economies of scale, reduced production costs, and greater specialization.

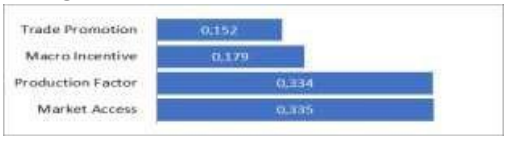


Chart IV.1. Perception of Priority for Increasing Coffee

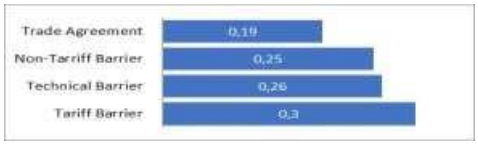


Chart IV.2. AHP *Market Access*

ii. Incentive Framework

In Aceh and South Sumatera, the tax that burdens exporters the most is 10% Value Added Tax (VAT). However, those 10% VAT has already been imposed on green beans. Therefore, the one who bear the VAT are the farmers since both of the exporters and buyers (importers) do not want to bear the tax. In addition to the 10% VAT, other taxes that are considered heavy are the increase in income tax which is initially amounted to 2.5%, currently rising to 7.5%.

Unlike North Sumatera, the main priority in terms of macro incentives is the depreciation of the rupiah against the dollar which affects coffee export activities. If the value of the rupiah depreciates, it actually provides incentives for farmers, exporters to increase production. In Lampung, business regulation, administrative procedures of local governments created problems in carrying out export activities.

iii. Factor Conditions

Coffee land productivity in Sumatra reaches 759 kg per hectare, which is higher than the national productivity average of 721 kg per hectare. However, the number is still far below the productivity of other competing exporting countries such as Brazil, Vietnam, and Colombia that have reached above 900 kg per hectare. The low productivity is caused by several reasons such as damaged and old plants, domination by smallholder plantations, human resources and infrastructure barriers.

Accordingly, one of the main obstacles in the development of coffee in Indonesia is the condition factor, which is related to financial access, especially to increase land productivity. The current situation of coffee business is that

entrepreneurs in Aceh are having difficulties to obtain trust from local banks because loans installments are often paid irregularly by the debtor. This is caused by uncertain harvest period and weather which affect income in the coffee business. In addition, the unavailability farmers' assets to borrow money become an obstacle when these entrepreneurs want to loan some money from a bank. Another obstacle is the kinds of seeds. The availability and the use of superior seeds that is suitable with the contours, climate, and nutrients in the plantation area is still not widely recognized and the process of fertilization and weeding is still rarely done.

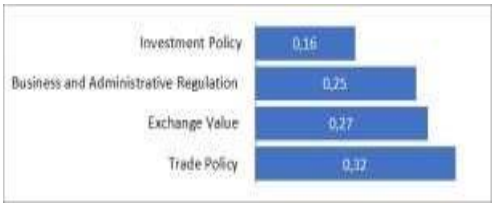


Chart IV.3. Perception of Macro Incentive Priorities

Chart IV.4. Perception of Condition Priority of Production Factors

The next problem is regarding work regulation and skills which are related to farmer cultivation. The current condition of farmers who understand the method and process of cultivating good coffee is limited as it affects the increase in coffee productivity. In South and North Sumatra, it is farmers' bad habit to harvest coffee that is still green or not fully immature. This habit is called rainbow picking which means harvesting coffee that is still immature to get money from exporters immediately. The selling value of unripe coffee is lower than the ripe ones and its quality decreases. Planting and harvest method are still in traditional way and the insufficient technology that is caused by the high cost of production equipment such as the grain threshing machine would affect the coffee productivity.

Some of the problems that often occur in coffee plantations include land management that is still traditionally performed by farmers both in the process of watering, fertilizing, picking and weeding on a small economic scale. This results in a relatively small volume of good quality products by farmers. Apart from the technical side of coffee cultivation, there are obstacles in terms of products and the development of processed coffee products. This problem includes the mismatch between supply and demand where the composition of the Sumatra coffee plant is still dominated by Robusta coffee (reaching 73% of total production), while the majority of international market demand is Arabica coffee. Indonesia faced difficulties to produce high value-added coffee products, with processing coffee into the form of extracts, essences and concentrates is still limited to around 4.1% of the total Sumatra coffee exports. **Furthermore, there are obstacles in terms of Trade and Logistics Facilities related to infrastructure.** The main

obstacle in this case is that the production centers are situated in the mountains resulting in high transportation costs.

iv. Trade Promotion Infrastructure

The coffee quality does not fully meet international standards. This occurs in Aceh, North Sumatra and South Sumatra. It is inseparable from the cultivation pattern and the application of technology which is still limited and the limited cultivation capabilities during the process of managing the coffee plantation land. In Lampung, the absence of foreign or domestic investment in coffee production hampered the growth of the technology used in the coffee industry. Currently, coffee production is still carried out in the traditional way which might hinder the mass production while the demand for coffee exports for both domestic and abroad are increasing.

c. Policy Priority to Increase Export Competitiveness

i. Market Access

The export share of Sumatran roasted coffee is still low at only 0.01% of Sumatra's coffee exports. This happens as a consequence of several things such as 1) Roasted coffee's durability is only about 1 month, much shorter than green beans which can reach 1 year, 2) The imposition of higher import tariffs, 3) Market preferences that cannot be met, and 4) Uncompetitive prices and a too long export chain. **One strategy that can be done is the development of roasted coffee market access to Asian countries**. This aims to cope with shorter roasted coffee durability. Currently, roasted coffee exports are still intended for the United States and Japan. Next, efforts are needed to expand the roasted coffee market access to Asian countries such as China, Singapore, and Malaysia.

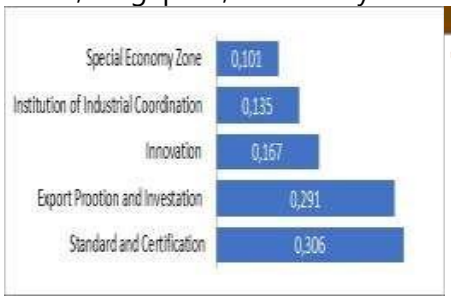


Chart IV.5 Perception of Trade Promotion Priorities

Chart IV.6. *Sumatran Roasted* Coffee Export Destination

ii.Incentive Framework

The main obstacle associated with macro incentives in the development of coffee commodities is the imposition of Plantation Taxes (10% VAT). With this policy, there is a tax obligation of Rp4.8 billion which is equivalent to 200 tons of Robusta coffee or 81 tons of Arabica coffee. These taxes certainly can influence coffee commodities in the way that it reduces margins at the farmer level.

In other coffee-producing countries such as Vietnam, there is a mitigation of tax rules for home businesses and plantation commodities. Some tax incentives that support the development of coffee exports are: 1) Exemption from income tax for home businesses in agriculture, 2) Exemption of 5% VAT for several plantation commodities, 3)2% subsidies of the increase in annual export value to exporters, 4) Reduction of land tax for agricultural and plantation companies up to 50%, 5) Irrigation costs are waived for farmers, and 6) Reduction of purchase tax for input factors by 5% (fertilizers, pesticides, plant medicines, etc.). Similar case with Brazil as it provides incentives in the form of tax mitigation for coffee and guaranteed prices. Some of the incentives provided are: 1) The final export tax of coffee is 2% to fund Funcafe - the Coffee Economic Defense/development Fund in Brazil (3 cent/lbs), 2) It is guaranteed to get a price of 85% from FOB, 3) Exemption of income tax and VAT from coffee exports, 4) the reduction of import taxes on raw goods and components is used for export products.

iii.Factors of Conditions, Rejuvenation and Replanting of Coffee

Increasing productivity of coffee can be realized through two things, replanting and rejuvenation. Replanting coffee plantation land is intended to replace old/unproductive plants aged over 20 years old with new plants. Meanwhile, rejuvenation is a partial pruning of stems from coffee plants aged 10-20 years old to increase the productivity of existing trees. Rejuvenating coffee plantations can increase productivity twice while replanting can increase productivity three times the current productivity.

Replanting coffee plantations takes 2.5-3 years until the trees produce optimally. Therefore, adequate funding is needed and crop diversification is needed to maintain farmers' income during replanting. Unlike palm oil, currently there is no national program related to replanting coffee. Funding that help replanting coffee is Special KUR (loans given by bank to businesses that are feasible but not bankable yet) funds that provide grace period.

Simulations conducted on replanting coffee for both Robusta and Arabica coffee showed results with a reasonable Internal Rate of Return (IRR). The simulation results for replanting 1 hectare Robusta coffee land require funds of Rp43 million to be financed through Special KUR. The IRR obtained by 15.7% is above the Sumatran investment interest in September 2018 which is 11.98%. Meanwhile, replanting Arabica coffee plantations yielded a much higher IRR of 38.9%.

Furthermore, rejuvenation of coffee plantations has the potential to increase land productivity up to 1.8 tons/ha from the current conditions of only 0.7-0.8 tons/ha. The area of land that can be rejuvenated with coffee is estimated to be 200 thousand ha. If this rejuvenation goes well, then farmers' income is expected to increase by 12.8%/year for Arabica farmers and 13.5% per year for Robusta farmers.

In addition, the loan to small businesses or Kredit Usaha Rakyat (KUR), a program that can be used as a financing source to increase agricultural productivity through replanting and rejuvenation. Several financial technology companies engaged in agricultural sector including plantations such as Igrow, Tanifund, Crowde, and Eragano.

iv.Improvement of Road Infrastructure and Production Support Facilities.

The problem in road infrastructure is access from production centers which is mainly in highland areas. The condition of plantation infrastructure in Sumatera is still far from perfect, unpaved roads in several regions are one of the problems and it hampers the production especially during rainy season. Therefore, the role of government is needed to improve access to production. The different conditions occur in Brazil and Vietnam with road access to production centers are relatively better. Meanwhile, in Vietnam, the government made infrastructure investments in five coffee-producing provinces with an 80% share of exports (Dak Lak, Dak Nong, Lam Dong, Gia Lai, Kon Tum) known as the *Vietnam Coffee Belt*. Meanwhile, to improve the quality of coffee drying, Dome Infrastructure as a cover for the coffee drying area is recommended.

v.Trade Promotion Infrastructure

From an institutional standpoint, a comprehensive institution is needed from upstream to downstream of the coffee sector by strengthening state owned enterprises as has been applied in Vietnam. Learning from coffee development in Vietnam, Sumatera may learn the marketing strategy in domestic and international markets. Also, learn from Vinacafe, a Vietnam state-owned company, who engaged in upstream to mainstream coffee chains who do the buying, processing and exporting coffee beans, roasted robusta as well as importing fertilizers. The strategy is to strengthen the role of state owned enterprise such as developing a core plasma scheme applied by smallholders, strengthening the role of PT Perkebunan Nusantara.

5. Discussion and Conclusion

Based on TOA, the addition of new export destination countries from Indonesia and Vietnam tends to be bigger in number than Brazil. In terms of quality, Indonesia's coffee export products are similar to Brazil and Vietnam and are dominated by Arabica. Base on AHP analysis, the main determinants to develop Sumateran coffee's competitiveness are 1). Market access, such as tariff barriers from export destination countries and non-tariff barriers such as certification standards; and 2). Condition factors to overcome low productivity, limited infrastructures, and lack of human resource capabilities; 3) the infrastructure for trade promotion and the macro incentive. We recommend to establish a comprehensive institution, from upstream to downstream of the

coffee sector chain, and strengthening the role of state owned enterprises as has been successfully applied in Vietnam. To increase productivity of coffee plantation through replanting and rejuvenation is considered most effective. Accordingly, farmers can use existing Special KUR funds. Meanwhile, for cost efficiency, it is necessary to support road infrastructure improvements in coffee plantations as well as the application of post-harvest technology through dome infrastructure such as drying houses. It is necessary to review the application of 10% VAT for plantation products as well as the macro incentives that can be given to farmers and coffee entrepreneurs as done by other coffee producing countries. Moreover, institutional support from the upstream to downstream parts of the coffee sector is also needed through strengthening state owned enterprise.

References

1. Bogliacino and Pianta, 2013. *The Dynamics of profits and wages* : *Technology Offshoring & Demand.*
2. UNCTAD, 2008. *Export Competitiveness and Development in LDCs.*
3. Farole, Thomas & Winkler, Deborah, 2012. *Export Competitiveness In Indonesia's Manufacturing Sector*
4. Limao & Venables, 1999. *Infrastructure, Geographical,Disadvantage & Transport Costs.*
5. World Bank, 2011. *Trade Competitiveness Diagnostic Toolkit*
6. Zhang, Kevin Hongling, 2014. *What Drives Export Competitiveness? The Role of FDI*

Household consumption expenditure patterns in Malaysia

Nurul Fatihah Mohamad Ariffin, Siti Nor Amalina Ghazali, Siti Rohani Anuar
Department of Statistics Malaysia

Abstract

The purpose of this study is to show the patterns of household consumption expenditure amongst Malaysians. The definition of household expenditure used in this survey is based on the concepts and guidelines by United Nations as published in A System of National Accounts, 2008 and Framework for Statistics on the Distribution of Household Income, Consumption and Wealth, 2013 published by the Organisation for Economic Co-operation and Development (OECD). Only 12 out of 13 household expenditure items were taken into account for this study, i.e. household consumption expenditure. Data collected were extracted from the Household Expenditure Survey 2016 report published by Department of Statistics Malaysia. A graph of composition of household consumption expenditure shows which expenditure items have the highest spending by household. The compounded annual growth rate (CAGR) of the mean monthly household consumption expenditure is also calculated.

Keywords

Household Expenditure Survey; Compounded Annual Growth Rate; Department of Statistics Malaysia

1. Introduction

The phrase 'cash is king' is relevant in a world where almost everything is doable if you have the money. This is why getting a job to get income doesn't only apply to adults but students too to survive in the ever-changing world where everything has its price. How much the people of a country spend contributes to the determination of Gross Domestic Product (GDP) of the country.

GDP is one of the measures of economic growth of a country in the sense of monetary expenditure (Divya K. H. & Devi V. R., 2014). There are three ways of calculating GDP of a country and one of them is through calculated expenditure, of which one of its components is the household expenditure (Landefeld J. S., Seskin E. P. & Fraumeni B. M., 2008).

Household refers to a group of people whether related or unrelated who usually live together in a living quarters and make provisions (expenses) for

food and necessities of life together. On the other hand, household expenditure can be broken down into two types; household consumption expenditure is the value of consumer goods and services acquired, used or paid for by a household through direct monetary purchase, own-account production, barter or as income in kind for the satisfaction of the needs and wants of its members, meanwhile household non-consumption expenditure refers to payments made by payers for services that cannot be identified and aimed to increase government revenue as well as payments that have no direct relation to the acquisition of services received, such as membership fees and gifts to charity donations.

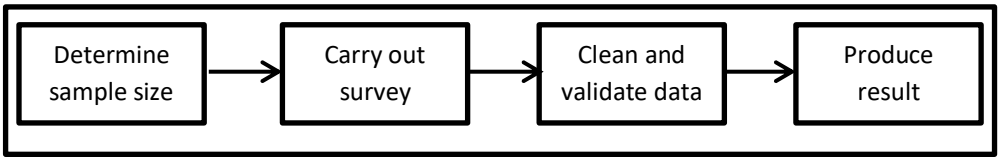
Household expenditure can further be classified into thirteen (13) main groups (United Nations, 2000). However, in this paper, only twelve (12) items in household consumption expenditure will be included, namely;

a) food & non-alcoholic beverages;
b) alcoholic beverages & tobacco;
c) clothing & footwear;
d) housing, water, electricity, gas & other fuels;
e) furnishings, household equipment & routine household maintenance;
f) health;
g) transport;
h) communication;
i) recreation services & culture;
j) education;
k) restaurants & hotels; and
l) miscellaneous goods & services.

There are three measurements of expenditure. The first one is acquisition, which is the value of purchasing during the acquisition of goods and services without taking into account whether they are fully used or not, or paid in full or not in the reference period. Acquisition was extended to include the estimated value of own production of non-durable goods and services and those given or received as its kind. The second measurement is consumption; an approach used in certain conditions for durable items that can last long and the use of utility such as water and electricity. Next is payment. This involves advance payment to obtain goods and services where the goods and services have yet to be received in the reference month. For example, the down payment we made when purchasing a car.

This paper aims to outline the patterns of household consumption expenditure in Malaysia as well as its CAGR between two survey years to show changes in how much Malaysians use their income to support their daily needs.

2. Methodology



Step 1: Determining the sample size

The process of determining the sample size is required to represent overall population at the analysis level. The sample size has been considered the following elements:

-Findings from previous HES report
-Level of sampling design; and
-Desired error

Procedures for estimating the sample size is calculated independently in each stratum (urban and rural). Simple Random Sampling Method (SRS) is used to take into account the design effect from the previous investigation, response rate and level of study. The optimum sample size was estimated at the level of EBs with regard to homogeneity characteristic variables and the costs involved.

Sample size calculation for sub population *j*, n_{1j} is calculated as in Equation 2.1:

$$n_{1j} = \frac{n_{0j}}{1+\frac{n_{0j}}{N_j}} \quad ; \quad \text{j = 1, 2, 3, …, k} \tag{2.1}$$

where;

N the number of element units in the population;

and n_{0j} can be calculated as:

$$n_{0j} = \frac{z^2 p_j(1-p_j)}{{d_j}^2} \tag{2.2}$$

where;

z level of confidence

p labour force participation rate

d desired error.

To satisfy the assumptions in the Stratified Sampling, Equation 2.3 was used where the design effect (D.E.) factor is taken into account:

$$D.E. = \frac{Variance\ for\ complex\ sample}{Variance\ for\ SRS} \tag{2.3}$$

Sample size taking into account D.E. for sub population *j*, n_{2j} is given as in Equation 2.4:

$$n_{2j} = n_{1j} \times D.E. \tag{2.4}$$

Next, taking into account the rate of response of the last survey, the overall sample size for sub population *j*, n_{3j} is as follows:

$$n_{3j} = n_{2j} \times \frac{1}{Response\ Rate} \tag{2.5}$$

Thus the total sample size, *n* was calculated by using Equation 2.6.

$$n = \sum_{j-1}^{k} n_{3j} \tag{2.6}$$

Step 2: Carrying out the survey

The survey was conducted face-to-face with respondents, or *isi rumah* (IR), in rounds of twelve; one round per month for the whole year, starting from January. Selected respondents were visited three times or more per week by interviewers to collect information on demography, income and the expenditure by classification of goods and services using a set of questionnaires.

Step 3: Cleaning and validating data

Quality checks were made by experienced officers from the DOSM state office to detect and correct any error or omission of the data. The review processes were also implemented for selected household to ensure the quality of the collected data.

Step 4: Producing results

Analysis of data was done by first separating the expenditures into respective group and providing a graph that shows the composition of each group from the total mean of monthly household expenditure by selected year. Then, the compounded annual growth rate (CAGR) of household expenditure was calculated and compared. CAGR can be calculated based on the exponent function as follows:

$$CAGR = \frac{\ln \frac{Y_t}{Y_0}}{t} \tag{2.7}$$

where

Y_t current year household monthly expenditure

Y_o previous year household monthly expenditure

t period between two subsequent survey year.

3. Results

The main purpose of this paper is to lay out the patterns of Malaysian household consumption expenditure according to their respective expenditure group. The data collected are from the year 1993/94 until the most recent 2016. The table below shows the composition of the mean monthly household consumption expenditure by expenditure group.

Expenditure group	1993/94*	1998/99*	2004/05	2009/10	2014	2016
	RM					
Food and non-alcoholic beverages	276	368	393	444	676	726
Alcoholic beverages and tobacco	26	30	35	48	83	98
Clothing and footwear	41	56	59	75	124	136
Housing, water, electricity, gas and other fuels	245	363	430	495	853	969
Furnishings, household equipment and routine	65	84	83	89	137	168
Health	21	29	27	29	59	75
Transport	168	227	314	327	523	553
Communication	24	59	103	124	189	203
Recreation services and culture	53	70	92	101	174	200
Education	17	31	38	31	41	54
Restaurants and hotels	145	209	213	239	454	540
Miscellaneous goods and services	78	105	167	190	266	312
Mean monthly household consumption expenditure	**1,161**	**1,631**	**1,953**	**2,190**	**3,578**	**4,033**

Table 3.1: Composition of mean monthly household consumption expenditure by expenditure group

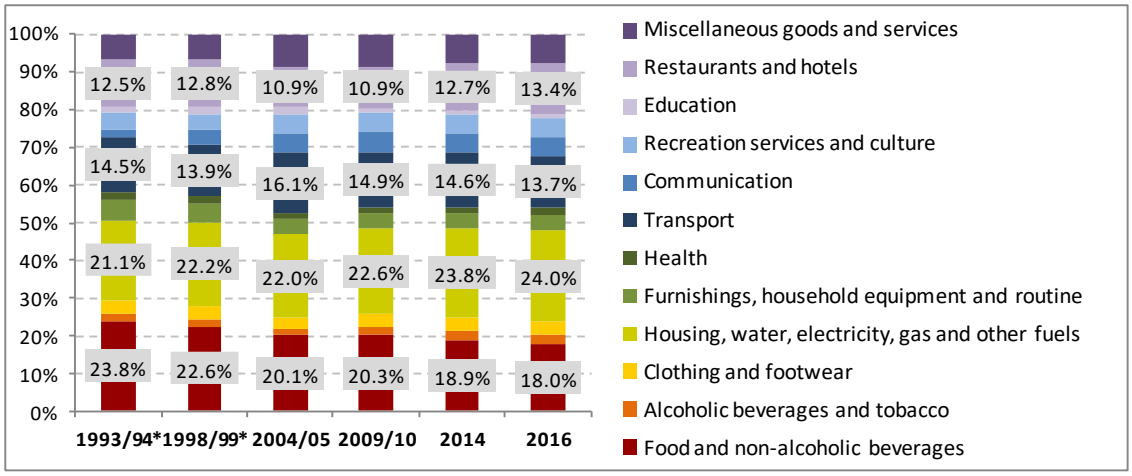


Figure 3.1: Composition of mean monthly household consumption expenditure by expenditure group

The table below shows the components or subgroups for the highest spending for expenditure group by household, which is housing, water, electricity, gas and other fuels for the year 2016.

Subgroups of expenditure	Amount (RM)	Percentage (%)
Imputed rent	580.77	59.93
Rental paid	153.75	15.86
Electricity	114.19	11.78
Water supply	37.70	3.89
Services for the maintenance and repair of the dwelling (including materials)	37.63	3.88
Gas	19.23	1.98
Materials for the maintenance and repair of the dwelling	15.75	1.62
Other services relating to the dwelling n.e.c.	7.12	0.73
Sewage collection	2.19	0.23
Refuse collection	0.38	0.04
Other fuels	0.32	0.03
Liquid fuels	0.11	0.01
Total	**969.13**	**100.00**

Table 3.2: Subgroups for housing, water, electricity, gas and other fuels for the year 2016

The CAGR between two subsequent survey years were also calculated at nominal value using Equation 2.7 and then recorded as follows:

Year	1998/99*	2004/05	2009/10	2014	2016
CAGR	6.8%	4.5%	2.3%	9.8%	6.0%

Table 3.3: CAGR of mean monthly household consumption expenditure

4. Discussion and Conclusion

The CAGR shows a decline in growth from the survey year 2004/05 to 2009/10 as well as from 2014 to 2016. In 2016, the mean monthly household consumption expenditure for Malaysia increased from RM3, 578 in 2014 to RM4, 033 in 2016 which grew at 6.0% per annum at nominal value. In terms of real value which refers to constant price using the Consumer Price Index (CPI) with the base year 2010 as the deflator, annual growth rate is 3.9% for the same period.

Based on Figure 3.1, it can be seen that for the first two survey years, expenditure item food and non-alcoholic beverages tops the composition with 23.8% in 1993/94 and 22.6% in 1998/99. However, for the next four years, Malaysians were seen to spend the most on expenditure item housing, water, electricity, gas and other fuels with 22.0%, 22.6%, 23.8% and 24.0% respectively. Additionally, referring to Table 3.2, the highest contribution to

this spending in 2016 is the subgroup imputed rent, with a percentage of 59.93 from the total amount spent, and followed by rental paid (15.86%).

In spite of that, the expenditure groups with the highest spending by household have been consistent every year. They would spend the most for the following five expenditure items:

1. Housing, water, electricity, gas and other fuels
2. Food and non-alcoholic beverages
3. Transport
4. Restaurants and hotels
5. Miscellaneous goods and services

In conclusion, the decline in CAGR from 1998/99 to 2016 showed that Malaysian spent less on household consumption expenditure. However, looking at the household consumption expenditure groups individually, some of the groups still showed an increment in mean monthly household consumption expenditure.

References

1. Divya K. H. & Devi V. R. (2014). A Study on Predictors of GDP: Early Signals. Procedia Economics and Finance.
2. Landefeld J. S., Seskin E. P. & Fraumeni B. M. (2008). Taking the Pulse of the Economy: Measuring GDP. Journal of Economic Perspectives
3. United Nations. (2000). Classification of Expenditure According to Purpose. United Nations Publication.

Assessment in an introductory statistics course – The challenge of consensus

Lina Schelin, Jessica Fahlén
Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

Abstract
In this work, we present results and insights from a pedagogical project carried out at the Department of Statistics at Umeå University, Sweden. The overall aim of the project was to reach a higher level of consensus within assessment teams assessing students' knowledge and skills in an introductory statistics course. Specifically, the focus was on the part of the course that was assessed using a written exam. The pedagogical project consisted of two parts that we present in more details; focus group discussions and the development of an examination guide. The main conclusions of the project are that there is room for improvement regarding the level of consensus, that an examination guide can be an aid working towards consensus, and that it is important to have a continuous discussion about assessment within the assessment teams.

Keywords
Statistical reasoning; statistical thinking; pedagogical project; written exam; examination guide

1. Introduction

The interest in statistics education has grown during the last 30 years with new ideas on how students learn, what topics that are important and how these topics should be taught (Moore, 1997; Garfield & Ben-Zvi, 2007). A theoretical framework for describing statistical understanding has been developed by Ben-Zvi and Garfield (2004), where they describe the concepts of statistical literacy, statistical reasoning, and statistical thinking. An earlier framework that defines five dimensions of understanding is presented in Putnam, Lampert and Peterson (1990). Overall, this educational reform, with a focus on understanding, has resulted in that more student active learning methods have been proposed and used. This is also the case for introductory statistics courses at our university. Several student active learning activities have successfully been incorporated in our courses.

As the objective of introductory statistics courses partly has shifted, along with the learning methods, the natural continuation of such process is to also change the assessment techniques (Chance, 1997). More authentic assessment techniques (e.g., journals, team projects, minute papers, portfolios) may be better suited to assess students statistical thinking and

reasoning (Garfield & Chance, 2000). But, as described in Chance (2002) also traditional examinations can be used to assess such type of understanding. For instance, to address the challenges identified by Garfield and Gal (1999) a test designed to measure students' conceptual understanding of important statistical ideas has been developed (delMas, Garfield, Ooms, & Chance, 2007).

In our introductory courses, we use several assessment techniques, including a written exam. The written exam consists of several tasks, and during the marking of the exam each task is rewarded with a point. The total score on the written exam determines a student's grade on the whole course. Garfield (1994) states that "*the primary purpose of any student assessment should be to improve student learning*". Although this is something we definitely agree on, it is not the primary purpose of the written exams in our courses. The total score on the written exam, is instead used as an indicator of each student's success in reaching the expected learning outcomes of the course and used for grading.

One other important aspect of all assessment is that it should be consistent and fair (Chance, 1997). This can be difficult, especially when several teachers are involved in the assessment; hence, *the challenge of consensus*.

At our department, the marking of the written exams on introductory courses, is performed individually by several teachers in assessment teams; usually the teachers correct 1-2 tasks each. During the years, there have been several cases where we during such marking processes have noted that the corrections deviate between different teachers, and between different assessment teams. The tendency is that this is a problem that grows with the size of the assessment teams. To prevent such situations and to further develop the written exams, we initiated this project where the main aim was to reach a higher level of consensus among the teachers at our department in the assessment process. Prior to the start of the project, preliminary discussions among the faculty indicated that the assessment teams usually agreed on the overall grade of an exam, while at the same time they did not agree on all details. Hence, a shift of focus from awarding each task on the exam with a point (on a rather fine scale) towards something that is similar to a rubric would be beneficial. We had previously tried to employ a rubric, with less successful results. Basically, it was too difficult, also for the more experienced teachers, to apply. The main reason why we want the shift towards rubrics is that it enables a holistic grading. In our experience, awarding points on a fine scale, does not provide a good measure of the students' ability of statistical reasoning and thinking. Instead, it awards students with fragmented knowledge.

We will now present the course and the written examination in more detail before presenting the project, the results and some conclusions.

2. The course

Here, the focus is on the first module of the introductory course Statistics A1. This course is the single largest course at the department. It is taught four times every year with 150, 70, 60 and 150 students respectively, and involves more than 10 different teachers. The expected learning outcomes for the first module of the course are:

- describe the basic statistical terminology that is relevant for each respective module of the course
- apply statistical methods such as point estimation, confidence interval and hypothesis testing
- based on a stated question and the scales of measures choose an appropriate method of statistical analysis among those presented in the course
- interpret derived statistical results and contextualize them

The core of every introductory statistics courses should be statistical thinking and reasoning (Garfield, 2002). The expected learning outcomes were constructed with this in mind. Further, the written exam is supposed to assess these expected learning outcomes.

3. The written exam

The written exam on Statistics A1 is performed on paper, in an exam room, during a maximum of four hours. The maximum score on the exam is 35 and the students have to obtain a score of 21 to be awarded the grade Pass. A score over 28 implies the grade Pass with distinction. Each task has a maximum point and a student is usually awarded a point between 0 and the maximum point, in increments of 0.5 or 1. A smaller part of the exam is related to terminology, but most tasks are such that they focus on statistical thinking and reasoning. For one example of a task, see Figure 1. The assessment teams are expected to discuss and agree on a marking template before marking the exams.

Task 4 (6 points)

Using recycled materials for new products is environmentally friendly and saves resources. However, a common perception is that products from recycled materials have lower quality compared to completely new products.

In order to examine consumers' perception of the quality of coffee filters made from recycled paper, a survey of 133 randomly selected consumers was made. Respondents, among other things, answered the question of whether coffee filters from recycled paper are of *higher*, *equivalent* or *lower* quality compared to coffee filters from non-recycled paper.

The survey showed that 36 of the 133 respondents buy coffee filters made from recycled paper and 97 did not.

a) Is there any empirical support for an association between the perception of the quality of coffee filters made from recycled paper and whether one buys such filters? Perform an appropriate test at 5% significance level. Describe and motivate all steps in the hypothesis test based in the information in the Minitab output below. The decision rule you use should be based on the critical value (do not use the p-value method).

	Higher quality	Equivalent quality	Lower quality	All
Buys filters from recycled paper	20	7	9	36
Does not buy filters from recycled paper	29	25	43	97
			Total	133

Pearson Chi-Square = 7,638

b) The observed value of the test statistic is 7,638. State the p-value for the test as accurately as possible. Make a drawing of the distribution of the test statistic when the null hypothesis is true and color what area that correspond to the *p*-value.

c) Calculate the expected number of people who buy coffee filters made from recycled paper but who think that coffee filters made from recycled paper are of lower quality assuming that there is no relation between perception of the quality of coffee filters made from recycled paper and whether one buys such filters.

Figure 1. *One example of a task from a written exam.*

4. The project

As mentioned, the main aim of the project was to reach a higher level of consensus among the teachers at our department, especially for the examination of the Statistics A1 course. The two distinct parts of the project will be described in more detail next.

Focus group

The aim of the focus group discussions was to use so called "social moderation" to better understand how teachers in the assessment teams categorize different mistakes in students' solutions of exam tasks. Social moderation, here, imply teachers working together when assessing and grading (Thornberg & Jönsson, 2015). All participants were asked to mark and grade two students' solutions (of the whole exam) before the focus group meeting. Not all participants corrected the same solutions. In total, the group corrected six different exams. The participants were also asked to write down how they were thinking during their correction process, to better take use of the two hours allocated for the focus group discussions. We closely followed the discussions in the group to understand when the teachers were agreeing or not, and if they came to a consensus on the different student solutions or not.

Examination guide

Based on the discussions in the focus group an examination guide was developed. Particular focus was given to those things that was difficult to reach consensus about during the discussions. A preliminary examination guide was tested, in a pilot study, during the marking of the exams in the course given in the second part of the spring semester 2018. A sample of the exams was marked using the instructions in the examination guide, in parallel to the ordinary marking in the course.

5. Results and discussion

The results of the focus group discussions were that the participants usually agreed on the overall grade of the exams, but that the exact points for each task could differ quite a lot. Some of our teachers are using a holistic approach when correcting, while others use a more fragmented approach (giving points to specific parts without an overall perspective of the solution). Since the grades are awarded based on the total score on exam, students that do not reach the expected learning outcomes are sometimes awarded the grade Pass, or are close to being awarded grade Pass, when the more fragmented approach is used. This problem might be reduced by not allowing such a fine resolution of point increments as we do now. With only a few predefined levels (or points) the teachers are forced to categorize the students' solution taking the overall perspective into account. Hence, this is something we decided to incorporate in the examination guide.

During the discussions, we could also identify some other aspects that should be addressed in an examination guide, mainly related to trivial things, but still of importance for the students. For example, we realized that some

teachers use a symbol to indicate that something is correct, while others use the same symbol for the opposite.

The most relevant part of the examination guide that we constructed, based on the focus group discussions, is the assessment levels (see Table 1). This is our suggestion on how to go towardsmarking with a general rubric, while still being able to award each task with a point. The example of a task given in Figure 1 contains three sub-questions. Our intention is thus that the assessment levels in Table 1 is used to award the whole task with a point (level), since all sub-questions cover the same "topic". A holistic assessment is simplified if a templet is used in the construction of the exam. Hence, the examination guide also contains information related to the how the exam should be constructed (i.e., a short templet).

Table 1. *The part of the examination guide that describe the new suggested predefined levels used for marking the exams. Each level has a short description of the criteria that should be met, and each level corresponds to a point, resulting in points with less resolution than we presently use.*

LEVEL 1	POINT	DESCRIPTION*
Level 1	6	Correct (maximum point)
Level 2	5	Correct, but with a minor mistake.
Level 3	3	Some mistakes, but the solution indicate some understanding.
Level 4	1	Not correct, but the solution includes something to reward.
Level 5	0	Not correct, and includes nothing to reward

*A list exemplifying minor and major mistakes should be attached as a supplement.

During the project, the examination guide was tested in parallel to the ordinary marking on the course. A simple random sample of 30 exams were marked by us according to the instructions in the examination guide. The results were compared to the ordinary marking made by the assessment teamon the course, see Figure 2.

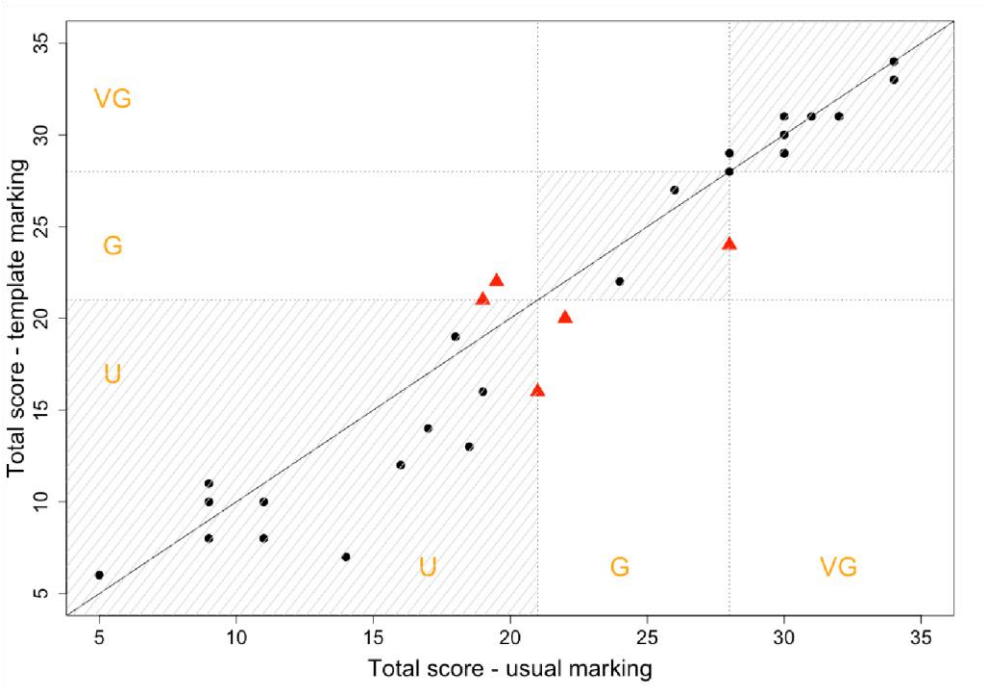


Figure 2. *The results of the parallel marking using the guide (y-axis) in relation to the correction made by the assessment team on the course (x-axis). Filled circles imply identical grades (U, G or VG) on the exam, while triangles imply that the final grades differ. The grade U corresponds to not passing, the grade G corresponds to Pass, and the grade VG corresponds to Pass with distinction.*

The experience of marking the exams using the assessment levels in Table 1 indicated that it makes marking more time efficient. Students that do not show that they fulfil the expected learning outcomes should not pass the exam. Still, it can happen that they get a score close to the numerical threshold for passing. This often leads to discussions, which are also time consuming. We hoped that using the guide would give less scores to students that don't pass the exam, and this seemed to be the case (c.f., Figure 2).

6. Conclusion

Although this project was related to one specific written exam of one specific introduction course, it could be adapted to other courses as well. We will definitely continue with regular discussions related to how we mark exams (similar to the focus group discussion), since we, and all participating persons in the discussions found it both meaningful and important.

The examination guide has so far been tested in a pilot study, with promising results. Specifically, we found it to be very time efficient. Not only did it save time during the actual marking, but we also believe that it will save time afterwards, since less students are close to Pass. Usually, at our

department, too much time is spent on discussions with students, that are close to Pass, after the exam has been marked.

Relevant parts of the examination guide were also presented for the students during a new teaching activity within the course. During the activity the students was supposed to use parts of the guide to mark four authentically solutions from previous students, as well as their own solution. The interaction with the students did not lead to any changes of the examination guide, but it was a successful experience, and has now been implemented as a regular teaching activity during the course.

Changing tradition takes time and could be challenging. Further, changing the conditions too often can be stressful for both students and faculty. Hence, before using the experience from this pedagogical project, and implement the examination guide in full scale, we want to have a well-tested and approved concept. Hence, the project continues, but in other forms.

References

1. Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Springer, Dordrecht.
2. delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 8-58.
3. Chance, B. L. (1997). Experience with authentic assessment techniques in an introductory statistics course. *Journal of Statistics Education*, *5*(3).
4. Chance, B. L. (2002). Components of statistical thinking and implications for instruction and
5. assessment. *Journal of Statistics Education*, *10*(3).
6. Garfield, J. (1994). Beyond testing and grading: using assessment to improve student learning. *Journal of Statistics Education*, *2*(1). Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*(1-2), 99-125. Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International statistical review*, *75*(3), 372-396.
7. Garfield, J. & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, 67, 1-12.
8. Moore, D. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review,* 65 (2). 123-137.
9. Putnam, R. T., Lampert, M., & Peterson, P. L. (1990). Alternative Perspectives on Knowing
10. Mathematics in Elementary Schools. In C. B. Cazden (Ed.), Review of Research in education (pp. 57-150). Washington, DC: AERA.
11. Thornberg, P., & Jönsson, A. (2015). Sambedömning för ökad likvärdighet?. *Educare-Vetenskapliga skrifter*, (2), 179–205.

Using predictive modelling to identify panel nonresponse

Jan-Philipp Kolb[1], Bernd Weiß[1], Christoph Kern[2]
[1] GESIS Leibniz Institute for the Social Sciences
[2] University of Mannheim

Abstract
Panel surveys are a valuable source of data to investigate a wide range of research questions. However, data quality can be negatively affected by nonresponse. Unit nonresponse is most critical when it is due to selective nonresponse patterns, which can lead to biased estimates. It is therefore essential to identify panellists with a high risk of nonresponse. If it is possible to locate these panellists, we could apply interventions in an adaptive survey design to motivate them to further participate in the panel study. However, identifying potential non-respondents is a challenging task given the wealth of information typically available in panel studies. In this study, we aim to utilize statistical learning methods with a diverse set of predictor variables to tackle panel attrition from a prediction perspective. We study nonresponse in the GESIS Panel, which is a bi-monthly probability-based mixed-mode access panel of the German population (n ≈ 4,700). In addition to socio-demographic and substantive variables, process-based para-data, as well as data from the panel management, are used as predictors. Feeding this information to supervised statistical learning methods offers a promising avenue for building a useful nonresponse prediction model, as these methods allow to model complex relationships across many features without the need of specifying the models' functional form in advance.

Keywords
Machine Learning; Panel Survey; Nonresponse; Feature Selection; Ensemble Methods

1. Introduction

Unit nonresponse can become a severe problem if it occurs due to patterns. It is the case, when the unit nonresponse is not completely random. This needs to be investigated, and especially in Panel Surveys, many variables need to be tested. This is where statistical learning methods come into play. These methods have their advantages when dealing with such a large number of variables. In this paper, statistical learning methods are tested using the Unit Nonresponse in the GESIS Panel.

The GESIS Panel is a *probability-based* mixed mode access panel (Bosnjak et al. 2018). Probability-based means that the panel participants were selected

from the German population register by a random sample. Every two years a panel refresher is carried out. The German general social survey (ALLBUS) is used for the data restocking. Except for an oversampling in Eastern Germany, the ALLBUS has the same sample design as the GESIS Panel. Participants in the ALLBUS are asked if they are willing to participate in the GESIS Panel. A restocking takes place every two years. So the first cohort is made up of the panellists who have been participating since the beginning of the panel. In the second cohort are people who have participated in the survey since 2016, and the panellists who have just been recruited are part of the third cohort.

Mixed mode means that there are two groups of participants-- the panellists have the choice between an offline questionnaire and an online version. These two modes ensure for example that people without an internet connection are also represented in the panel. It does not matter if they do not have the internet because of conviction or age.

Access panel means that researchers can submit study proposals. These proposals are then peer-reviewed, and if the review is positive, the submitters have a slot of about five minutes in the panel at their disposal. The proposed study can be either cross-sectional or longitudinal. The GESIS panel started in 2013 with more than 32 waves and over 40 studies running in the panel since then. The data is available as scientific use file, and it is also possible to use an extended version in the GESIS secure data center.

Our target is to predict unit nonresponse in the GESIS Panel. More specifically, we want to predict whether a person has a high likelihood to fall into one of the following AAPOR categories (Smith and others 2004).

- 211 Refusal
- 212 Break-off: Questionnaire too incomplete to process
- 319 Nothing ever returned
- 2112 Explicit refusal
- 211221 Logged on to survey, no item complete

On this basis, we e.g. flag a unit nonresponse if a person breaks off at the start so that the questionnaire is too incomplete to process. The result is a binary variable, which has value one for non-respondents and zeroes for respondents.

Potentially we could use more than 4000 variables as predictors. As a first block of potential predictors, we have substantive survey information from studies like "Lifestyles in everyday life" with study token aa or "threat perception and political trust" (study token bi). Examples for these variables are age, gender or educational attainment of the panelists, as well as distance from place of residence to a large city (Kolb and Weyandt 2018). Some of the variables are collected regularly if the study is longitudinal. Other variables are only available at one point in time.

Furthermore, a survey evaluation is carried out at the end of each wave questionnaire. The panelists are for example asked if they had difficulties with questionnaire comprehension or with finding an adequate answer on a question. Also, further survey-related questions are asked. Survey participation, for example, gathers the information whether the person participates in other surveys in addition to the GESIS Panel.

Also, it is possible to access process-based para-data because the GESIS panel is self-administered. We have information on, e.g. how long it took the online panelists to answer one specific question. However, some of these variables are only available for the online part. Nonetheless, this type of information might be of particular importance, since the impact of para-data for predicting nonresponse was highlighted recently (Lugtig and Blom 2018).

In addition, we have information about panel management. The GESIS panel maintains all contacts with panelists in a panel protocol database. This includes, for example, whether there was a complaint or concern from a panelist. For online users, it records whether they have identified a technical problem with the online portal. It is then also recorded whether the technical problem could be solved. Some panelists also have problems with the content of surveys and express these. All in all, this information provides a very interesting picture. In some cases it is quite easy to explain why there is a case of unit nonresponse. However, it must also be noted that the number of cases is unfortunately relatively small and that these variables therefore do not play a major role in the statistical learning models.

Another example of an administrative variable is the mode of invitation. Two ways of participation are available, online and offline. We also generated the variable survey participation, to control the frequency and regularity a person responds to the survey. We divided the number of waves a panelist responded by the number of waves he could have participated. This takes into account that panelists who have been in the panel since 2016 were able to participate in fewer waves than panelists who have been in the GESIS panel since 2013. We use wave fa for the analysis in this paper. Another interesting variable in this context is the latency. This variable is used to record how long it takes the panelists to respond to the invitation to participate in the questionnaire. This can be recorded relatively precisely for online users. For offline panelists, the time span between the invitation and the day on which the questionnaire is filled out and returned is recorded.

2. Methodology

We applied two types of techniques. On the one hand, we have parametric methods like logit or lasso regression. That means that, in this case, we have a predefined additive and linear functional form. In the model, we use a logarithmic function for the link between probability and logits. Thus, we have

the linearity between predictors and log odds. For the Logit model, the search routine for adequate predictors is external to the model. These techniques aim to determine whether a particular independent variable affects the dependent variable. If there is an effect, it is the target to estimate the magnitude of that effect. We have the following loss function that should be minimized (Pereira, Basto, and Silva 2016):

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x_i \beta}\right) \right]$$

y_i is the binary outcome and β the vector of regression coefficients to be estimated. The outcome variable takes the value one if the panelist has not responded. It would take the value zero if the panellist responded.

When we use all the information available, logit regression becomes unfeasible given the dimensionalty of the prediction problem . Least absolute shrinkage and selection operator (LASSO) (Kyrillidis and Cevher 2012) is a feature selection method (T. Hastie, Tibshirani, and Friedman 2009). LASSO regression has inbuilt penalization functions to reduce overfitting. The loss function that should be minimized (Pereira, Basto, and Silva 2016) is then displayed in the nest formula:

$$l_\lambda(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - log\left(1 + e^{x_i \beta}\right) \right] - \lambda \sum_{j=1}^{p} |\beta_j|$$

x_i the i-throw of an matrix of n observations with p predictors. β is the column vector of the regression coefficients. y_i is the binary outcome and λ is the shrinkage parameter.

We also applied tree-based methods where the built-in feature selection combines the predictor search algorithm with the parameter estimation. In these cases, the estimation is usually optimized with a target function like the likelihood. By using decision trees, we go from observations about an item represented in the branches to conclusions about the item's target value which is represented in the leaves. The predictor space is recursively split into disjoint regions R_j

$$T(x; \Theta) = \sum_{j=1}^{J} \gamma_j \, I(x \in R_j)$$

with the tree parameters $\Theta = \{R_j, \gamma_j\}$. In the following the applied tree-based techniques are listed.

Conditional inference trees (ctree) use significance test procedures to select variables instead of selecting the variable that maximizes an information measure (Hothorn, Hornik, and Zeileis 2015).

The random forest (rf) technique generates and combines different decision trees (Breiman 2001). The goal is to get a more accurate and stable

prediction. The resulting "forest", is an ensemble of decision trees. Normally the "bagging" method is used for building an ensemble of trees. Random Forest adds additional randomness to the model while growing the trees. The algorithm does not search globally for the optimal feature, but for the best feature from a random subset of characteristics when the node is split. This proceeding often leads to better models.

In this paper, we also apply gradient boosting machines (GBM) which grow e.g. a sequence of trees that use updated residuals (J. Friedman et al. 2000). Gradient boosting is often used since a single decision tree fails to include predictive power from multiple, overlapping regions of the feature space. . In boosting, an ensemble of classifiers is built incrementally. In each step, a new sub-model is added that tries to compensate for the errors made by the other sub-models applied previously.

We used the R programming language to implement the outlined methods (R Development Core Team 2008). The R package caret was used to train and test the models within a statistical learning environment (Kuhn et al. 2018). We used exhaustive grid search to tune the hyper-parameters of the various predictive models (Kuhn and others 2008).

3. Result

In a first step we have looked at the importance of the used variables. In all statistical learning methods, the latency is important. For the lasso method it is the most important variable. Other important variables are the cohort, the number of complaints or the mode. The importance rank for the variables varies across the different statistical learning techniques.

In Figure 1 the classifier performances are displayed. We have results for four statistical learning techniques on the y-axis and five performance indicators in the five grids of the graphic. The number of correct predictions divided by the number of all predictions is the accuracy. We have a rather high accuracy for all techniques. However, since we have unbalanced data (more respondents than non-respondents), this indicator is by itself not particularly meaningful.

In the second grid, Cohen's Kappa is displayed as a performance measure on the x-axis. We used that measure because we have very imbalanced classes. Cohens Kappa tells us how much better our classifier is performing over the performance of a classifier that guesses at random according to the frequency of each class. Cohen's kappa is bounded by zero and one (although values smaller than zero can occur). Values close to zero or below indicate that the classifier is useless.

The receiver operating characteristic (ROC) is an excellent way to visualize the performance of a classifier and to select a decision threshold (Bradley 1997). We create the ROC curve by plotting the sensitivity against the false

positive rate (FPR) at various threshold settings. It is possible to calculate an indicator (ROC-AUC) for that curve, and the values for this indicator are also bounded between zero and one. Values over .5 indicate that the used methods perform better than flipping a coin.

The sensitivity (Sens) also called the true positive rate (TPR) measures the proportion of actual positives that are correctly identified as such. As we have many more respondents than non-respondents, the values for this sensitivity indicator are very close to one in our case.

Specificity (Spec) is also called the true negative rate. It measures the percentage of real negatives that are correctly identified as such. Here the performance varies for the statistical learning techniques.

Sensitivity and specificity are inversely proportional to each other. So when we increase specificity, sensitivity decreases, and vice versa.

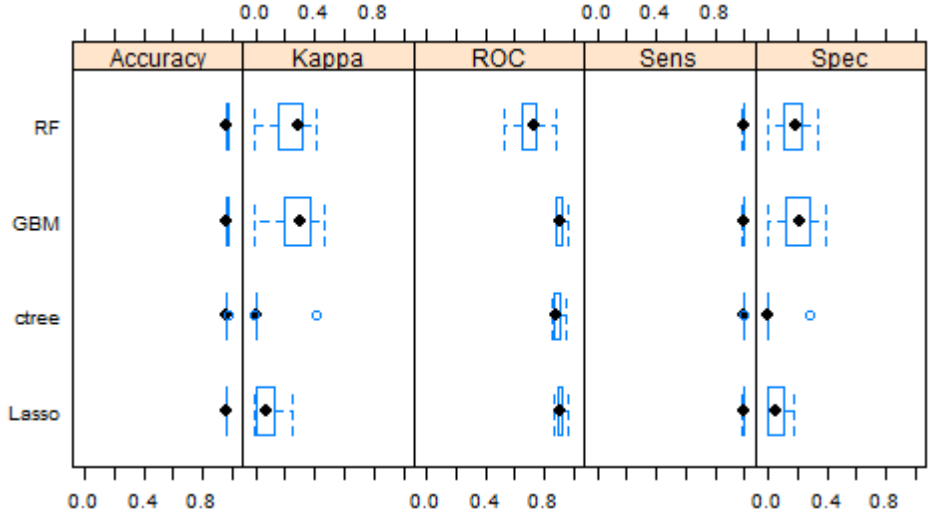


Figure 1 Boxplot with results for four methods and five metrics

4. Discussion and Conclusion

The present study investigates the potential of using features from diverse sources for predicting nonresponse in the GESIS Panel. We investigated which predictors do play a role for the risk of nonresponse. Also, different statistical learning techniques are employed and compared regarding prediction accuracy. We can use the results of this approach as a guideline for developing a useful model for predicting panel nonresponse in advance.

More specifically, preliminary findings suggest that variable importance varies across different statistical learning techniques. Para-data like the latency is essential for the prediction of panel nonresponse. Random forests exhibit the best results regarding precision and recall.

In further analysis, we plan to take the longitudinal character of the GESIS panel more into account. So far we divided one panel wave into test and

training data. In the future, we want to use the complete data as training data and the next wave dataset as test data. Also, we want to consider more information about the questionnaire structure. For example, it would be possible to use the information on questionnaire scales and their characteristics. How significant is the number of questions classified as sensitive? How many open questions are in the questionnaire? We further plan to analyse nonresponse patterns.

References

1. Bosnjak, Michael, Tanja Dannwolf, Tobias Enderle, Ines Schaurer, Bella Struminskaya, Angela Tanner, and Kai W Weyandt. 2018. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The Gesis Panel." Social Science Computer Review 36 (1). SAGE Publications Sage CA: Los Angeles, CA: 103–15.
2. Bradley, Andrew P. 1997. "The Use of the Area Under the Roc Curve in the Evaluation of Machine Learning Algorithms." Pattern Recognition 30 (7). Elsevier: 1145–59.
3. Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1). Springer: 5–32.
4. Friedman, Jerome, Trevor Hastie, Robert Tibshirani, and others. 2000. "Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors)." The Annals of Statistics 28 (2). Institute of Mathematical Statistics: 337–407.
5. Hastie, Tibshirani, R Tibshirani, and J Friedman. 2009. The Elements of Statistical Learning. NY: Springer. doi:https://doi.org/10.1007/978-0-387-84858-7.
6. Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2015. Ctree: Conditional Inference Trees. Vignettes of the Comprehensive R Archive Network. https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf.
7. Kolb, Jan-Philipp, and Kai Weyandt. 2018. "GESIS Panel Incremental Codebook - Related to ZA5664 and ZA5665 (23-0-0)." Codebook. Mannheim: GESIS.
8. Kuhn, Max, and others. 2008. "Building Predictive Models in R Using the Caret Package." Journal of Statistical Software 28 (5): 1–26.
9. Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, et al. 2018. Caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret.
10. Kyrillidis, Anastasios, and Volkan Cevher. 2012. "Combinatorial Selection and Least Absolute Shrinkage via the Clash Algorithm." In Information Theory Proceedings (Isit), 2012 Ieee International Symposium on, 2216–20. IEEE. doi:10.1109/isit.2012.6283847.

11. Lugtig, P., and A. Blom. 2018. "It's the Process Stupid! Using Machine Learning to Understand the Relation Between Paradata and Panel Dropout."
12. Pereira, Jose Manuel, Mario Basto, and Amelia Ferreira da Silva. 2016. "The Logistic Lasso and Ridge Regression in Predicting Corporate Failure." Procedia Economics and Finance 39. Elsevier: 634–41.
13. R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.
14. Smith, TW, and others. 2004. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys." Lenexa, KS: American Association for Public Opinion Research.

Tagged and numbered: Big data and central banking

Pamela Kaye A. Tuazon[1]
Bangko Sentral ng Pilipinas

Abstract

Big Data is a current global buzzword which gained traction due to its novel sources, processing, and applications. Highfalutin words associated with Big Data include algorithms, machine learning, artificial intelligence, coding and programming. However, the true value of Big Data is not the deluge of information but rather the ability of the end-user to calibrate and harness insights to guide future strategies and policies. Central banks worldwide are incorporating Big Data as an unconventional supplement to official statistics to produce leading and near-real time economic and financial indicators. The paper is organized as follows: Section I provides an overview of Big Data, Section II explores Big Data in the Philippines, Section III elaborates on the potential applications of Big Data in Central Banking, and Section IV concludes with the challenges ahead and recommends routes for the Big Data pilot project in the Philippine central bank.

Keywords

Smart Data; Nowcasting; Digitalization; Real-time market surveillance; Leading indicators

1. World Digitalization and Big Data

Imagine a world tagged in numbers – every footstep, every tap of your transportation card in the bus and train stations, every email sent, every tweet, every swipe of your credit card, or even just your mere location – every tiny deed leaves digital crumbs that allow analysts to predict behavior and recommend future strategies.

Indeed, the emergence of the knowledge economy rests upon the digitalization of operations, physical infrastructures, and even human interactions resulting into information-rich digital footprints. These datasets when harnessed and analyzed, easily provides both policy-makers and businesses with ample real-time insights. Big Data, therefore, became the

[1] Ms. Pamela Kaye A. Tuazon (Bank Officer IV) from the Department of Economic Statistics from the Bangko Sentral ng Pilipinas and is part of the Department's ad hoc and interim Big Data Team. Errors and omissions are the sole responsibility of the author and do not necessarily reflect that of the BSP.

catchword for voluminous, structured and unstructured data sets, which are hard to process, analyze, and store using traditional means.

The *International Monetary Fund (IMF)'s 2017 Staff Discussion Note on Big Data* posited that there is no agreed-on definition of Big Data yet.[2] However, it is usually characterized by the 3Vs of **high volume** (Exabyte data), **high-velocity** (speed of data), and **high-variety** (mixed sources and types).

According to the *United Nations Economic Commission for Europe (UNECE),* the sources of Big Data ranges from **Social Networks** (human-sourced information), **Traditional Business Systems** (process-mediated data), and **Internet of Things** (machine-generated data). Due to the complex nature of Big Data, advanced technologies and skills set are needed to transform these random numbers into valuable insights. Hence, Big Data should not just be viewed simply as a large set of raw information, but should be processed and analyzed to achieve its end-purpose of guiding decision-making – towards Smart Data (Marr, 2015).[3] The unprecedented flooding of information from various sources may seem daunting and overwhelming, but the value of Big Data is its capacity to produce **actionable intelligence**.[4]

2. Big Data in the Philippines

In the 2018 Global Digital Report[5], the Philippines maintained its standing as the population who spend the greatest amount of time on social media (i.e. *4 hours average daily time spent on social media*), which provides an abundant source of information regarding the general populace and consumer base.

Big Data for the **private sector** means having a more vivid view of the customers and being able to fine tune and personalize their products and services to maintain and attract business. For instance, Big Data captured through point-of-sale (POS) machines collects real-time information on operations, enabling businesses to employ targeted marketing programs. E-commerce, led by online shopping platforms[6], are now heavily patronized by Filipino consumers; while Financial Technology (FinTech) start-ups capture Big Data on trade and payment statistics, and even offers the services of matching borrowers and lenders.[7]

For the **public sector**, the **Task Force (TF) on Big Data for Official Statistics**, spearheaded by the Philippine Statistics Authority (PSA), has

[2] Hammer, C. et al. (2017). "Big Data: Potential, Challenges, and Statistical Implications," International Monetary Fund Staff Discussion Note SDN/17/06.
[3] Marr tagged this concept of transforming Big Data into useful data as "Smart Data".
[4] Source: Exist Software Labs on turning Big Data into Actionable Intelligence (A Big Data consulting and implementing firm).
[5] Source: We Are Social, retrieved from https://digitalreport.wearesocial.com/
[6] Shoppee and Lazada, among others.
[7] LoanSolutions.ph and LendMe.ph

conducted various initiatives on Big Data, such as workshops, pilot studies, draft of the Philippine Big Data Classification System (PBDCS) and provided comments to the House and Senate Bills on Big Data (*House Bill No. 3056, s.2016 and Senate Bill No. 688, s.2016, "An Act Institutionalizing the Establishment of the Philippine Big Data Center"*). The TF aims to accomplish the compilation of Big Data initiatives of the Philippine Statistics System (PSS) agencies, other government agencies and the private sector, as well as recommend business models on Big Data.

Other notable public sector initiatives on Big Data include the World Bank Group's **OpenRoads Philippines** which aims to provide real-time decision making and priority tagging on infrastructure investments in the country through opensourcing. Moreover, the health sector is utilizing online health-tech platforms[8] to collect real-time health data (e.g., medicine sales, patients' demographics, prescription slips) which are rich sources for Big Data as input to policy-making.

In the **field of research**, among the earlier initiatives on Big Data in the Philippines include **Asian Development Bank's (ADB) Use of Nighttime Lights as Social and Economic Indicators**. The study used satellite images to track the luminosity to indicate economic production and consumption activities (e.g., transportation of people and goods, mass media consumption), which eventually correlates to socio-economic measures (i.e., Gross Domestic Product, poverty, employment and population).

Evidently, public and private sectors must be closely coordinated to reap the full benefits of Big Data. On one hand, the public sector needs to access proprietary information to guide policy-making; on the other hand, the private sector must rely on the public sector's initiatives to facilitate the use of Big Data in its daily operations (i.e., data privacy laws, data governance framework, technological infrastructures, high internet speed, and ease of doing business). And given the projected high demand for data scientists in the near future, colleges and universities in the Philippines are also starting to incorporate Data Science courses in their graduate and post-graduate curriculum to close the competency gaps and fill the skills sets needed by the local economic sectors.

3. Big Data and Central Banking

Similar to the business models of large enterprises, central banks could employ Big Data to drive strategies, streamline operations, abate risks, and effectively deliver their mandates of monetary and financial stability.

[8] Source article retrieved from ttps://www.forbes.com/sites/techonomy/2015/07/30/how-big-data-can-make-people-healthier-in-emerging-markets/2/#4e23c41e2bc0

The primary role of Big Data in central banking operations is to complement official statistics to address gaps in data sources, time lags, and provide more timely insights for prudent policy making. The application of Big Data include gathering sentiments from social media on central bank communications, policy announcements, and consumer and/or business expectations on the economic and financial conditions in the near-term. Leading indicators or early warning indices may also be developed using Big Data through scraping of information from the web, scanner codes, sensors, or even satellite images.[9] Given the emerging interest in exploring the Philippine central bank's potential use of Big Data, surveying cross-country experiences is indispensable to build a suitable framework:

Countries	Big Data and Central Banking
South Korea[10]	Bank of Korea's (BOK) Big Data Research Section started operations in 2017 to explore incorporation of Big Data in generating official statistics (e.g., measuring GDP and consumer-related data though social media platforms and mobile devices)
Indonesia	Bank Indonesia's (BI) Transformation Program[11] uses Big Data to transform information system towards an innovative approach in decision-making. For instance, prior to interest rate announcement, BI scours social media and news sites to gather public perceptions and interest rate expectations to be forwarded to the Governor and the Board.[12]
Singapore[13]	Monetary Authority of Singapore (MAS) uses Big Data to gear its financial sector towards the new digital economy through the establishment of its Data Analytics Group.[14] Case studies include using social media accounts to predict market movements, defaults prediction, and consumer credit risk; and the use of machine learning algorithms in trading patterns to detect money-laundering transactions.
Japan	Bank of Japan (BOJ) uses Big Data in evaluating a series of economic statistics in improving the accuracy of its forecasts

[9] Asian Development Bank study on nighttime light as a development indicator.

[10] Source article retrieved from http://koreajoongangdaily.joins.com/news/article/article.aspx?aid=3036799

[11] Transformation Program of Bank Indonesia will be implemented through 5 phases: (1) Policy excellence, (2) Outstanding execution, (3) Institutional Leadership, (4) Motivated Organization, (5) State of the art technology.

[12] Source article retrieved from http://www.straitstimes.com/business/economy/big-data-helps-bank-indonesia-plug-gaps-in-official-figures

[13] Sources: Monetary Authority of Singapore official website and various news and publications

[14] Source article retrieved from www.techinasia.com/mas-data-analytics-group

	(Matsuo, 2014), for instance nowcasting of service consumption via Google trends.[15]
United States[16]	Fed's Billion Prices Project (BPP) index is an alternate and unconventional measure of retail price inflation though web-scraping techniques and scanner code data.
United Kingdom	Bank of England created a data council for Big Data and uses text analytics on social media posts, news, central bank governor's speeches, and financial contracts among others, to assess monetary and financial stability. Another case is the search data on labor and housing markets through webscraping of key words such as "jobs", "unemployed", "sell house" among others.

In the case of the Philippine central bank, some of the potential use cases of Big Data through its Department of Economic Statistics (DES) are the following:

Current Central Bank Statistics	Potential Big Data Sources
Property Prices: Residential Real Estate Price Index Commercial Property Price Index	Web scraping of online housing prices on sites such as www.lamudi.ph, Property24.com, OLX.ph, Propertyfinder.ph
Household Debt	Online credit card transactions
Services Account: Travel Services	Mobile call data record, road censors, credit card transactions on travel-related services, accommodation-based data
Expectations Survey: Business Expectations Survey Consumer Expectations Survey	Sentiment analysis from social media from Bangko Sentral ng Pilipinas' official social media accounts, or posts with related tags; Text analysis from bank communications (i.e., announcement of policy rate decisions)
Labor Market Statistics: Supply and Demand of Labor Average Wage Levels	Online job search engines such as Jobstreet, JobsDB, PhilJobNet

Other potential applications of Big Data for the Philippine central bank includes nowcasting of macroeconomic indicators, development of sentiment indices from social media posts and text mining analysis, early warning system models through real-time market surveillance to detect financial misconduct and manage risks.

[15] Relates household consumption activities and internet use (i.e., searches related to travel, dining, etc.)

[16] Sources: Federal Reserve Board of Governors/Econresdata

4. The Challenges Ahead for Big Data Application in the Philippine Central Bank

In a Harvard Business Review (HBR) article entitled *"Big Data at Work: Dispelling the Myths, Uncovering the Opportunities",* one of the key recommendations is to adapt the **DELLTA Framework**[17] in incorporating Big Data in business operations. The framework suggests the two (2) phases of Discovery and Production, emphasizing that Big Data initiatives must be unified across the entity through effective leadership. Targets must be well-defined, technological infrastructures must be suited for Big Data, and data scientists must be hired and empowered.

For the Philippine central bank, the extensive benefits on the use of Big Data range from its potential applications as innovative approach for economic and financial analysis to unconventional leading economic and financial indicator that will support its core mandates. However, institutions exploring the possible applications of Big Data will have to face the concomitant challenges that include the following:

1. **External Data Access:** One of the major sources of Big Data are proprietary information from the private sector. Given the central bank's limited clout on gathering data, it needs to enter into bilateral agreements with its identified data sources, or solicit the support of the national statistics agency in accessing the needed information. Some proprietary information are made available at a cost and is fast becoming a highly-valuable asset to the private sector. The cost of accessing Big Data may increase over time and negotiations in establishing public-private partnership agreements is crucial in the near-term.
2. **Data Quality:** Ensuring and verifying the reliability of the statistics/indicators derived and obtained from Big Data is important to minimize, if not totally eliminate, the risks associated with the use of Big Data. From a statistical standpoint, Big Data may **not necessarily cover and represent random samples of the target population**. Hence, thorough examination on the soundness of the methodology and metadata must be undertaken to ensure data quality. Continuity of the data series may also be a concern since Big Data are mostly sourced from the private sector (as a by-product of their daily business operations) and they operate in a continually changing competitive environment. Hence, **statistical comparability of time series** could potentially be affected. Moreover, **outliers and missing information** in the Big Data time series must be clearly detected in order to be statistically resolved with imputations or sound estimates.

[17] DELLTA Framework: Data, Enterprise, Leadership, Targets, Technology, and Analytics

The members of the Philippine Statistical System (PSS) must remain closely coordinated in researching and compiling best practices in the statistical techniques and methodologies to address Big Data's inherent veracity and volatility issues.

3. **Data Governance:** The Philippine central bank must be able to establish a single Data Governance framework which will serve as guidelines to direct work flow, processes and data management with regard to how data should be accessed, used, and protected. Without a single framework, incoherent practices on Big Data management may pose risks to data quality and security. The path towards a wider and more inclusive data-sharing and collaboration in accessing information-rich data sets must be clearly linked with data security. The extent and coverage of the legalities involved in Big Data is not limited to consumer rights and protection, intellectual property rights, copyrights, and licensing arrangements with partner providers. Liabilities for Big Data breaches must be discussed and clearly defined.
4. **Data Confidentiality:** Big Data usually consist of highly personal and private information that the BSP must safeguard against cyber security threats and confidentiality risks. Although the Philippine central bank adheres to the guidelines set by the National Privacy Commission (NPC), specifically the Data Privacy Act of 2012 and its Implementing Rules and Regulations (IRR), the BSP must still enhance its data security layers (i.e. cryptography, user access) to mitigate reputational risks and loss of trust from its key data providers.
5. **Cyber Security:** Given the high volume and high velocity nature of Big Data, cyber security measures must be stringent and reliable. Preventing cyber-attacks must be real time, swift, efficient, and effective. Threats, therefore, must be countered proactively rather than reactively given the wide variety of data sources and the data's high level of sensitivity.
6. **Capital Resources:** Organizational readiness in terms of BSP's capital resources is vital in keeping up with Big Data developments and opportunities. Human and technological resources must be compatible and/or up to speed with the demands of Big Data to ensure operational efficiency and effectiveness in data collection, management, analysis, and timely dissemination to BSP management for policy-making.
    a. **Technological Capacity:** One of the pressing cost implications of Big Data is the expenditure related to the acquisition, installation, and maintenance of digital infrastructures (e.g., hardware and software compatible with Big Data). The BSP's current IT infrastructure may need to be upgraded to keep pace with the developments in collecting, processing, storing, and managing Big Data.

b. Human Capacity[18]: Investing in human capital, through trainings and scholarships, is vital in harnessing the benefits of the latest technological advancements in Big Data. The BSP has always been an advocate of pursuing further learning through courses and postgraduate studies[19] by means of scholarships and fellowships. Promoting data-driven courses and programs to the BSP staff will significantly narrow the gap in competencies and technical expertise needed in analyzing Big Data. The Bank may hire experienced data scientists to jumpstart pilot projects and who will provide guidance and train the BSP's existing staff. It may also consider harnessing talents from the undergraduate and graduate levels in key universities who are now offering courses and graduate studies on Big Data.

Big Data is a multi-dimensional discipline that requires pooling of talents or experts in the field of computational science, mathematics, statistics, programming, and data management, among others. Given these features, setting up a unit or division within the Bank composed of human resources with expertise in the areas enumerated above is optimal. In building Big Data skills and capabilities, implications on organizational structure and/or human resource assignment must be rationalized, and the details on the number of positions needed, qualifications, and competencies for employee selection, as well as the duties and responsibilities of the organizational unit, must be clearly identified.

Indeed, Big Data is replete with policy applications which warrants strong management support in overcoming its concomitant challenges. For the Philippine central bank, Big Data may supplement its official statistics, provide new insights for policies, improve market surveillance, and provide an innovative source for research for a more effective and efficient delivery of its core mandates – monetary and financial stability that delivers a high quality of life for all Filipinos.

[18] Staff involved in Big Data must be able to identify Big Data opportunities for the BSP, gather and clean data, apply statistical methodologies, communicate insights to management through expertise in programming, domain knowledge, and project management. Skills on database management, data modelling, developing algorithms, and data visualization are also required.

[19] Currently, the BSP Institute (BSPI) is offering courses and scholarships for graduate studies on Data Science.

References

1. ASEAN UP. 2017. FinTech startups in the Philippines. https://aseanup.com/fintech-startups-philippines/ (accessed 28 December 2017).
2. Asian Development Bank (ADB). 2016. Key indicators for Asia and the Pacific 2016. Mandaluyong City, Philippines. ADB.
3. Barrios, E.B. 2017. Big data, official statistics, analytics: Philippine context. Presentation at the *Big Data Analytics: Applications in Policy, Development, and Governance*, 26 July 2017, Asian Institute of Management (AIM).
4. De Costo, S.B. 2017. Philippine Task Force on Big Data. Presentation during the United Nations Expert Group Meeting on International Statistical Classifications on 6-8 September 2017 in New York City, United States. ESA
5. Marr, B. (2015). *Big Data: Using Smart Big Data Analytics and Metrics to Make Better Decisions and Improve Performance.* United Kingdom: John Wiley & Sons Ltd.
6. Matsuo, Y. (2014, March 7). Retrieved from Asia Nikkei: http://asia.nikkei.com/Politics-Economy/Economy/Big-data-makes-big-changes-in-how-Japan-s-central-bank-looks-ahead

Calibration for the Census of Agriculture in Albania

Alban Çela[1], Klajd Shuka[2]
[1] Agriculture and Environment Statistics Director at Institute of Statistics Albania
[2] Finance and Supporting Services Director at Institute of Statistics Albania

Abstract
Agriculture is one of the main economic activities, impacting gross domestic product of the country, contributing with around 20 percent. Accurate agriculture indicators are crucial for the European integration of Albania. The Institute of Statistics in Albania is conducting every 10 years the Census of Agriculture. This census is one large scale statistical activity for the collection, processing and publication of data on the structure of agricultural holdings. The 2012 census of agriculture data comparative analysis showed deviations from the historic administrative data. This paper presents calibration as an alternative method to estimate agriculture figures, based on census of agriculture data and administrative data. Fisher Ideal Index is used for calculation of trend components. This method adjusted the figures for different type of bias such as under-coverage and non-response. As a result of this analysis calibration process to improve the quality and accuracy of agriculture data that will be used for the creation of the statistical Farm Register.

Keywords
Estimation; Calibration; Bias; Administrative data

1. Introduction

INSTAT is the leader of the National Statistical System and main producer of official statistics in Albania. During the implementation of official statistics programmes many surveys and censuses have been conducted. Usually censuses in Albania are conducted every ten years. For censuses usually the indicators are produced based on the collected data taking into consideration the effects of under-coverage. On the other hand, for surveys estimations are made taking into consideration under-coverage and different types of non-response. To reduce the negative impact of the previously mentioned problems the estimates are adjusted during weights calculation and calibration phase. For the Census of Agriculture (CA) of 2012 for the first time INSTAT has been calibrating census data, due to differences with other administrative historical data. High quality and accurate data from CA are the basis for the creation of the statistical Farm Register and have a big impact on the quality of indicators produced using future surveys on agriculture.

2. Census of Agriculture

The last Census of Agriculture in Albania was conducted in October 2012, with the objective to collect process and publish data on the structure of Agricultural Economic Units (farms) operating in Albania. The CA is an integral part of the agricultural statistics system, the methodology of which complies with the international standards of the Food and Agriculture Organization (FAO) and with EU legislation: Regulation (EC) 1166/2008 of the Parliament and of the European Council on farm structure surveys and agricultural production methods. Agriculture Censuses in Albania usually have a 10 years frequency. The main purpose of the CA was to: Provide up-to-date and accurate statistical information on all entities that carry out agriculture activities in the territory of Albania; Emphasize the main structural features of agricultural and livestock units such as the management system, organizational and legal forms, land surface and soil cultivation methods, cultivated crops and number of livestock heads, used mechanical tools, work forces and other aspects; Establish the Statistical Farm Register as a basis for carrying out statistical surveys on Agricultural Economies.

3. Problems faced

The 2012 CA results were delayed as to ensure that totals were in line with administrative data available from Ministry of Agriculture, Rural Development and Water Administration (MARDWA). The administrative data available were not fully compliant with requirements and standards of the EU and had to be reviewed and revised accordingly. Comparative analysis showed that the outcomes of the CA of 2012 were deviating from the historic data from the Ministry. The main differences were due to the process of creation of the frame for CA, as a result of the lack of existence of a unique identifier for agriculture units in Albania. Preparation of frame has been based on information from 2011 Population and Housing Census; farm animal identification campaign conducted in 2011 by the Food and Veterinary Safety Institute; National Registration Centre; the list of public and religious institutions that own or have access to agricultural land or exercise livestock activities. However, apart from the existing doubts about the quality of the sampling frame to be used, there was also the practical issue which choice has to be made about the organization of the data collection in the field. Negotiations have been initiated with the Ministry intensifying the cooperative arrangements and assigning regional staff for the statistical data collection, because of their experience in the sector and their knowledge of the local circumstances in the field.

A Memorandum of Understanding[1] in the field of statistics between INSTAT and Ministry of Agriculture was signed in March 2016. On the basis of this agreement, INSTAT and MARDWA secure the data exchange produced within the scope of respective institutional statistical activities. To resolve differences between statistical figures and administrative data INSTAT and MARDWA proposed conducting two surveys: Calibration of CA (Village Survey) and Annual Agriculture Survey (AAS).

AAS was a sample survey, whose sample frame was prepared based on the CA database and updated with information from administrative sources for the large farms. The questionnaire used was the one prepared for the usual AAS with small adjustments. The survey provided annual data on crop and livestock production as well as some other data needed for calculation of estimated agriculture area, since the sampling frame was prepared based on CA with potential under coverage, the data needed to be calibrated with village survey data.

Village survey was based on the sample of enumeration areas selected from population and housing census of 2011 database. The sampling frame was prepared taking into consideration the following information for enumeration areas: Number of households; Share of households with agriculture activity; Utilised agricultural area - UAA (in square meters). The questionnaire included filter questions on agricultural activity, question on land use and livestock number as in the CA. The survey was used to calibrate CA data and also to calibrate the AAS.

4. Census calibration

The calibration of the CA was based on two surveys:

- Annual Agricultural Survey (AAS) and
- Agricultural Census calibration (ACC)– Village Survey

The strata taken into account for the calibration process were at level of the prefecture and household/farm size. In Albania there are 12 prefectures and four strata of the farm size. Stratification by size is shown in table 1.

[1] http://instat.gov.al/en/about-us/activities/other/instat-and-ministry-of-agriculture-have-signed-the-memorandum-of-understanding-in-the-field-of-statistics/

Table 1: Stratification by size of the household/farm based on 14 auxiliary variables

Sampling Design				
Variables	Size 1	Size 2	Size 3	Size 4
UAA	if B2_09_a>80000	else if B2_09_a>9200	else if B2_09_a>2000	else if B2_09_a<=2000
Cows	if D1_06_2>20	else if D1_06_2>7	else if D1_06_2>1	else if D1_06_2<=1
Sheeps	if D2_11>150	else if D2_11>40	else if D2_11>20	else if D2_11<=20
Goats	if D3_15>150	else if D3_15>40	else if D3_15>14	else if D3_15<=14
Saws	if D4_17>9	else if D4_17>5	else if D4_17>1	else if D4_17<=1
Green House	if GH>=4000	else if GH>2500	else if GH>1500	else if GH<=1500
wheat	if C1_10>=25000	else if C1_10>10000	else if C1_10>2000	else if C1_10<=2000
Root	if root>=2200	else if root>1000	else if root>300	else if root<=300
Indust	if indust>=7500	else if indust>5000	else if indust>1000	else if indust<=1000
Dried	if dried>=9000	else if dried>5000	else if dried>300	else if dried<=300
Harvest	if harvest>=50000	else if harvest>6000	else if harvest>2000	else if harvest<=2000
Arable Land	if arable_land>=70000	else if arable_land>10000	else if arable_land>2000	else if arable_land<=2000
Citrus	if Citrus>=10000	else if Citrus>2000	else if Citrus>700	else if Citrus<=700
Poultry	if Poultry>400	else if Poultry>100	else if Poultry>50	else if Poultry<=50

The main idea for calibration is to use ACC data set as a source of the household/farm size strata 2, 3 and 4 and form the survey ASS to use an existing part, household/farm size strata 1. The ACC sampling was based on a population census, and the ASS was based on the agricultural census plus data from the MARDWA which is collecting from the big farm on monthly bases. Analysing coefficients of variances and standard error of the auxiliary variables was chosen the most suitable methods for the variable Arable Land and Number of Farms.

Marginal tables were corrected for ratios between coefficients for the two most important auxiliary variables.

Table 2: Ratios between marginal coefficients between Number of Farms and Arable Land by Prefecture

Prefecture	**Marginal Coefficients Ratio Number of Farms / Arable Land**
Total	**0.9969**
Berat	0.9759
Dibër	0.9826
Durrës	0.9729
Ebasan	1.0542
Fier	1.1041
Gjirokastër	0.6118
Korçë	0.7569
Kukës	0.8031
Lezhë	0.9239
Shkodër	0.7188
Tiranë	1.7382
Vlorë	1.4266

Table 3: Ratios between marginal coefficients between Number of Farms and Arable Land by stratum Size of farms

Stratum Size	1	2	3	4	Total
Marginal Coefficients Ratio Number of Farms / Arable Land	0.9143	1.1222	1.1496	1.4495	0.9969

For the calculation of the trend component was used the Fisher Ideal index[2]. The Fisher Ideal index is a measure of change in volume from period to period. It is calculated as the geometric mean of a chain Paasche volume index and a chain Laspeyres volume index. Using information from Arable land and Number of Farms between census and calibrated database we used the Fisher Ideal index formula as follows:

$$F = \sqrt{LP}$$

based on Laspeyres Index that includes Arable Land

$$\text{Laspeyres Index} = 100 \text{ x } \frac{\sum PnQo}{\sum PoQo}$$

and Paasche Index that includes No of farms

$$\text{Paasche Index} = 100 \text{ x } \frac{\sum PnQn}{\sum PoQn}$$

In our situation L=1.0319 and P=1.0309 and calculated F=1.0314. Combination of two main parameters showed a yearly change of +3.14%. Final results are given in Table 3 after excluding cumulative trend association.

Table 3. Final weight adjusted coefficients for final calibration of the Agricultural Census 2012.

Prefecture/Stratum Size	1	2	3	4
Berat	2.1644	1.0388	1.0813	0.8727
Dibër	2.0375	0.9779	1.0178	0.8215
Durrës	2.7294	1.3099	1.3635	1.1005
Ebasan	1.9931	0.9566	0.9957	0.8036
Fier	2.5810	1.2387	1.2893	1.0407
Gjirokastër	2.2117	1.0615	1.1049	0.8918
Korçë	1.5297	0.7341	0.7642	0.7274
Kukës	1.8040	0.8658	0.9012	0.7274
Lezhë	2.8060	1.3467	1.4017	1.1314
Shkodër	1.6606	0.7970	0.8296	0.6696
Tiranë	2.6064	1.2509	1.3020	1.0509
Vlorë	2.1905	1.0513	1.0943	0.8832

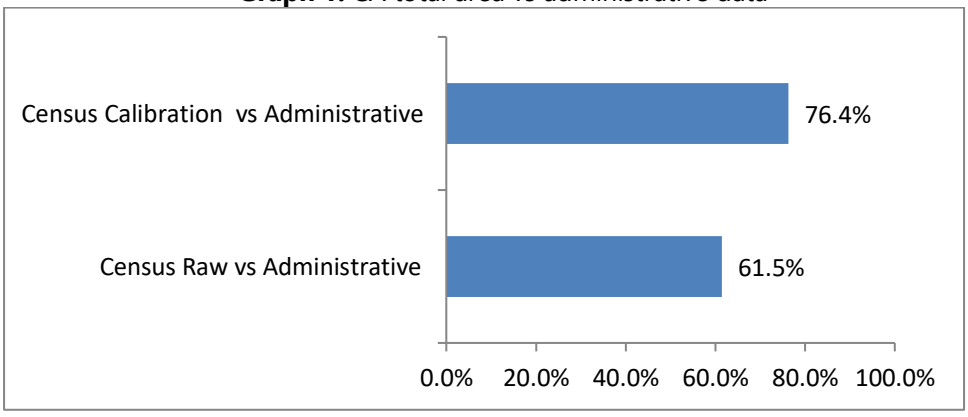
[2] https://stats.oecd.org/glossary/detail.asp?ID=989

5. Improvement of quality in the CA estimation

Calibration estimation, whereby sampling weights are adjusted to reproduce known population totals, is commonly used in survey sampling. As described in the article by Deville and Särndal (1992), calibration can be treated as an important methodological instrument, especially in large-scale production of statistics.

In Graph 1 is presented the difference between Census Raw data versus Administrative data provided by MARDWA, as well as the difference between Census Calibrated data versus Administrative ones for the total area. The total area is defined as the cumulative sum of Kitchen garden; Arable Land; Permanent grassland; Permanent crops; Unutilized agricultural area; Wooded area; Non-agricultural area.

Graph 1: CA total area vs administrative data



Results have been improved approximately 14.8 percent from the calibration process. These results will directly impact positively the sampling frame for future agriculture surveys.

6. Conclusions

Taking into consideration problems faced for the CA of 2012 as well as the actions taken to improve quality of the indicators produced the following conclusions are reached:

- Calibration uses for CA improved the CA estimation and they are more in line with the historic data from the Ministry.
- In the preparatory phase of CA, the alternative methods for the logical controls, extreme value analysis, post data collection checking of the farms, post stratification needs to be evaluated, to improve the quality of the estimations.
- The frame for the next Census of Agricultural needs to be based on an Administrative data for the agricultural enterprises/big farms and Population Census for household/farms

- In order to evaluate better the problem of under coverage and different type of non-response in Albania it is needed to established the Address system
- Unique identifiers for farms needs to be established

References

1. http://instat.gov.al/en/about-us/activities/other/instat-and-ministry-of-agriculture-have-signed-the-memorandum-of-understanding-in-the-field-of-statistics/
2. https://stats.oecd.org/glossary/detail.asp?ID=989
3. Calibration Estimators in Survey Sampling. Deville and Särndal. JASA, 1992, 376-382
4. Law on Official Statistics No 17, dated 05/04/2018 http://instat.gov.al/media/3972/law-no17-2018-on-official-statistics.pdf

Relationship of inflation with imports and exports in Malaysia

Wan Siti Zaleha Wan Zakaria, Siti Nuraini Rusli, Nur Amirah Daud
Department of Statistics, Malaysia

Abstract

Inflation refers to a common raise in prices which lead to drop in the purchasing value of money. The most commonly used indicators as a proxy to inflation are the Consumer Price Index (CPI), Wholesale Price Index (WPI) and the Gross Domestic Product (GDP) deflator. This paper aims to examine the relationship between CPI and imports and exports of goods in Malaysia. The study is motivated by demand-pull and cost-push theory of inflation using monthly time series data from the Department of Statistics, Malaysia (DOSM) for the 336 months between January 1990 and December 2017. This study applies unit root test to check the stationarity of variables of time series; co-integration test to examine possible correlations among variables in the long term; and the causality test to check causality between the pair of variables in a time series. By employing co-integration technique, it is observed that the long-run relationship does not exist between CPI & imports and CPI & exports. However, causality analysis suggests the existence of short-run relationship between CPI & imports and CPI & exports. The CPI had caused the increase in imports and the exports had caused the increase in CPI.

Keywords

Machine Learning; Panel Survey; Nonresponse; Feature Selection; Ensemble Methods

1. Introduction

Inflation refers to a general increases in prices which lead to reduction in the purchasing power. Inflation can be measured through CPI or WPI or GDP deflator. Inflation measures the increase in the cost of living in a country as when prices go up, monetary value declines and lead consumers to spend less on goods and services. Inflation mainly due to either demand or supply or both factors. Demand side factors result in demand-pull inflation and supply side factors lead to cost-push inflation.

Demand-pull theory implies that the inflation happens when aggregate demand for goods and services increases so much leading to increased pressure on limited resources. When demand surfeits supply, prices of goods and services will go up and create inflation. Among the reasons of demand-pull inflation are depreciation of the exchange rate; higher demand from a

fiscal stimulus; monetary stimulus to the economy; and fast growth in other countries. Cost-push theory indicates that the inflation occurs when the producers counter the increasing in costs by raising prices to save guard their profit margins. Among the factors of cost-push inflation are raising labour costs, expectation of inflation, higher indirect taxes, a fall in the exchange rates and monopoly employers/profit push inflation.

As Malaysia is an open economy, the exports and imports play important roles to the inflation. When demand exceeds the domestically produced goods and services, the gap becomes larger. This will result in the inflationary situation. In order to fulfil the demand, the country may import and ease the inflation. On the contrary, the country may export when the domestic supply of goods and services surpasses the demand and avoid the deflation. A depreciation of the exchange rate raises the price of imports and cuts the foreign price of exports. External trade also can cause inflation by the competition of local production as compared to imported items.

2. Methodology

2.1. The data

To carry out the study on the relationship of inflation with imports and exports of goods in Malaysia, the variables involved are the Malaysia's CPI to measure inflation; and Malaysia's merchandise imports & exports. The study uses the monthly data from January 1990 to December 2017 which were obtained from the DOSM.

2.2. Unit root test

The variables in the regression model have to be stationary in order to prove that the standard assumptions for asymptotic analysis are valid. To investigate the stationarity of the data, a univariate analysis of each of the time series was carried out by testing for the presence of a unit root. This study used Augmented Dickey-Fuller (ADF) test as stationary test. If the series (level) are non-stationary, the data should first be transformed into stationary data (using first (or higher) differences) so that further statistical analysis can be applied.

2.3. Co-integration test

The co-integration test identifies the existence of long-run relationship between the variables under study. This study used both Johansen's Maximum Eigenvalue test and the trace test for investigating co-integration of the time series. If the test indicates the absence of co-integration relation, the model VAR will be used. If otherwise, the model VECM will be used.

2.4. Causality test

The causality test investigates link between pairs of variables where it determines whether one time series is useful in forecasting another time

series. This study used Granger causality test to examine the existence of short-run relationship between the pairs of variables under study.

3. Results

a. Unit Root Test

This stage involves establishing integration order by using ADF test at constant without trend and constant with trend. Table 3.1 represents the results of unit root test for CPI, imports and exports.

Table 3.1: Results of Unit Root Test

Tests	Augmented Dickey Fuller	
Models	Constant without Trend	Constant with Trend
Level		
CPIs	-1.9484(5)	-2.8268(1)
Imports	-1.6411(12)	-2.5821(12)
Exports	-2.1877(12)	-1.7335(12)
First Difference		
CPI	-13.9582***(0)	-9.4433***(0)
Imports	-4.4121***(12)	-4.4810***(12)
Exports	-4.3969***(12)	-4.6709***(12)

Notes: *** (**) and * denotes significant at 1%, (5%) and 10% significant level, respectively.

3.2. Co-integration Test

This test will determine the existence of long-run relationship between CPI & imports and CPI & exports. The co-integration result between CPI & imports is in Table 3.2.1 and CPI & exports is in Table 3.2.2.

Referring to Table 3.2.1, Trace test shows that there is co-integration between CPI & imports. Meanwhile, Maximum Eigenvalue test indicates no co-integration between CPI & imports. The Trace and Maximum Eigenvalue test yield different results, in this case the results of Maximum Eigenvalue test should be preferred (Banerjee et, al., 1993). Based on that, there is no co-integration between CPI & imports.

Table 3.2.1: Cointegration Test Between CPI & Imports

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
Results of Unrestricted Cointegration Rank Test (Trace)				
None *	0.0402	15.9375	15.4947	0.0429
At most 1	0.0083	2.6974	3.8415	0.1005
Trace test indicates 1 cointegrating eqn(s) at the 0.05 level * denotes rejection of the hypothesis at the 0.05 level **MacKinnon-Haug-Michelis (1999) p-values				

Results of Unrestricted Cointegration Rank Test (Maximum Eigenvalue)				
None	0.0401	13.2401	14.2646	0.0721
At most 1	0.0083	2.6974	3.8415	0.1005
Max-eigenvalue test indicates no cointegration at the 0.05 level * denotes rejection of the hypothesis at the 0.05 level **MacKinnon-Haug-Michelis (1999) p-values				

According to Table 3.2.2, Trace test and Maximum Eigenvalue test indicates no co-integration between CPI & exports. Thus, this study find that there is no long run relationship between CPI & imports and CPI & exports.

Table 3.2.2: Cointegration Test Between CPI & Exports

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
Results of Unrestricted Cointegration Rank Test (Trace)				
None *	0.0359	11.8238	15.4947	0.1656
At most 1	5.55E-05	0.0180	3.8415	0.8934
Trace test indicates no cointegration at the 0.05 level * denotes rejection of the hypothesis at the 0.05 level **MacKinnon-Haug-Michelis (1999) p-values				
Results of Unrestricted Cointegration Rank Test (Maximum Eigenvalue)				
None *	0.0359	11.8059	14.2646	0.1181
At most 1	5.55E-05	0.0179	3.8415	0.8934
Max-eigenvalue test indicates no cointegration at the 0.05 level * denotes rejection of the hypothesis at the 0.05 level **MacKinnon-Haug-Michelis (1999) p-values				

3.3. Causality Test under the Multivariate Vector Autoregressive Model (VAR) Framework

The co-integration test defined the non-existence of the long-run relationship between CPI, imports & exports. In order to define the direction of Granger causality among the variables, a regression model with Vector Autoregressive Model (VAR) should be established. The estimation of VAR requires selection of a suitable lag length. The number of lags in the model is determined accordingly based on the smallest Akaike Information Criterion (AIC). The smallest AIC for this study is lag 12 for CPI & imports and CPI & exports. The VAR equations are as follows:

$$\begin{aligned}\Delta CPI = {} & 0.0141 + 1.2375CPI_{t-1} - 0.2682CPI_{t-2} + 0.0011CPI_{t-3} \\ & - 0.0151CPI_{t-4} - 0.1072CPI_{t-5} + 0.1060CPI_{t-6} \\ & + 0.1439CPI_{t-7} - 0.1645CPI_{t-8} + 0.0013CPI_{t-9} \\ & + 0.10406CPI_{t-10} - 0.0116CPI_{t-11} + 0.0306CPI_{t-12} \\ & + 0.0069Imports_{t-1} - 0.0004Imports_{t-2} - 0.0035Imports_{t-3} \\ & - 0.0033Imports_{t-4} - 0.0002Imports_{t-5} + 0.0039Imports_{t-6} \\ & + 0.0007Imports_{t-7} + 0.0041Imports_{t-8} - 0.0041Imports_{t-9} \\ & + 0.0030Imports_{t-10} - 0.0040Imports_{t-11} \\ & - 0.0016Imports_{t-12}\end{aligned}$$

$$\begin{aligned}\Delta CPI = {} & 0.0141 + 1.2375CPI_{t-1} - 0.2682CPI_{t-2} + 0.0011CPI_{t-3} \\ & - 0.0151CPI_{t-4} - 0.1072CPI_{t-5} + 0.1060CPI_{t-6} \\ & + 0.1439CPI_{t-7} - 0.1645CPI_{t-8} + 0.0013CPI_{t-9} \\ & + 0.10406CPI_{t-10} - 0.0116CPI_{t-11} + 0.0306CPI_{t-12} \\ & + 0.0069Imports_{t-1} - 0.0004Imports_{t-2} - 0.0035Imports_{t-3} \\ & - 0.0033Imports_{t-4} - 0.0002Imports_{t-5} + 0.0039Imports_{t-6} \\ & + 0.0007Imports_{t-7} + 0.0041Imports_{t-8} - 0.0041Imports_{t-9} \\ & + 0.0030Imports_{t-10} - 0.0040Imports_{t-11} \\ & - 0.0016Imports_{t-12}\end{aligned}$$

$$\begin{aligned}\Delta CPI = {} & 0.0043 + 1.2096CPI_{t-1} - 0.2386CPI_{t-2} - 0.0199CPI_{t-3} \\ & - 0.0126CPI_{t-4} - 0.0985CPI_{t-5} + 0.0862CPI_{t-6} \\ & + 0.1791CPI_{t-7} - 0.1767CPI_{t-8} + 0.0049CPI_{t-9} \\ & + 0.0516CPI_{t-10} - 0.0136CPI_{t-11} + 0.0302CPI_{t-12} \\ & + 0.0090Exports_{t-1} + 0.0013Exports_{t-2} - 0.0054Exports_{t-3} \\ & - 0.0048Exports_{t-4} + 0.0037Exports_{t-5} + 0.0011Exports_{t-6} \\ & + 0.0033Exports_{t-7} - 0.0006Exports_{t-8} - 0.0050Exports_{t-9} \\ & + 0.0009Exports_{t-10} - 0.0016Exports_{t-11} \\ & - 0.0029Exports_{t-12}\end{aligned}$$

$$\begin{aligned}\Delta Exports = {} & 0.1552 + 2.2320CPI_{t-1} - 2.1922CPI_{t-2} + 1.1066CPI_{t-3} \\ & - 0.3156CPI_{t-4} - 0.4721CPI_{t-5} + 0.3130CPI_{t-6} \\ & - 3.3625CPI_{t-7} + 2.0842CPI_{t-8} + 1.0225CPI_{t-9} \\ & + 0.3774CPI_{t-10} - 1.9642CPI_{t-11} + 1.2175CPI_{t-12} \\ & + 0.03713Exports_{-1} + 0.3129Exports_{t-2} + 0.1869Exports_{t-3} \\ & - 0.0919Exports_{t-4} - 0.0216Exports_{t-5} - 0.0751Exports_{t-6} \\ & + 0.0388Exports_{t-7} + 0.0247Exports_{t-8} + 0.01540Exports_{t-9} \\ & - 0.1541Exports_{t-10} - 0.1386Exports_{t-11} \\ & + 0.3627Exports_{t-12}\end{aligned}$$

Next, the short-run relationship between CPI & imports and CPI & exports is determined using Granger-Causality test at 5 per cent significant level. The results are presented in Table 3.3.1.

Table 3.3.1: Results of Granger Causality Tests

Null Hypothesis	F-Statistic	Prob.	Decision	Causality
Imports does not Granger Cause CPI	17.0241	0.1487	Do Not Reject H_0	No

CPI does not Granger Cause Imports	25.6096	0.0122	Reject H_0	Yes
Exports does not Granger Cause CPI	23.7083	0.0223	Reject H_0	Yes
CPI does not Granger Cause Exports	17.3147	0.1381	Do Not Reject H_0	No

Following the results in Table 3.3.1, the null hypothesis for CPI does not granger cause imports (p-value at 0.0122) and exports does not granger cause CPI (p-value at 0.0223) is rejected at 5 per cent significance level. However, the the null hypothesis for imports does not granger cause CPI and CPI does not granger cause exports failed to be rejected at 5 per cent significant level. Thus, conforming a uni-directional causality running from CPI to imports, supporting the results from Dewan Muktadir-Al-Mukit (2013). Besides that, a uni-directional causality also run from exports to CPI.

4. Discussion and Conclusion

The main purpose of this study was to discover whether imports and exports have an impact on inflation in Malaysia The study employed the following econometrics techniques which are the unit root test for testing the stationarity in variables; the co-integration test to examine for the long-run relationship between variables; and causality test to identify short-run relationship between the pairs of variables. These techniques were employed on monthly Malaysia's CPI and imports and exports of goods data for the period between January 1990 and December 2017. Based on the findings, it can be concluded that the long-run relationship does not exist between CPI & imports and CPI & exports. However, there is short term linkage relationship between CPI & imports and CPI & exports. Based on the Granger causality test, the CPI had caused the increase in imports and the exports had caused the increase in CPI.

Result from Granger causality analysis indicates the existence of unidirectional causality running from CPI to imports which points out that CPI lead imports condition for the Malaysia economy. Based on the cost-push theory of inflation, the result suggests that Malaysia's inflation more attributable to domestic supply which subsequently, influences imports. On the other hand, the result also shows the existence of unidirectional causality running from CPI to exports which points out that exports lead CPI condition for the Malaysia economy. Based on the demand-pull theory of inflation, Malaysia's inflation may due to the need of fulfilling the demand externally.

The domestic resources need to be utilized at optimal to fulfil the domestic demand in order to ease the inflation. Goods that satisfy domestic needs which can be produced more inexpensively or efficiently by other countries should to be identified and imported to lower the price. Concurrently, the domestic production should complete the domestic demand before export. Policymakers should also concentrate on other factors such as money supply, interest rate, national expenditure and exchange rate which are also contributors to inflation.

References

1. Ahmed, Rizwan Raheem; Ghauri, Saghir Pervaiz; Streimikiene, Dalia and Vveinhardt, Jolita (2018). An Empirical Analysis of Export, Import and Inflation: A Case of Pakistan. Romanian Journal of Economic Forecasting – XXI (3), 117-196
2. Bank Negara Malaysia (BNM) (2011). Determinants of Inflation in Malaysia. BNM Annual Report 2010. Kuala Lumpur: Percetakan Nasional Berhad, 50-53.
3. Department of Statistics Malaysia. Malaysia Consumer Price Index. Putrajaya.
4. Department of Statistics Malaysia. Malaysia External Trade. Putrajaya.
5. Islam, Rabiul; Abdul Ghani, Ahmad Bashawir; Mahyudin, Emil; and Manickam,Narmatha (2017). Determinants of Factors that Affecting Inflation in Malaysia. International Journal of Economics and Financial Issues, Vo. 7(2), 355-364.
6. Lim, Yen Chee and Sek, Siok Kun (2015). An Examination on the Determinants of Inflation. Journal of Economics, Business and Management, Vol. 3, No. 7, 678-682
7. Mukit, Dewan and Shafiullah, Abu Zar (2013). Inflation Led Import or Import Led Inflation: Evidence from Bangladesh. Asian Business Review, Volume 2, Number 2/2013 (Issue 4), 5.
8. Mukit, Dewan and Shafiullah, Abu Zar (2014). Export, Import and Inflation: A Study on Bangladesh. Amity Global Business Review Vol. 9, 46-55.

The importance of SUA as part of cost of living

Siti Zakiah Muhamad Isa, Manisah Othman, Nur Khairunniza Harun
Department of Statistics Malaysia

Abstract

The Supply and Utilization Accounts (SUA) for selected agriculture commodities is a balance account comprising the elements of supply and utilization. Supply accounts include opening stock, production and imports while utilization accounts include exports, seeds, feeds, waste, processing, closing stock and food. The compilation of SUA is used to calculate three agricultural indicators of self-sufficiency ratio which explains the extent to which the production of agricultural commodities for a country is sufficient to meet domestic needs. While, the second indicator is the imports dependency ratio which explains a country's dependence on imports of agricultural commodities to meet domestic needs. Imports dependency ratio shows the higher rates the more supply of agricultural commodities to be imported. The third indicator is per capita consumption which refers to the amount of food consumed by each person per year and it is measured in kilograms. These three indicators can be used to analyze the trend of supply and utilization of agricultural commodities.

Keywords

Supply accounts; utilization accounts; self-sufficiency ratio; domestic needs; per capita consumption

1. Introduction

The Department of Statistics, Malaysia (DOSM) has published The Supply and Utilization Accounts (SUA) since 2010. This coverage was expanded starting with nine agricultural commodities (2010) to 33 as published in 2018 (Exhibit 1). Agricultural commodities selected in the compilation of SUA are commodities listed in the National Agro-Food Policy (NAP) 2011-2020 and it focused on food that have direct impact on consumers. However, the coverage of the commodity depends on the availability and stability of the data for production, import and export. Primary data sources are secondary data obtained from international trade data and survey conducted by DOSM and also include production data from agricultural agencies.

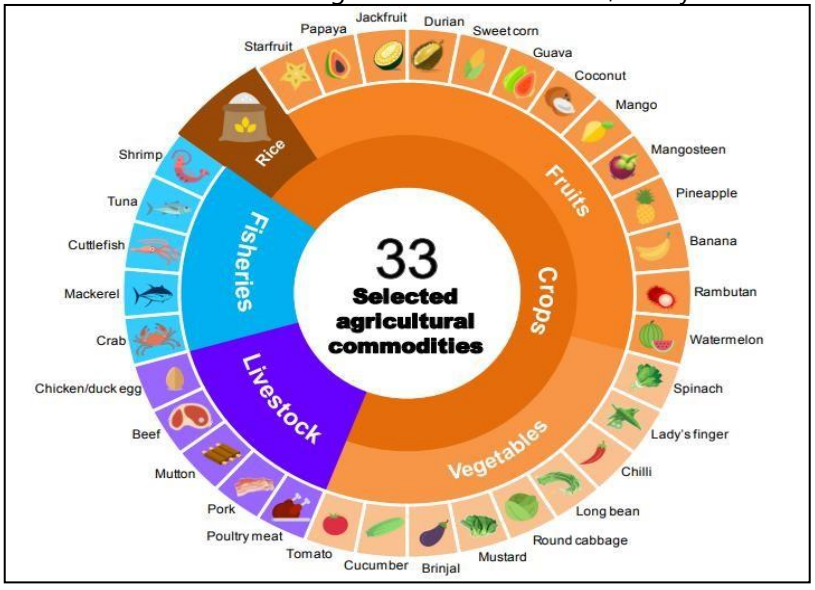
SUA is a balancing account that contains elements of supply and utilization. This paper focuses on agriculture-related statistics and indicators which is an important indicator to the national agriculture situation,

particularly in determining sufficient food supply security. Indicators generated through SUA could assist not only researchers or government agencies but also policy makers in assessing the current situation and planning the development of the agricultural sector in order to advance the nation and the well-being of the people.

NAP 2011-2020 has provided guidelines for the agro-food industry in contributing to the country's economic growth. Agro-food industry plays a role in meeting the growing needs of food and manufacturing industries. Major challenges in this industry include rising cost of production as a result of rising agricultural input prices, the need to comply with international standards, lack of labour force and the use of less extensive technology. However, the agro-food industry also has the potential and wide range of opportunities based on population growth.

Agro-food productions not only improve the country's economy, but it is also a source of our daily food. Consumption of a variety of fruits and vegetables in daily diet can help meet the vitamins and minerals needs as well as to prevent disease and increase body resistance. Based on the Malaysia Food Pyramid, fruits and vegetables are categorized into food groups which need a lot to eat. Every day, the number of fruits and vegetables should be taken according to the prescribed recommendations. In addition, we are also advised to take fish, chicken, ducks, meat and legumes because the food is rich in protein. Per capita consumption for this item will be further elaborated. The latest publication of SUA is for the series of 2013 to 2017.

Exhibit 1: Selected agricultural commodities, Malaysia



2. Methodology

SUA is a balancing account that contains elements of supply and utilization. Elements of supply include opening stock, production and imports. While, elements of utilization include exports, seed, feed, waste, processing, closing stock and food. Three indicators can be generated i.e. self-sufficiency ratio (SSR), imports dependency ratio (IDR) and per capita consumption (PCC) through which able to analyze the trend of supply and utilization of agricultural commodities thus evaluate the country's agricultural performance especially in determining sufficient food supply security.

Self-sufficiency ratio (SSR) can explain the extent to which the supply of agricultural commodities can meet domestic demand. SSR that reach 100 per cent or more indicates that production is sufficient to meet domestic needs. While, the import dependency ratio (IDR) indicate the degree of dependence of a country on the import of agricultural commodities to meet domestic needs. The higher the IDR, the more supply of agricultural commodities to be imported. Meanwhile, for per capita consumption (PCC) refers to the amount of food consumption by each person per year. PCC is measured in kilograms a year.

3. Results

Summary of findings is described in three categories: crops, livestock and fisheries. Referring to the 2013-2017 publication, there are 33 selected agricultural commodities listed in SUA. These include 22 selected agricultural commodities for crops, 5 selected agricultural commodities for livestock and fisheries respectively. Other commodity is reviewed base on the importance of the commodity on that particular of time.

3.1 Crops

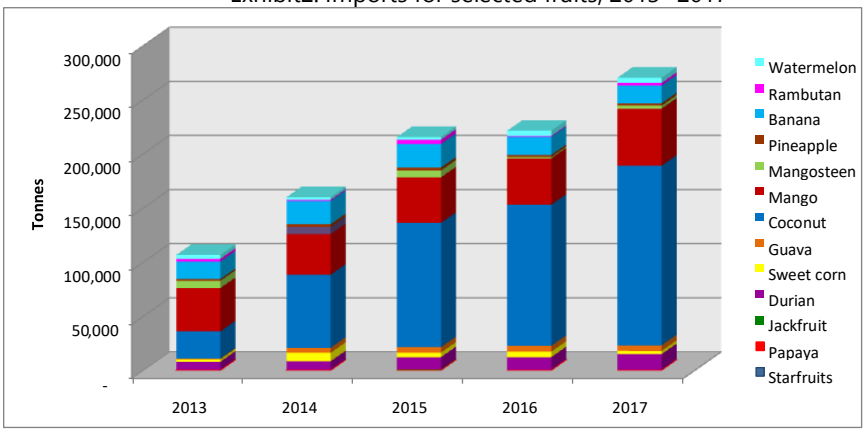
Out of 22 crop items, 13 are fruit items and 9 are vegetable items. Selected fruits consist of coconut, banana, pineapple, durian, watermelon, guava, papaya, sweet corn, jackfruit, rambutan, mango, mangosteen and starfruit. Of the 13 items of the fruits, the SSR for watermelon is highest with 151.1 per cent. This indicates that watermelon production is sufficient and has exceeded domestic demand.

Data in 2017 showed that 177,048.5 tonnes comprising production and import, of 172,275.4 tonnes and 4,773.1 tonnes respectively. The Federal Agricultural Marketing Authority (FAMA) shows that the average retail price for the first week of April, July and October 2017 ranges from RM3.00 to RM3.15 per kilogram; stable throughout the year.

However, the opposite scenario faced for mango, where SSR is 25.3 per cent, the lowest among 13 items of fruits. The signal indicates that mango production is insufficient to meet domestic needs. With the lowest SSR for mango, statistics shows a high impact on IDR where mango IDR is 78.8 per

cent, the highest in the selected fruit category. The production of mango in 2017 was 16,912.6 tonnes, decreased from 17,429.7 tonnes in 2016. Likewise, imports grew by 23.1 percent (2017; 52,751.5 tonnes & 2016; 42,843.1 tonnes).

Exhibit2: Imports for selected fruits, 2013 -2017



SSR for pineapple is also high at 106.3 per cent while IDR pineapple is only 0.7 per cent. This is supported by the statement that Sarawak is aiming to become a major pineapple producer of the country through a commercial pineapple cultivation initiative at Tanjung Manis Halal Hub, operated by Saramanis Sdn Bhd. The success of the company as a pioneer in the project proved that pineapple plants in Tanjung Manis not only expand but also provide job opportunities to locals as well as intend to venture downstream industries through the participation of local residents. Furthermore, it is anticipated that the export of pineapples to Hong Kong will be realized in September and will increase the downstream pineapple industry by encouraging the participation of locals through contract farming programs.

Round cabbage most likely the popular vegetable category as it shows the highest per capita consumption of 5.4 per cent which mean indirectly every person consumed 5.4 kilograms round cabbage per year. Supply of round cabbage recorded 41.8 per cent, dropped from 61.3 percent in 2016. Due to the reduced of supply, as an alternative measure, round cabbage imports have increased (IDR; 58.7 per cent compared to 39.2 per cent the previous year) to meet domestic needs.

Exhibit 3: Total production and imports of round cabbage, 2013 – 2017

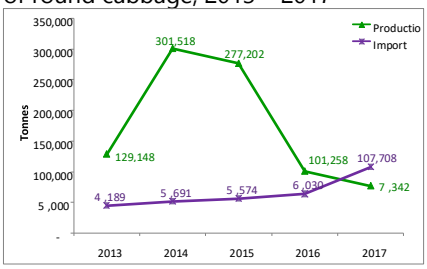


Exhibit 4: Total imports of selected vegetables, 2013 - 2017

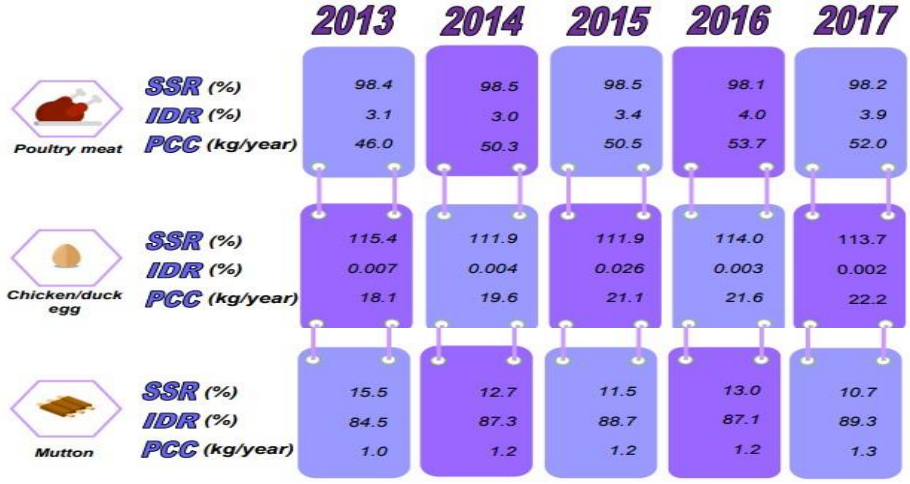
Selected vegetables	Imports (Tonnes)				
	2013	2014	2015	2016	2017
Spinach	513.9	461.8	426.2	469.5	606.4
Lady's finger	42.0	115.5	156.9	296.5	274.1
Chili	36,324.0	41,229.4	47,670.5	49,069.0	47,127.7
Long bean	587.9	753.7	545.6	842.0	521.6
Round cabbage	45,188.7	51,691.3	56,574.1	64,029.5	107,707.6
Mustard	1,333.2	4,622.5	7,032.5	6,011.2	7,709.0
Brinjal	1,595.3	5,737.5	3,886.6	4,097.3	3,493.1
Cucumberx	6,951.4	11,154.8	16,841.9	11,460.1	12,785.1
Tomato	4,156.3	5,609.1	5,987.6	2,576.6	3,618.0

3.2 Livestock

Referring to the Economic Census findings in 2016, the highest value of the gross output is the raising, breeding & production of chicken, broiler, followed by the production of chicken eggs and raising, breeding & production of swine/pigs. This is in line with the highest statistics of production of poultry meat followed by chicken/duck eggs.

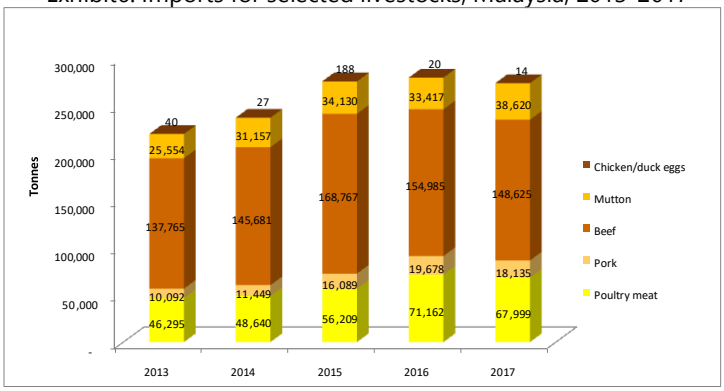
SSR for chicken/duck eggs in 2017 is 113.7 per cent, which is 13.7 per cent exceeding domestic demand, unlike mutton where SSR is low at only 10.7 per cent. This has a direct impact on the IDR where IDR for mutton increased to 89.3 per cent compared to 87.1 per cent in 2016.

Exhibit 5: SSR, IDR and PCC for selected livestock, 2013-2017



The per capita consumption of poultry meat in 2017 is 52 per cent, which means every person consumed 52 kilograms of poultry meat. From the above statistics, "Is the production of poultry meat sufficient to meet domestic needs?". Data showed, since 2013, SSR for poultry meat exceeded 98 per cent and it can be concluded that the production of poultry meat is sufficient and not dependent on imports. Exhibit 6 shows that imports of poultry meat decreased in 2017 (-4.4%) over the previous year.

Exhibit6: Imports for selected livestocks, Malaysia, 2013-2017



3.3 Fisheries

In this paper, fisheries consist of five selected items namely shrimp, mackerel, tuna, cuttlefish and crab. The mackerel species include Temenong/Pelaling, Tenggiri and Kembung. Since 2013, SSR for shrimp and cuttlefish show the supply of this commodity exceeds domestic demand. Similarly, tuna and crab SSR are approaching 100 percent. This indicator shows sufficient supply for shrimp, cuttlefish, tuna and crab. Unfortunately, mackerel SSR is lower (82.9%) compared to other species. Therefore, to fulfil the demand for domestic needs, imports is necessary for these species. The IDR indicator showed an increased from 14.7 per cent in 2016 to 18.3 per cent in 2017. Referring to the Annual Market Intelligence Report 2017 issued by the Malaysian Fisheries Development Authority (LKIM), the highest quantities of imported species are Kembung (Table 1).

Table 1: Import quantities for 8 selected species

Spesies	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Kembung	8,927	8,397	10,705	6,857	7,435	7,742	7,359	7,171	7,004	8,428	11,452	11,089	102,566
Selayang	6,812	6,598	6,650	6,547	6,245	6,059	6,835	6,821	6,862	6,441	8,114	9,222	83,206
Aya/Tongkol	3,786	4,484	4,296	3,867	4,028	3,960	4,399	4,482	4,417	5,165	4,626	4,474	51,984
Cencaru	3,701	4,173	4,668	4,038	4,103	3,954	4,199	3,986	4,079	4,416	3,748	4,241	49,306
Tenggiri	3,234	3,931	3,982	2,659	2,697	2,389	2,466	2,246	2,414	2,911	2,803	2,878	34,610
Pelaling	477	419	439	436	382	322	393	416	498	360	403	382	4,927
Bawal Putih	345	408	448	346	354	286	309	316	399	285	122	100	3,718
Udang Putih Besar	472	77	316	73	130	405	146	152	14	72	48	179	2,084
Jumlah	27,754	28,487	31,504	24,823	25,374	25,117	26,106	25,590	25,687	28,078	31,316	32,565	332,401

Source: Annual Market Intelligence Report 2017, LKIM

It is important not only to provide adequate provision but also to provide sustainable sources of fisheries. It is expected that the demand for the national fish will grow in tandem with the increase in population. Therefore, the demand should be balanced with sufficient supply. In accommodating the

food needs of these protein sources, the government also focuses on aquaculture.

This is in line with what was stated in NAP 2011-2020 where the contribution of aquaculture activities has shown significant growth in the production of fish in the country. This industry development is driven primarily by the development of Aquaculture Industrial Zone equipped with basic infrastructure and effective support services. Aquaculture program will focus on producing quality fish and prawn seedlings, cost-effective livestock feed and aquaculture-based processing activities. The total fish demand is expected to increase from 1.3 million tonnes in 2010 to 1.9 million tonnes by 2020 with growth of 3.8 per cent per annum. The per capita consumption of fish is expected to increase from 46 kilograms to 55 kilograms with growth of 1.9 per cent per annum. Aquaculture production is projected to increase to 790 thousand tonnes (41%) of total fish demand in the country by 2020.

4. Discussion and Conclusion

It is importance to recognise statistics as a tools and evidence based for development planning. Therefore, timely, accurate and reliable statistics are essential for the purpose of providing SUA. The SUA statistical framework for food commodities is a useful reference because it displays statistics on the supply and utilization simultaneously in the form of balance equations. The statistics and indicators provided illustrate the country's agricultural position, especially in determining adequate food supply security.

The Food and Agriculture Organization of the United Nations (FAO) recommends the work program of SUA according to list of commodities classified by major food groups for the provision of SUA to all involved in agricultural data and analysis thereof. Malaysia is among the countries that adopted the program. The compilation of SUA started with nine agricultural commodities in 2010 and expanded to 33 commodities in 2017. However, the number of commodities is not an issue because it is based on its suitability in a country. For the Philippines, the 31st edition of SUA publication, reference year 2015 to 2017 covers 82 agricultural commodities. The publication also displaying accounts in the form of balance equations include elements of supply and usage.

Problems arise primarily on the accuracy and availability of data. For the provision of SUA, secondary data is collected from various agencies. Therefore, it is suggested that cooperation between agencies should be further enhanced. For example, having a workshop, with the presence of representatives from all data provider agencies, detailed discussions can be made regarding the availability and sharing of data. In addition, it is recommended that for the calculation of per capita consumption, the number

of residents in a country should refers to the total population and not only meant for the citizens.

A high impact agriculture project under the NAP is The Permanent Food Production Park (TKPM) where the TKPM's areas are gazetted and rented to entrepreneurs for commercial planting of fruits and vegetables. This program has been developed to focus on food production, especially the production of fruits and vegetables. Various efforts have been undertaken to promote the production of organic vegetables at premium prices, increasing the area for organic farming as well as enhancing the implementation of the Malaysian Organic Scheme certification. The planted area under the TKPM project will be increased to 38 thousand hectares, through increased crop intensity as well as the use of efficient agricultural practices, high yielding breeds and the use of latest technology.

To ensure adequate food supply, various efforts have been taken such as strengthening research and development (R&D) activities, innovation and enhancing technology utilization. This was stated in NAP 2011-2022 under eight key ideas to support the transformation process of the agro-food industry, off which is Modernization of R&D, technology and innovation. Among the efforts undertaken was to enhancing R&D collaboration with the industry through smart collaboration. Examples, MARDI have successfully produced Starfruits hybrid clones (MSTAR 1), *Bintang Mas*. MARDI also produce Fertigation System to protect agricultural production, including during droughts.

In conclusion, self-sufficiency ratio, imports dependency ratio and per capita consumption of selected agricultural commodities can be generated from Supply and Utilization Accounts. Self- sufficiency ratio of 100 per cent or more is good as this indicates sufficient production to meet domestic needs, in contrast, for the import's dependency ratio. The higher the rate of imports dependency ratio indicates the greater supply of agricultural commodities that need to be imported to meet domestic needs. On the other hand, per capita consumption shows the amount of food consumed by each person within a year. Overall, in 2017, Malaysia was able to supply agricultural commodities to meet domestic needs. Of the 33 selected agricultural commodities, 16 recorded SSR exceeding 100 per cent. However, high imports dependency ratio is only for three agricultural commodities namely mango (78.8%), beef/buffalo (76.4%) and goat/lamb (89.3%). Malaysia still manage to balance the needs as the country does not have to rely on imported agricultural commodities to meet domestic needs.

References

1. Food Balance Sheets - A Handbook published by the Food and Agriculture Organization of the United Nations (FAO), Rome, 2001.
2. Supply and Utilization Accounts Selected Agricultural Commodities, Malaysia, 2013-2017.
3. National Agro-Food Policy (NAP) 2011-2020.
4. Annual Market Intelligence Report 2017, Fisheries Development Authority of Malaysia (LKIM).
5. Supply Utilization Accounts (SUA) of Selected Agricultural Commodities, Philippines, 2013-2017.
6. Website, Federal Agricultural Marketing Authority (FAMA), www.fama.gov.my.

Model Averaging on Household Income to Examine Poverty in Malaysia

Siti Aisyah Mohd Padzil, Khuneswari Gopal Pillay, Rohayu Mohd Salleh
Department of Science and Mathematics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Kampus, KM 1 Panchor Road, 84000, Muar Johor Malaysia

Abstract

Household income for Malaysian vary due to many factors as Malaysian are multiracial and multi religious. Malaysia is divided into two regions which are east and west Malaysia. Due to different geographical area, there's a contrast on the development between these two regions. Even though the rate of poverty in Malaysia had been reduce yearly, the poverty of Malaysian still has not been eradicated. This paper aims to examine the consequences of households differences (age, gender, marital status, education, activity and family size), states, region and net income to the cause of poverty among Malaysian in year 2012 using data obtained from the Department of Statistic Malaysia (DOSM). For statistical modelling, this paper will show a comparison between Model Averaging (MA) as well as Alternative Model Averaging (AMA) based on $AICc$ and BIC. The performance of best model from each method will be evaluated using accuracy such as RMSE, MSE and MAE). The analysis indicates that computing weights based on $AICc$ produce a slightly better performance model for both MA and AMA approach. Model formed using AMA based on $AICc$ is picked as the best model of household income data as it has minimum error among other model. In conclusion, the best model summarizes that state, household differences as well as net income are the causes of poverty among Malaysian. There's a negative relationship between household size with probability of not poor which means as household size increases, risk of poverty increases.

Keywords

Household Income; AICc; BIC; Household

1. Introduction

Household poverty is defined by Poverty Line Income (PLI) which measures the capacity of household with the minimum requirement for food and non-food consumption (Saidatulakmal, 2014). Household with gross income under the national PLI will be categorized as poor. PLI in Malaysia varies according to region, urban or rural and it changes over the time. Pasthiban (2018) classified the factors of poverty into two aspects which are income/expenditure-based factors and secondly on non-income factors (example: household characteristic, health, state, amenities and etc.). This

paper will be focusing on obtaining the relationship between household characteristic and state with poverty to examine whether or not it effects among Malaysian's household poverty.

To highlight the causes of household poverty, this paper will provide a study on household income data and uses two modelling methods which are Model Averaging and Alternative Model Averaging. Model Averaging is a well-known statistical modelling approach which aims to solve issues of MS regarding underestimation of parameter estimates. MA decreases the estimates of a weaker variables by averaging weights of all possible model. In other word, MA helps in making inferences based on the entire set of candidate models (Posada and Buckley, 2004). Since MA will give best model that includes all variables being studied, AMA is applied to shows a different modelling approach. AMA approach is build based on MA approach except for elimination of insignificant variables in the final model.

2. Methodology

2.1 Household Income Data

To point out the causes of poverty in Malaysia, data of Household Income retrieved from Department of Statistic Malaysia was studied. The poverty level is measured using net income and Malaysia's PLI as in (Saidatulakmal, 2014). The survey includes 13232 observations and nine independent variables as in Table 1.

Table 1: Description of Household Income

Variable	Description
Y	Poverty Level 1: Not Poor 0: Poor
X_1	State 01: Johor 02: Kedah 03: Kelantan 04: Melaka 05: Negeri Sembilan 06: Pahang 07: Pulau Pinang 08: Perak 09: Perlis 10: Selangor 11: Terengganu 12: Sabah 13: Sarawak 14: Wilayah Persekutuan Kuala Lumpur 15: Wilayah Persekutuan Labuan 16: Wilayah Persekutuan Putrajaya
X_2	Household Age (H.Age) Age of the head of household
X_3	Household Gender (H. Gen) 1: Male 2: Female
X_4	Household Marital (H. Mar) 1: Never married 2: Married 3: Widowed 4: Divorced 5: Separated
X_5	Household Education (H. Edu) Highest level of formal education
X_6	Household Activity (H. Act) 1: Employer 2: Government employee 3: Private employee 4: Own account worker 5: Unpaid family worker 6: Unemployed 7: Housewife/looking after home 8: Student 9: Pensioner 10: Others 11: Child not at school
X_7	Household Size (H. Size) Total number of household member
X_8	Region 1: Peninsular Malaysia 2: Sabah (including Labuan) 3: Sarawak
X_9	Net Income (N. Inc) Net total is the total amount of income

2.2 Multiple Binary Logit

Multiple Binary Logit (MBL) or often called as Logistic Regression is a variation to ordinary linear regression and is used when the dependent variable is binary taking on values 0 and 1. The basic general model of MBL was defined by (Kutner *et al.* 2008) s in Equation (1).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_q X_{qi} + u_i \quad (1)$$

where Yi represent the response variable, β0 is the constant coefficient, q is the number of covariates and ui is the random error for the model.

The results of success/ failure for MBL is presented in forms of probability values which is in between 0 and 1. For example, probability of 0.75 explain that there are 75% chances of success (1) to occur and vice versa. Unlike Multiple regression, the Yi do not gives the probability results, instead the probability, Pi can be obtain using Equation (2).

$$P_i = \frac{exp^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_q X_q}}{1 + exp^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_q X_q}} \quad (2)$$

2.3 Model Averaging and Alternative Model Averaging

Model Averaging discussed in (Claeskens and Hjort, 2008) was proposed to deal with underestimation of parameter estimates issue which come from Model Selection. MA applies weights to all possible models, so that the final best model will include all variables being studied and covariates with higher importance will receive more weight. According to Posada and Buckley (2004), MA will shrink the estimates of a weaker variables. In this study, the weight calculation will be based on $AICc$ and BIC as in (Hurvich and Tsai,1989) and (Schawarz, 1978) respectively. The formula for weight, Wm is as in Equation (3).

$$W_m = \frac{\exp(\frac{I_m}{2})}{\sum_{m=1}^{M} \exp\frac{I_m}{2}} \quad (3)$$

where m is all possible models, m = 1, 2, 3 …, M and Im is the model selection criterion. MA aims to incorporate the estimates of potentially good model by averaging the weights of all possible models to produce estimator for the best model. Equation (4) shows the formula to obtain coefficients estimates for each covariate which average the weights for all possible models.

$$\hat{\beta}_p = \sum_{m=1}^{M} w_m \hat{\beta}_{(p,m)} \quad (4)$$

where $\hat{\beta}(p,m)$ is the estimate of βp under model for m = 1,2,..., M.

Figure 1 summarized the step and procedure of obtaining MA best model based on (Aisyah *et al.* 2018) as well as AMA procedure.

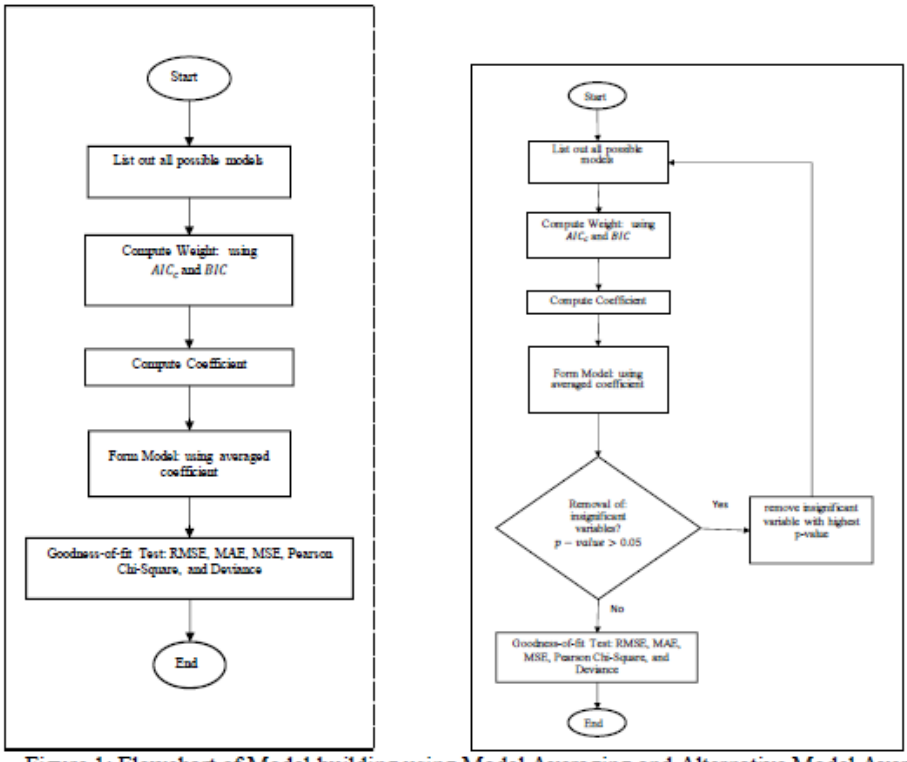


Figure 1: Flowchart of Model-building using Model Averaging and Alternative Model Averaging.

MA approach differs from MS as there is no elimination of insignificant variables in the model-building. Best model of MA will include all covariates being studied in final model even if it's insignificant. As an alternative approach, AMA will produce final model which comprises of only significant variables as minimization of covariates in the final best model will results in numerical stability and reliable results (Bursac, 2008).

Based on Figure 1, the process of AMA is similar with MA except for elimination of insignificant variables. After model was obtain from the combination of computed coefficients, independent variable with -value larger than 0.05 which indicate insignificant variable will be eliminated one by one. This process only allows one insignificant variable to be eliminated at a time and after a covariate is remove, the process of obtaining model will be rerun until best model with only significant variable is obtain.

To compare the accuracy of predictive model produced by traditional MA and alternative method, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Square Error (MSE) are computed. Equation 12, 13 and 14 presents the formula for prediction error as suggested by (Chai and Drexler, 2014).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}{N}} \quad (12)$$

$$MAE = \frac{\sum_{i=1}^{N}|\hat{Y}_i - Y_i|}{N} \quad (13)$$

$$MSE = \frac{1}{N}(Y_i - \hat{Y}_i)^2 \quad (14)$$

where, N is the total number of sample, Y is the actual value of dependent variables and $\hat{Y}_i$ is the estimated value of Y.

3. Result

This paper will only show the results for the full models produced using MA and AMA. Khuneswari *et al.* (2018) had provide a clear, step by step illustration of MA guidelines for further understanding. Table 2 and Table 3 shows the best models formed using MA and AMA with two different weights ($AICc$ and BIC) for Household Income data.

Table 2. Full Model using Model Averaging

Model Selection Criteria	Full Model
AIC_c	$\hat{Y}_i = -0.1030 + 0.001952X_1 + 0.000598X_2 + 0.02808X_3 + 0.01716X_4 + 0.0007363X_5 + 0.01717X_6 - 0.0074845X_7 + 1.023e^{-4}X_8 + 6.098e^{-7}X_9$
BIC	$\hat{Y}_i = -0.1022 + 0.001957X_1 + 0.0005739X_2 + 0.02803X_3 + 0.01729X_4 + 0.000736X_5 + 0.01725X_6 - 0.007483X_7 - 7.49e^{-6}X_8 + 6.098e^{-7}X_9$

Table 3. Full Model using Alternative Model Averaging

Model Selection Criteria	Full Model
AIC_c	$\hat{Y}_i = -0.1029 + 0.001963X_1 + 0.0005984X_2 + 0.02809X_3 + 0.01716X_4 + 0.0007364X_5 + 0.01717X_6 - 0.0078425X_7 + 6.101e^{-7}X_9$
BIC	$\hat{Y}_i = -0.1022 + 0.001959X_1 + 0.000574X_2 + 0.02803X_3 + 0.01729X_4 + 0.000736X_5 + 0.01725X_6 - 0.007483X_7 + 6.101e^{-7}X_9$

In order to select the final best model of Household Income data, as well as to examine which method will produce better performance model, accuracy measure as in Table 4 is computed.

Table 4: Accuracy Measure

Modelling Approach	Model Selection Criteria	RMSE	MAE	MSE
MA	AIC_c	0.1928294	0.0864072	0.0371832
	BIC	0.1928307	0.0865210	0.0371832
AMA	AIC_c	**0.1928272**	**0.0859525**	**0.0371823**
	BIC	0.1928274	0.0859573	0.0371824

Results above shows a small difference of accuracy measure among models. When comparing the performance of model selection criteria, both MA and AMA approach shows a smaller error when $AICc$ is used for weight computation. For the performance of modelling method, AMA gives a slightly lower error in the best model which concludes that the final best model of household income to determine the poverty status is

$$\hat{Y}_i = -0.1029 + (0.001963)\text{State} + (0.0005984)\text{H. Age} + (0.02809)\text{H. Gen} + (0.01716)\text{H. Mar} + (0.0007364)\text{H. Edu} + (0.01717)\text{H. Act} - (0.0078425)\text{H. Size} + (6.101e^{-7})\text{N. Inc}$$

By using equation 2, when assuming the value for all covariates are zero, the probability of the household to be considered as poor (Y=0) is 0.5257 an vice versa. The calculation for P_i is as follows.

$$P_i = \frac{exp^{-0.1029}}{1 + exp^{-0.1029}} = 0.4743$$

$$P_i(Y_i = 0) = 1 - (0.4743) = 0.5257$$

4. Discussion and Conclusion

In comparison of modelling methods, the performance of best model using MA and AMA does not significantly differ. This is because, AMA is build based on MA approach. Even though MA is proven to have overcome underestimation of parameter estimate issues, by eliminating insignificant variables in the final model, it will result in slightly lower error and thus producing a better performance model. Also, when the researchers aim in modelling is to pinpoint the most influential factors, AMA comes in handy.

The results from the final model formed using AMA concludes the characteristic of poor household in Malaysia to be state, household age, household gender, household marital, household education, household activity, household size and net income. Previous study by (Anyawu 2014; Saidatulakmal 2014; Rasyid *et al.* 2018) had highlighted a similar contributing factor of poverty. Household education, household activity and net income are factors that are related as education will affect the head of household occupation and hence the net income. A study by Rasyid et al. (2018) shows that majority of the family head for poor household have a poor education. Household Income shows a negative relationship toward non-poor (Y=1) which means that as household size increases, the probability of becoming non-poor decreases (becoming poor increases). This proven that family planning does have an effect with household poverty.

This paper had shown the application of MA and AMA in modelling household income data. The results could help Malaysian's to plan their future as well as their next generation's to eradicate poverty in Malaysia.

References

1. Aisyah. M.P.S., Khuneswari, G.P. and Rohayu, M.S. (2018). Model-Building of Multiple Binary Logit using Model Averaging. International Journal of Engineering and Technology, pp. 224-228.
2. Anyanwu, John C. (2014). Marital status, household size and poverty in Nigeria: evidence from the 2009/2010 survey data. African Development Review, 26(1) pp. 118-137.
3. Bursac, Z. et al. (2008). Purposeful Selection of Variables in Logistic Regression: Macro and Simulation Results. Section on Statistical Computing, pp. 1886-1891.
4. Chai, T. and Drexler, R. R. (2014). Root Mean Square Error or Mean Absolute Error? Arguments against avoiding RMSE in the literatures. Geoscientific Model Development, 7, pp. 1247-1250.
5. Claeskens, G. and Hjort, N, L. (2008). Model Selection and Model Averaging. United Kingdom: University Press, Cambridge.
6. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, pp. 461–464.
7. Hoeting, J. A., D. Madigan, and A. E. Raftery. (1999). Bayesian model averaging: A tutorial. Statistical Science, (14), pp. 382–417.
8. Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* ,76, pp.297-307.
9. Khuneswari, G.P., Aisyah. M.P.S. and Rohayu, M.P. (2018). Model Selection and Model Averaging on Mortality of Upper Gastrointestinal Bleed Patients. IOSR-JDMS, 11(8), pp. 68-78.
10. Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2008). Applied Linear Regression Models. 4th edition. (McGraw-Hill Inc.: Singapore.
11. Parthiban, S.G. (2018). Poverty measurement revisited from a multidimensional perspective among Universiti Sains Malaysia's B40 poor students. *Malaysian Journal of Society and Space,* 14(4), pp.299-307.
12. Posada, D. and Buckley, T, R. (2004) Model Selection and Model Averaging in Phylogenetics:advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic Biology, 53(5), pp. 793-808.11.
13. Raftery, A. E. (1999). Bayes factors and BIC: Comment on A critique of the Bayesian information criterion for model selection. *Sociol. Methods Res*, 27, pp. 411–427.12.
14. 14.Rasyid, Rusman. et al. (2018). Problem and solution of poverty household in Makassar city, Indonesia. *Academy of Strategic Management Journal,* 7(6).
15. 15.Saidatulakmal, Mohs. (2014). Poverty Issues Among Malaysian Elderly.*Proceeding of the Social Sciences Research,* pp. 123-132.

Causal inference in sport date ~Causal model of fly ball revolution in NPB

Yoshimitsu Morinishi[1], Kazunori Yamaguchi[2]
[1] Graduate School of Business, Rikkyo University
[2] Nishi-Ikebukuro, Toshima-ku, Tokyo

Abstract
In case of Japanese professional baseball NPB, we construct causal inference and causal model from sports data. Specifically, in the theme of "flyball revolution" which is sweeping the major league in the United States in recent years, we construct causal model on whether we can adapt at NPB and examine whether the "flyball revolution" at NPB is correct or not. I will also describe the usefulness of causality reasoning in sports and the usefulness of causal reasoning in observational research in future society. Study the literature of past causal inference and establish a method to extract causal effects from data. Formulate and verify a causal model for baseball 'flyball revolution at NPB'. Construct a causal model and adopt SEM (structural equation model) as a measure method of causal effect. From the result of the constructed causal model, we evaluate and verify flyball revolution at NPB and verify the effectiveness of causal reasoning in sports. From the result of the causal model, the significance of the contribution of each explanatory variable (both latent variable).

Keywords
Causal inference; sport data; Saber Metrics

1. Introduction

Today, with the development of information technology and the spread of the Internet, many data are accumulated, many organizations possessing large amounts of data such as big data, and the era of utilizing data everywhere is about to come. With such changes in society as a background, as the value of data itself rises, it is considered that building a model that can discover a causal relationship between data is extremely valuable. In this research, we will study methodology to clarify causal mechanism underlying natural phenomena and human behavior from sports data, from the above problem consciousness and motivation.

2. Purpose

In case of Japanese professional baseball NPB, we construct causal inference and causal model from sports data. Specifically, in the theme of "flyball revolution" which is sweeping the major league in the United States in

recent years, we construct causal model on whether we can adapt at NPB and examine whether the "flyball revolution" at NPB is correct or not. I will also describe the usefulness of causality reasoning in sports and the usefulness of causal reasoning in observational research in future society.

3. Method

Study the literature of past causal inference and establish a method to extract causal effects from data. For details about causality inference, refer to Rubin (1974, 1976), Rosebaum and Rubin (1983), Rosenbaum (2002), Iwasaki (2015), Hoshino(2009). Formulate and verify a causal model for baseball 'flyball revolution at NPB'. Construct a causal model and adopt SEM (structural equation model) as a measure method of causal effect. From the result of the constructed causal model, we evaluate and verify flyball revolution at NPB and verify the effectiveness of causal reasoning in sports. Specifically, analysis was carried out according to the following procedure.

1. Basic calculation of results for each batting: NPB Aggregate "one bat data" in 2016 and 2017 and calculate the batting result (out or not) or batting strategy (whether frying), Batter player ID, pitcher player ID, and so on.
2. Basic calculations of pitcher · batter's grades: results of pitchers necessary for constructing a causal model (four-ball rate, ball speed etc.) Basic calculate the results of the batter (batting rate, base rate, base stolen, etc.) and combine it with the data of 1.
3. Replacement of swing speed with striking number and homerun count: In this research, the objective is to verify whether "flyball revolution" is effective for each player's swing speed. However, the data of the two years of NBL used this time did not include the batter's swing speed. Therefore, it was necessary to find a substitute variable to replace the swing speed. In the interview active baseball club members, knowledge that the number of strikes and the number of homeruns may have a positive correlation with the swing speed and a causal effect was obtained, so in this study, Principal component analysis was performed on two variables, number and homerun count, and we decided to use a factor common to the two variables as a swing speed substitute variable. Principal component analysis was carried out as described above and the main component score of the obtained factor called swing speed was combined with the data set of 2.
4. Data set creation for 2 left and right batting groups and 10 swing speed groups, totalling 20 groups: Data sets created in 3. Are divided into right and left batting seats, then divided into 10 groups with slow swing speed, fast respectively. Create a total of 20 data datasets.

5. Convert explanatory variable to categorical variable and Convert to polychoric Correlation Matrix: Considering that the objective variable used this time is 01 data, in order to increase the goodness of fitting, all the variables to be used are categorical variables After that, we calculated the polychoric correlation matrix which is the correlation matrix of the categorical variables.
6. Formulation of a causal model for each bat by SEM (structural equation model): Using the SEM (structural equation model) function of JUSE · Stats Works package of Nikka Giken Statistics usiness analysis package, batting strategy We verify the causal effect of the fly ball out of each swing speed verified in this research by verifying the causal effect against the out of the fly ball and comparing

between the groups. SEM is described in detail in Toyoda (1998, 2000) and so on.

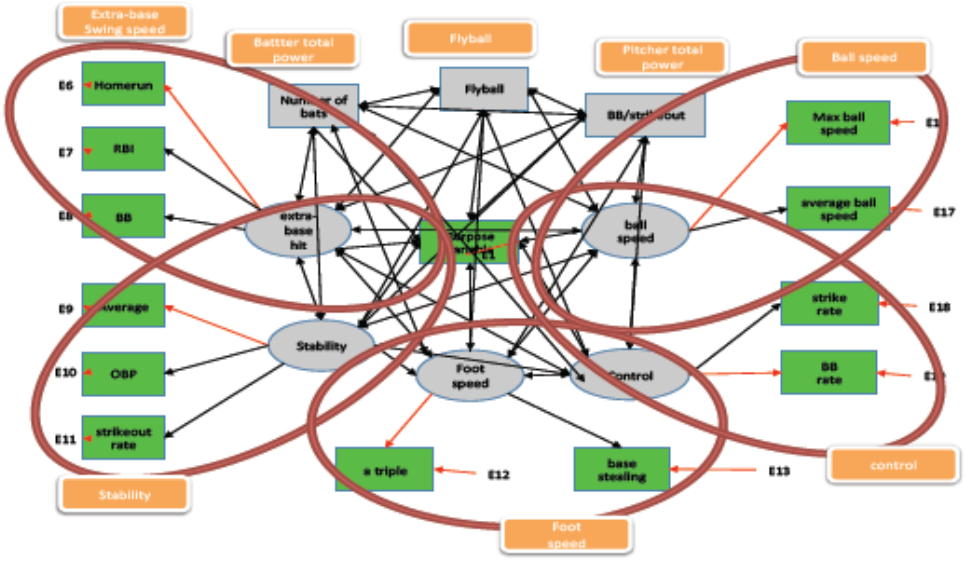


Figure 1 SEM model

4. Outcome

From the result of the causal model, the significance of the contribution of each explanatory variable (both latent variable and observed variable) to the objective variable was confirmed together with the concept. It is thought that it is effective in explanation on the site that it can explain by not only a prediction model as a simple black box but also a model unique to SEM. Moreover, as the result of SEM model was not sufficiently adapted, we adopted logistic regression analysis as the final model, and we were able to verify the effectiveness of "flyball revolution" at NPB. Specifically, I found the following.

Flyball in batting strategy has the effect of reducing out. Right batter is easier to receive flyball benefits in batting strategy compared to left batter. Especially the group that the slow swing speed and the group with the fast swing speed benefit from the flyball. A group with a standard swing speed is likely to benefit from the flyball even though the right batter still has the benefit of flyball but the left batter has less benefit of flyball and there is a possibility that some players will not benefit from the flyball in the fundamental batting strategy There was a need to verify in more detail.

As a conclusion of this study,

1. For the entire NPB, like the MBL, the fly leather revolution should be promoted.
2. The right batteries especially benefit from the fly-ball revolution.
3. It may be necessary to consider whether to adopt the Fribore Revolution in consideration of the swing speed of each player. With this conclusion, we have been able to fulfill "verification of the fly leather revolution at NPB" which is an empirical study in this research. However, future emerging issues can be considered for this empirical study.

5. Task

1. Construction of a more persuasive model: In addition to the explanatory variables currently being used, batting results and pitcher performance, it is considered that factors that can explain the batting situation are left behind. In order to evaluate the flyball in an accurate batting strategy, it is considered necessary to improve the model in consideration of defense capability and stadium.
2. Small range of swing speed group, individual: In this study, we constructed a causal model by dividing into 10 groups of left and right batting and swing speed into a total of 20 groups, but in a smaller range, ultimately, It is thought that it is necessary to construct a causal model to judge whether flyball is positive or negative.
3. Evaluation of flyballs by circumstances: In this study, we do not define the situation and verify the causal effect of the flyball for each swing speed, but we think that the strategy required by the runner situation and outcounting is different in the actual game Therefore, it is required to verify the causal effect of the flyball in the bait track for each situation.
4. Construction of causal model using hitting angle and swing speed: In this research. For the convenience of data, we analyzed the striking number and the number of homeruns as the principal component analysis and analyzed the principal component score as the swing speed, but in the theory of "barrel zone" which is originally the basis of

the "flyball revolution", the hitting angle and hitting ball Since it depends on the speed and thus the swing speed, it may be possible to analyze the accuracy by constructing a causal model using the data of the hitting angle and the swing speed.

By working on the above problem and constructing a causal model it is thought that we can verify the causal effect in the batting strategy more precisely and make it a valuable thing.

6. Conclusion / Discussion

Through a series of analyze, the effectiveness of flyball revolution in NPB was verified. Although this series analysis still leaves room for improvement, it seems that we could verify with certain value. Construction of a causal model in sports is effective, and it is considered to be very useful for formulating a sports strategy. It is considered effective to build a causal model by the following procedure.

1. Observation of events for which causal effects are to be examined.
2. Definition of events for which causality effects are to be investigated.
3. Preliminary research and interview surveys of events for which causality effects are to be examined.
4. Prediction of causal effects of events for examining causality effects.
5. Construction of a causal model (definition of independent variables and explanatory variables and covariates).
6. Measurement of causal effects.
7. Verification of obtained causal effects.

   Also, in the AI era widely, estimating the causal effect from the construction of the causal model can be said to be very meaningful.

References

1. ROSENBAUM,2002, Observational studies, Springer
2. ROSENBAUM, P.R. and RUBIN, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), pp. 41-55.
3. RUBIN, D.B., 1976. Inference and missing data. Biometrika, 63(3), pp. 581-592.
4. RUBIN, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational psychology, 66(5), pp. 688.
5. Manabu Iwasaki, 2015. Statistical causal reasoning Statistical analysis standard Asakura Shoten
6. Hideki Toyoda, 1998. Covariance structure analysis primer Structural equation modeling (statistical
7. library). Asakura Shoten
8. Hideki Toyoda, 2000. Covariance structure analysis application. Structural equation modeling
9. (statistical library). Asakura Shoten
10. Takahiro Hoshino, 2009. Statistical science of survey observation data - causality inference selection vias data fusion, probability and information science.

Tests for mean vector using approximate degrees of freedom with two-step monotone missing data

Tamae Kawasaki, Takashi Seo
Tokyo University of Science, Tokyo, Japan

Abstract

In this study, we consider testing for the mean vector with two-step monotone missing data. Many statistical methods have been developed to analyse data with missing values. Additionally, the monotone missing data have been widely studied in the past. Kawasaki and Seo (2016) derived the asymptotic expansion of the Hotelling's type test statistics for the case where the sample size is large with two-step monotone missing data. The asymptotic first two moments are obtained using stochastic expansion. The goal of our research is to propose approximate solutions, which are simpler and better convenience than previous studies. We approximate the distribution for the Hotelling's *T2* type test statistics by constant times an F distribution by adjusting the degrees of freedom. The method of adjusting the degrees of freedom are estimated unknown parameters of degrees of freedoms of the F distribution using the asymptotic expansion of the Hotelling's *T2* type test statistic by Kawasaki and Seo (2016). The accuracy of the approximation is investigated using Monte Carlo simulation.

Keywords

asymptotic expansion; *F* approximation; missing data; multivariate normal

1. Introduction

In almost all statistical analyses, missing data is a constantly occurring problem. In this study, we consider the problem of testing for normal mean vectors when the data set has two-step monotone missing observations.

Let $\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_{N_1}^{(1)}$ be distributed as the multivariate normal $N_p(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_{N_2}^{(2)}$ be distributed as the multivariate normal $N_{p_1}(\boldsymbol{\mu}_1, \Sigma_{11})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and two-step monotone missing data are drawn from a multivariate normal population of the form

$$\begin{pmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p_1}^{(1)} & x_{1p_1+1}^{(1)} & \cdots & x_{1p}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p_1}^{(1)} & x_{2p_1+1}^{(1)} & \cdots & x_{2p}^{(1)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{N_1 1}^{(1)} & x_{N_1 2}^{(1)} & \cdots & x_{N_1 p_1}^{(1)} & x_{N_1 p_1+1}^{(1)} & \cdots & x_{N_1 p}^{(1)} \\ x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1p_1}^{(2)} & * & \cdots & * \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{N_2 1}^{(2)} & x_{N_2 2}^{(2)} & \cdots & x_{N_2 p_1}^{(2)} & * & \cdots & * \end{pmatrix}$$

where , $N = N_1 + N_2, p = p_1 + p_2$, and '*' indicates missing data. We partition $\boldsymbol{x}_j^{(1)}$ into a $p_1 \times 1$ random vector and a random vector as $\boldsymbol{x}_j^{(1)} = (\boldsymbol{x}_j^{(1)'}, \boldsymbol{x}_{2j}^{(1)'})'$, where $j = 1,2, \ldots, N_1$. We define the sample mean vectors as

$$\overline{\boldsymbol{x}}^{(1)} = (\overline{\boldsymbol{x}}_1^{(1)'}, \overline{\boldsymbol{x}}_2^{(1)'})' = \frac{1}{N_1}\left(\sum_{j=1}^{N_1} \boldsymbol{x}_{1j}^{(1)'}, \sum_{j=1}^{N_1} \boldsymbol{x}_{2j}^{(1)'}\right)',$$

$$\overline{\boldsymbol{x}}^{(2)} = \frac{1}{N_2}\sum_{j=1}^{N_2} \boldsymbol{x}_j^{(2)},$$

and the sample covariance matrices are given as

$$S^{(1)} = \begin{pmatrix} S_{11}^{(1)} & S_{12}^{(1)} \\ S_{21}^{(1)} & S_{22}^{(1)} \end{pmatrix} = \frac{1}{N_1 - 1}\sum_{j=1}^{N_1} (\boldsymbol{x}_j^{(1)} - \overline{\boldsymbol{x}}^{(1)})(\boldsymbol{x}_j^{(1)} - \overline{\boldsymbol{x}}^{(1)})',$$

$$S^{(2)} = \frac{1}{N_2 - 1}\sum_{j=1}^{N_2} (\boldsymbol{x}_j^{(2)} - \overline{\boldsymbol{x}}^{(2)})(\boldsymbol{x}_j^{(2)} - \overline{\boldsymbol{x}}^{(2)})'.$$

We consider the following hypothesis test problem:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{ vs. } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ is a specified vector. We can construct the Hotelling's T2 type test statistic as

$$T^2 = (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)'\{\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\mu}})\}^{-1}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0),$$

Where

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_1 \\ \widehat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{N}(N_1 \overline{\boldsymbol{x}}_1^{(1)} + N_2 \overline{\boldsymbol{x}}^{(2)}) \\ \overline{\boldsymbol{x}}_2^{(1)} - \widehat{\Sigma}_{21}\widehat{\Sigma}_{11}^{-1}(\overline{\boldsymbol{x}}_1^{(1)} - \widehat{\boldsymbol{\mu}}_1) \end{pmatrix},$$

$$\widehat{\Sigma} = \begin{pmatrix} \frac{1}{N}(W_{11}^{(1)} + W^{(2)}) & \widehat{\Sigma}_{11}(W_{11}^{(1)})^{-1}W_{12}^{(1)} \\ W_{21}^{(1)}(W_{11}^{(1)})^{-1}\widehat{\Sigma}_{11} & \frac{1}{N_1}W_{22\cdot 1}^{(1)} + \widehat{\Sigma}_{21}\widehat{\Sigma}_{11}^{-1}\widehat{\Sigma}_{12} \end{pmatrix},$$

$$\widehat{\mathrm{Cov}}[\widehat{\boldsymbol{\mu}}] = \begin{pmatrix} \frac{1}{N}\widehat{\Sigma}_{11} & \frac{1}{N}\widehat{\Sigma}_{12} \\ \frac{1}{N}\widehat{\Sigma}_{21} & \widehat{\mathrm{Cov}}[\widehat{\boldsymbol{\mu}}_2] \end{pmatrix},$$

$$\widehat{\mathrm{Cov}}[\widehat{\boldsymbol{\mu}}_2] = \frac{1}{N_1}\left(\widehat{\Sigma}_{22} - \frac{N_2}{N}\widehat{\Sigma}_{21}\widehat{\Sigma}_{11}^{-1}\widehat{\Sigma}_{12}\right) + \frac{N_2 p_1}{N N_1(N_1 - p_1 - 2)}\widehat{\Sigma}_{22\cdot 1},\ (N_1 > p_1 + 2),$$

and

$$W^{(1)} = (N_1 - 1)S^{(1)} = \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \end{pmatrix},\quad W_{22\cdot 1}^{(1)} = W_{22}^{(1)} - W_{21}^{(1)}(W_{11}^{(1)})^{-1}W_{12}^{(1)},$$

$$W^{(2)} = (N_2 - 1)S^{(2)} + \frac{N_1 N_2}{N}(\overline{\boldsymbol{x}}_1^{(1)} - \overline{\boldsymbol{x}}^{(2)})(\overline{\boldsymbol{x}}_1^{(1)} - \overline{\boldsymbol{x}}^{(2)})'.$$

Anderson and Olkin (1985) developed an approach to derive the maximum likelihood estimators (MLEs) of the mean vector and the covariance matrix with several missing patterns. Kanda and Fujikoshi (1998) proposed the distribution of the MLEs in the cases of two-step, three-step, and general monotone missing data. For a two-step monotone missing data, Seko, Yamazaki and Seo (2012) derived Hotelling's T^2 type statistic and an accurate simple approach to give the upper percentiles in one-sample problem, and Seko, Kawsaki and Seo (2011) provided Hotelling's T^2 type statistic of testing for two normal mean vectors and its approximate upper percentile. Kawasaki and Seo (2016) derived the asymptotic expansion of the Hotelling's T^2 type test statistics for large sample and proposed the Bartlett corrected statistics with one-sample problem. Their results are theoretical results; however, the equation is slightly complicated.

The aim of this study is to propose simple and convenient approximations with two-step monotone missing data by adjusting the degrees of freedom. For adjusting degrees of freedom, Yanagihara and Yuan (2005) provided some approximate solutions to the multivariate Behrens-Fisher problem that are two F approximations with approximate degrees of freedom for complete data. Kawasaki and Seo (2015) proposed some new approximate solutions by deriving the asymptotic expansions up to the term of order N^{-2} for the moments of test statistic under the multivariate Behrens-Fisher problem with complete data. Note that the asymptotic expansions up to the term of order N^{-1} for the moments of test statistic are obtained by Yanagihara and Yuan (2005). Krishnamoorthy and Pannala (1999) derived an approximate distribution of the Hotelling's type test statistic by a constant time an F distribution using the decompositions of the statistics.

In the following section, we propose approximate solutions by adjusting the degrees of freedom of F distribution. We perform Monte Carlo simulations in Section 3.

2. Methodology

In this section, we will consider approximate solutions with two-step monotone missing data. In this study, we employ the asymptotic expansion of the Hotelling's T^2 statistic by Kawasaki and Seo (2016) in a situation when

$$\gamma_i = \frac{N_i - 1}{N - 2} \to \text{ positive constants},$$

as (N_i-1)'s tend to infinity. In our derivation, we consider the stochastic expansions of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ in terms of

$$\boldsymbol{z}^{(1)} = \begin{pmatrix} \boldsymbol{z}_1^{(1)} \\ \boldsymbol{z}_2^{(1)} \end{pmatrix} = \sqrt{N_1 - 1} \begin{pmatrix} \overline{\boldsymbol{x}}_1^{(1)} \\ \overline{\boldsymbol{x}}_2^{(1)} \end{pmatrix}, \quad \boldsymbol{z}^{(2)} = \sqrt{N_2 - 1}\, \overline{\boldsymbol{x}}^{(2)},$$

$$V^{(1)} = \sqrt{N_1 - 1}(S^{(1)} - \Sigma) = \begin{pmatrix} V_{11}^{(1)} & V_{12}^{(1)} \\ V_{21}^{(1)} & V_{22}^{(1)} \end{pmatrix}, \quad V^{(2)} = \sqrt{N_2 - 1}(S^{(2)} - \Sigma_{11}).$$

Note that the stochastic expansions are derived under $\mu = \mu_0 = 0$ and $\Sigma = I_p$. Then, the Hotelling's T^2 type statistic can be written as

$$T^2 = \left(\sqrt{N-2}\ \hat{\boldsymbol{\mu}}\right)' \left\{(N-2)\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}})\right\}^{-1} \left(\sqrt{N-2}\ \hat{\boldsymbol{\mu}}\right),$$

and $\sqrt{N-2}\hat{\mu}$ and $(N-2)\widehat{\mathrm{Cov}}(\hat{\mu})$ can be expanded as

$$\sqrt{N-2}\ \hat{\boldsymbol{\mu}} = \begin{pmatrix} \sqrt{\gamma_1}\boldsymbol{z}_1^{(1)} + \sqrt{\gamma_2}\boldsymbol{z}^{(2)} \\ \frac{1}{\sqrt{\gamma_1}}\boldsymbol{z}_2^{(1)} \end{pmatrix} + \frac{1}{\sqrt{N-2}} \begin{pmatrix} \boldsymbol{0} \\ \frac{\sqrt{\gamma_2}}{\sqrt{\gamma_1}} V_{21}^{(1)} \boldsymbol{z}^{(2)} - \frac{\gamma_2}{\gamma_1} V_{21}^{(1)} \boldsymbol{z}_1^{(1)} \end{pmatrix}$$

$$+ \frac{1}{N-2} \begin{pmatrix} \left(\frac{1}{\sqrt{\gamma_1}} - 2\sqrt{\gamma_1}\right)\boldsymbol{z}_1^{(1)} + \left(\frac{1}{\sqrt{\gamma_2}} - 2\sqrt{\gamma_2}\right)\boldsymbol{z}^{(2)} \\ \frac{\gamma_2}{\gamma_1\sqrt{\gamma_1}} V_{21}^{(1)} V_{11}^{(1)} \boldsymbol{z}_1^{(1)} - \frac{\sqrt{\gamma_2}}{\gamma_1} V_{21}^{(1)} V_{11}^{(1)} \boldsymbol{z}^{(2)} \end{pmatrix} + \mathrm{O}_p((N-2)^{-\frac{3}{2}}),$$

$$(N-2)\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}})$$

$$= \begin{pmatrix} I_{p_1} & O_{12} \\ O_{21} & I_{p2}/\gamma_1 \end{pmatrix} + \frac{1}{\sqrt{N-2}} \begin{pmatrix} \sqrt{\gamma_1} V_1 1^{(1)} + \sqrt{\gamma_2} V^{(2)} & V_{12}^{(1)}/\sqrt{\gamma_1} \\ V_{21}^{(1)}/\sqrt{\gamma_1} & V_{22}^{(1)}/(\gamma_1\sqrt{\gamma_1}) \end{pmatrix}$$

$$+ \frac{1}{N-2} \begin{pmatrix} \gamma_1\gamma_2 \left(\frac{1}{\sqrt{\gamma_1}}\boldsymbol{z}_1^{(1)} - \frac{1}{\sqrt{\gamma_2}}\boldsymbol{z}^{(2)}\right)\left(\frac{1}{\sqrt{\gamma_1}}\boldsymbol{z}_1^{(1)} - \frac{1}{\sqrt{\gamma_2}}\boldsymbol{z}^{(2)}\right)' - 4I_{p1} & \frac{\sqrt{\gamma_2}}{\sqrt{\gamma_1}} V^{(2)} V_{12}^{(1)} - \frac{\gamma_2}{\gamma_1} V_{11}^{(1)} V_{12}^{(1)} \\ \frac{\sqrt{\gamma_2}}{\sqrt{\gamma_1}} V_{21}^{(1)} V^{(2)} - \frac{\gamma_2}{\gamma_1} V_{21}^{(1)} V_{11}^{(1)} & \frac{p_1\gamma_2 - 2}{\gamma_1^2} I_{p_2} - \frac{\gamma_2}{\gamma_1^2} V_{21}^{(1)} V_{12}^{(1)} \end{pmatrix}$$

$$+ \mathrm{O}_p(N^{-\frac{3}{2}}),$$

respectively. Therefore, the Hotelling's T^2 type statistic can be expanded as

$$T^2 = Q_0 + \frac{1}{\sqrt{N-2}} Q_1 + \frac{1}{N-2} Q_2 + O_p((N-2)^{-(3/2)}).$$

For details, see Kawasaki and Seo (2016). Using this results and assuming the standard regularity condition $N_i/N = O(1), i = 1,2,$ we can write

$$T^2 = \frac{\boldsymbol{w}'\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{w}/(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)'\{\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}})\}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)} = \frac{\boldsymbol{w}'\boldsymbol{w}}{U}.$$

By approximating the distribution of U as

$$U \approx \frac{\chi_\nu^2}{\phi},\ \boldsymbol{w} \sim N_p(\boldsymbol{0}, I_p),$$

we have

$$\frac{\nu}{p\phi}T^2 \approx \frac{\chi_p^2/p}{\chi_\nu^2/\nu} \sim F_{p,\nu}.$$

The constants ν and ϕ are decided using the first and second moments of U. That is, we can follow

$$\mathrm{E}[U] \approx \frac{\nu}{\phi},\ \mathrm{E}[U^2] \approx \frac{\nu(\nu+2)}{\phi^2}.$$

Therefore, using the asymptotic expansions of E[U]and E[U^2], the new approximation to the values of ν and ϕ are given. In addition, we propose another approximate solution by adjusting the degrees of freedom of F distribution. In the above results, we estimated the second degree of freedom ν and the coefficient ϕ , but further estimate the first degree of freedom ξ as

$$\frac{\nu}{\xi\phi}T^2 \approx \frac{\chi_\xi^2/\xi}{\chi_\nu^2/\nu} \sim F_{\xi,\nu},$$

and the constants ν, ϕ and ξ are decided using the first, second and third moments of T^2 . That is, we can follow

$$\mathrm{E}[T^2] \approx \frac{\nu\phi}{\nu-2},\ \mathrm{E}[(T^2)^2] \approx \frac{\phi^2\xi(\xi+2)}{(\nu-2)(\nu-4)},\ \mathrm{E}[(T^2)^3] \approx \frac{\phi^3\xi(\xi+2)(\xi+4)}{(\nu-2)(\nu-4)(\nu-6)}.$$

Therefore, using the asymptotic expansions of E[T^2], E[$(T^2)^2$] and E[$(T^2)^3$], the new approximation to the values of ν, ϕ and ξ are given.

3. Results

We give the two-step monotone missing data to be generated from multivariate normal distribution by Monte Carlo simulation (10^4 runs) and calculate the upper 100α percentile of T^2 type statistics and F approximations, and type I error with significant level α=0.05,0.01 where $p = 4((p_1, p_2) = (2,2))$, which are given by

$$\alpha_1 = \mathrm{P}\left(T^2 > x_p^2(\alpha)\right), \alpha_2 = P(T^2 > \frac{p\phi F_{p,v}(\alpha)}{v})$$

Note that the simulated values approach closer to the upper percentiles of chi-squared distribution when sample size become large. The type I error rates show that α2 are very good approximations.

Table 1

Upper percentiles and type I error rate

		α=0.05				α=0.01			
N_1	N_2	T2	F	α_1	α_2	T2	F	α_1	α_2
10	10	15.16	14.54	0.162	0.057	25.22	22.82	0.073	0.015
20	20	13.43	13.23	0.131	0.053	21.03	20.25	0.049	0.012
40	40	11.12	11.10	0.086	0.050	16.68	16.21	0.027	0.012

80	80	10.31	10.24	0.067	0.050	14.29	14.63	0.016	0.010

Note :$x_4^2(0.05) = 9.49, x_4^2(0.01) = 13.28$

4. Discussion and Conclusion

In this study, we have developed the approximate distribution of Hotelling's T^2 type test statistics by a constant times an F distribution by adjusting the degrees of freedom. The method of adjusting the degrees of freedom is estimated unknown parameters of the degrees of freedom of the F distribution using the asymptotic expansion of the Hotelling's T^2 type test statistic. The approximate values can be calculated easily and the approximation is much better than the chi-squared approximation, even when the sample size is small.

References

1. Anderson, T. W. and Olkin, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra and its Applications*, **70**, 147-171.
2. Kanda, T. and Fujikoshi, Y. (1998). Some basic properties of the MLE's for a multivariate normal distribution with monotone missing data, *American Journal of Mathematical and Management Sciences*, **18**, 161-190.
3. Kawasaki, T. and Seo, T. (2016). Bias correction for T2 type statistic with two-step monotone missing data, *Statistics*, **50(1)**, 76-88.
4. Krishnamoorthy, K. and Pannala, M. K. (1999). Confidence estimation of a normal mean vector with incomplete data, *The Canadian Journal of Statistics*, **27**, 395-407.
5. Seko, N., Kawasaki, T. and Seo, T. (2011). Testing equality of two mean vectors with two-step monotone missing data, *American Journal of Mathematical and Management Science*, **31(1- 2)**, 117-135.
6. Seko, N., Yamazaki, A. and Seo, T. (2012). Tests for mean vector with two-step monotone missing data, *SUT Journal of Mathematics*, **48(1)**, 13-36.
7. Yanagihara, H. and Yuan, K. (2005). Three approximate solutions to the multivariate Behrens-Fisher problem, *Communications in Statistics – Simulation and Computation*, **34**, 975–988.

Reduced *k*-means with nonlinear principal component analysis

Takatsugu Yoshioka[1],Masahiro Kuroda[2],Yuichi Mori[2]
[1] Graduate School of Informatics, Okayama University of Science, Japan
[2] Department of Management, Okayama University of Science, Japan

Abstract
Reduced *k*-means analysis (RKM) is a useful method for clustering objects in a low-dimensional subspace by conducting *k*-means clustering and dimension reduction simultaneously. Here RKM for categorical data and mixed measurement level data is considered. Although there have been several methods for categorical data based on RKM, a method of RKM which combines nonlinear principal component analysis (NLPCA) as dimension reduction with *k*-means clustering is proposed to deal with not only nominal variables but also ordinal and combination of categorical and numerical ones, and to provide the effective information for interpretation of the variables as well as the relationships between objects/clusters in the reduced subspace. A couple of numerical experiments demonstrate the performance of the proposed method.

Keywords
Dimension reduction; Clustering, Alternating least squares optimal scaling; Categorical data; Simultaneous estimation

1. Introduction

Reduced *k*-means analysis (RKM) is a method for clustering objects in a low-dimensional subspace, hat is, a simultaneous analysis in which *k*-means clustering and dimension reduction are conducted at the same time to obtain clustering of objects and low-dimensional subspace reflecting the clusterstructure (De Soete and Carroll, 1994).

The original RKM is developed for continuous data and there are several approaches/extensions based on RKM, e.g., Vichi and Kiers (2001), Timmerman et al., (2010), Vidal (2011), Yamamoto and Hwang (2014), which combine dimension reduction such as principal component analysis (PCA) with *k*-means clustering. When categorical data is analyzed by RKM, an appropriate quantification of categorical variables is necessary in the analysis. In such case, multiple correspondence analysis (MCA) is often used to quantify categorical variables. There are a number of studies combining MCA with *k*-means to find clustering objects consisting of categorical variables in a low-dimensional subspace with category quantifications, e.g., MCA *k* -means (Hwang et al., 2006), iFCB (Iodice D'Enza and Palumbo, 2013), Reduced *k*-means clustering

with MCA (Mitsuhiro and Yadohisa, 2015), and Cluster Correspondence Analysis (van de Velden et al., 2017).

From the aspect of categorical analysis, it is better to deal with not only single nominal variables but also multiple nominal, ordinal, and combination of measurement level including numeric variables. To do this, an optimal scaling by Alternating Least Squares (ALS) can be utilized. So-called Nonlinear PCA (NLPCA), which is carried out by algorithm PRINCIPALS (Young et al., 1978) or PRINCALS (Gifi, 1990), is one of possibilities for effective interpretation of the variables as well as the relationships between objects/clusters in the reduced subspace (GROUALS by Van Buuren and Heiser (1989) has the same concept for quantification). Considering RKM which includes PCA as a dimension reduction procedure and provides information to interpret the configuration of principal components scores of objects and the loading of each variable, NLPCA can be used as a dimension reduction method in RKM for data including categorical variables.

Thus, we propose a simultaneous analysis of *k*-means clustering and NLPCA (we refer this RKM with NLPCA) to find clustering objects in a low-dimensional subspace with category quantifications.

2. Methodology

We here introduce ordinary RKM and NLPCA briefly based on De Soete and Carroll (1994) and Mori et al. (2017), respectively, before proposing RKM with NLPCA for categorical data / mixed measurement level data.

2.1 Ordinary reduced *k*-means analysis

Ordinary RKM (for numerical data) is as follows.

Let X be the $n \times p$ centered data matrix, where n denotes the number of objects and p the number of variables, k be the number of clusters, r be the number of components (generally, $k \geq r+1$), **U** be the $n \times k$ membership matrix, and **A** be the $p \times r$ loadings matrix, **Y= XA** is the $n \times r$ object scores (component scores) matrix. RKM looks for centroids in a low-dimensional subspace that minimize the distance of the data points from the centroids. The RKM minimizes the following loss function

$$fRKM(\mathbf{U}, \mathbf{F}, \mathbf{A}) = \left\| \mathbf{X} - \mathbf{U}\mathbf{F}\mathbf{A}^{\mathrm{T}} \right\|^2, \quad (1)$$

where **F** is the $k \times r$ matrix collecting centroids. We can confirm which of k clusters each object belongs to (from the estimated **U**), where each centroid of k clusters is (from the estimated **F**), which direction of loading of each variable is (from the estimated **A**), and others in the r dimensional subspace. PCA, estimating **U**, and estimating **F** are alternately executed until convergence.

1. **Optimal scaling with alternating least squares**

NLPCA reveals nonlinear relationships among variables with different measurement levels and therefore presents a more flexible alternative to ordinary CA.

Let **X** consist of p categorical variables, each of which have k_j categories ($j = 1, \ldots, p$). Let $\mathbf{X}^*$be an optimal scaled matrix of data, i.e., j-th optimal scaled variable $\mathbf{x}_j^* = \mathbf{G}_j\mathbf{y}_j$, where $\mathbf{G}_j$ is $n \times k_j$ indicator matrix, $\mathbf{y}_j$ is $k_j \times r$ category quantifications of variable j. NLPCA can find solutions by minimizing two types of loss functions, a low-rank approximation and homogeneity analysis with restrictions. The former loss function is

$$\sigma(\mathbf{Z}, \mathbf{A}, \mathbf{X}^*) = \|\mathbf{X}^* - \mathbf{Z}\mathbf{A}^T\|^2, \quad (2)$$

Where **Z** is $n \times r$ object scores. The minimization of loss functions has to take place with respect to both parameters. The ALS algorithm is utilized to solve such minimization problem. We describe the general procedure of an ALS algorithm, PRINCIPALS (Young et al., 1978) that minimizes the loss function (2). PRINCIPALS accepts nominal, ordinal and numerical variables, and alternates between two estimation steps. The first step estimates the model parameters **Z** and **A** for ordinary PCA, and the second obtains the estimate of the data parameter $\mathbf{X}^*$ for optimally scaled data. Given the initial data $\mathbf{X}^{*(0)}$, PRINCIPALS iterates the following two steps:

Model estimation step:

By solving the Eigen-decomposition of $\mathbf{X}^{*(t)\mathrm{T}}\mathbf{X}^{*(t)}/n$ or the singular value decomposition of $\mathbf{X}^{*(t)}$, obtain $\mathbf{A}^{(t+1)}$ and compute $\mathbf{A}^{(t+1)} = \mathbf{X}^{*(t)}\mathbf{A}^{(t+1)}$.Update $\mathbf{X}(t+1)=\mathbf{Z}(t+1)\mathbf{A}(t+1)\mathrm{T}$. $\widehat{\mathbf{X}}^{(t+1)} = \mathbf{Z}^{(t+1)}\mathbf{A}^{(t+1)\mathrm{T}}$

Optimal scaling step:

Obtain $\mathbf{X}^{*(t+1)}$ by separately estimating $\mathbf{x}_j^*$ for each variable j. Compute for nominal variables as $\mathbf{y}_j^{(t+1)} = (\mathbf{G}_j^T\mathbf{G}_j)^{-1}\mathbf{G}_j^T\hat{\mathbf{x}}_j^{(t+1)}$. Re-compute $\mathbf{y}_j^{(t+1)}$ for ordinal variables using the monotone regression (Kruskal,1964). For nominal and ordinal variables, update $\mathbf{x}_j^{*(t+1)} = \mathbf{G}_j\mathbf{y}_j^{(t+1)}$ and standardize $\mathbf{x}_j^{*(t+1)}$. For numerical variables, standardize observed vector $\mathbf{x}_j$ and set $\mathbf{x}_j^{*(t+1)} = \mathbf{x}_j$.

2.3 RKM with NLPCA

We here propose a RKM with NLPCA based on RKM and NLPCA. Roughly speaking, RKM with NLPCA is a method obtained by replacing dimension reduction procedure (PCA step) in ordinary RKM by NLPCA which provides both quantification and dimension reduction. The loss function is

$$f_{RKM/NLPCA}(\mathbf{U}, \mathbf{F}, \mathbf{A}) = \|\mathbf{X}^* - \mathbf{U}\mathbf{F}\mathbf{A}^T\|^2 \quad (3)$$

The algorithm is as follows:

[Step1] *Initialization*: Determine the desired number of cluster k and components r, considering the relation $k \geq r+1$. If you want the reasonable number of r, you may apply NLPCA to the data once. Assign random values to **U**.

[Step2] *Quantification*: Quantify $\mathbf{X}^{(t+1)}$ by PRICIPALS and obtain quantified data matrix $\mathbf{X}^{*(t)}$.
[Step 3] *Clustering*: Minimize (3) to estimate the cluster allocations and centroids by k-means algorithm (obtain $\mathbf{U}^{(t)}$, $\mathbf{F}^{(t)}$ and $\mathbf{A}^{(t)}$).
[Step 4] *Termination*: Compute the difference between the current (t-th) value of loss function (3) and the previous ((t-1)-th) value of loss function (3). If it is sufficiently small or the maximum number of iterations has been reached, stop. Otherwise, set t=t+1 and go back to Step 2.

3. Results

We illustrate two examples, mild disturbance of consciousness (MDOC) data (Sano et al., 1977) and test score data. We apply RKM with NLPCA to MDOC data, which collected responses from 81 patients who have mild disturbance of consciousness. The responses are observational results on 23 items (see the list under Figure 1). All items are categorical: five levels in 21 items and two levels in other 2 items.

Figure 1 is a biplot of the first two dimensions drew by RKM with NLPCA with k=3 and r=2.

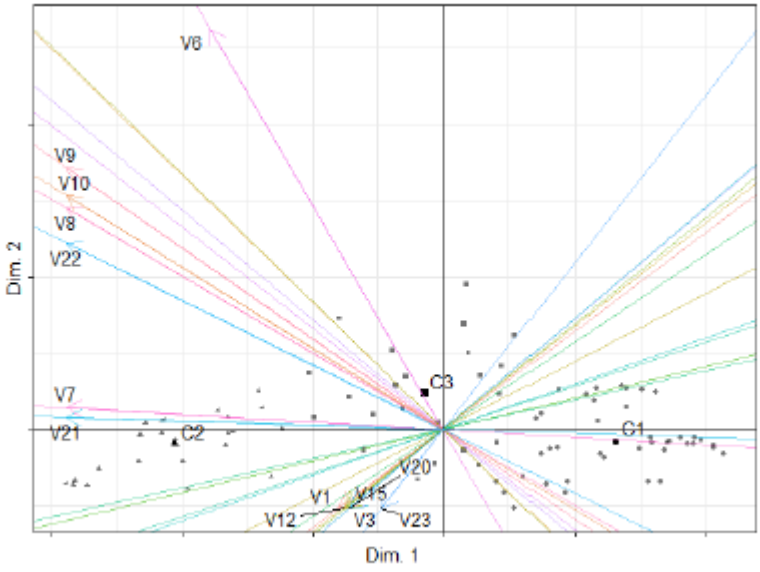


Fig1: Biplot of RKM with NLPCA including centroids (MDOC data, k=3, r=2)

V1: eating, V2: urinary incontinence, V3: response to calling or greeting, V4: orientation (place), V5:orientation (season), V6:orientation (date), V7:orientation (hour), V8: orientation (person). V9: grade of patient's insight, V10: volition, V11: knowledge, V12: response to command, V13: counting from 1 to 20, V14: calculation, V15: quality of voice, V16: facial expression, V17: attitude during examination, V18: spontaneous movement, V19: spontaneous speech, V20: attention, V21: tendency to perseveration, V22: stating date of birth, V23: stating name.

We can find three centroids that line up along the first axis and cluster 81 component points. We can also interpret the configuration of principal components and the meaning of variables from the loading vectors such that the proposed method derives two principal axes dividing 23 items to patients' attitudes and behaviours.

Next, we examine the performance of RKM with NLPCA using test score data. We got the data consisting of 40 students' test score (categorical, five levels rating) of 9 subjects (we refer this data TSc). Form this data, we generated a numerical data by assigning a random number to original categorical score according to the rating score (we refer this data TSo). Regarding this numerical data as a true continuous structure, we apply RKM with NLPCA to TSc and ordinary RKM to TSo with k=4 and r=3 and compare them.

Figure 2 is a biplot of ordinary RKM to TSo and Figure 3 a biplot of RKM with NLPCA to TSc.

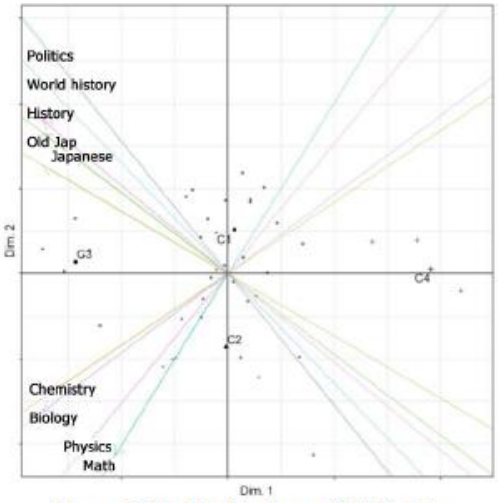


Figure 2: Biplot of ordinary RKM to Tso

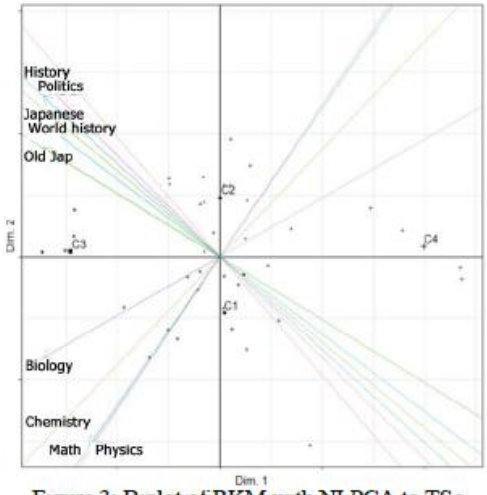


Figure 3: Biplot of RKM with NLPCA to TSc.

Table 1: Coordinates of centroids in three dimensions

Left: RKM to TSo, Right: RKM with NLPCA to TSc

Squares: Sum of square of the differences of coordinates for two dimensions and three dimensions

Cluster	Dim.1	Dim.2	Dim.3	Cluster	Dim.1	Dim.2	Dim.3
1	0.154	1.034	-0.127	1	0.113	-1.358	0.046
2	-0.038	-1.725	-0.082	2	0.008	1.426	0.066
3	-3.566	0.272	0.306	3	-3.644	0.126	-0.158
4	4.770	0.114	0.360	4	4.985	0.253	-0.178
				Squares	15.743	16.300	

Table 2: Number of objects belonging to each cluster

Method	Cluster1	Cluster2	Cluster3	Cluster4
RKM to TSo data	18	12	6	4
RKM with NLPCA to TSc data	16	14	6	4

We show the coordinates of centroids in three dimensions in Table 1 to check how close these configurations are. Left table is coordinates of TSo centroids estimated by RKM, and right table ones of TSc by RKM with NLPCA. We also illustrate the number of objects in each cluster on each dimension in Table 2.

We can see that the difference of these two configurations is totally very small and that component scores and cluster centroids are very similar, although the loading vector is slightly different and the numbers of objects in Cluster 1 and Cluster 2 are not the same. That is, it can be stated that RKM with NLPCA reproduces the original low-dimensional structure, if TSo is the true structure of TSc.

4. Discussion and Conclusion

We propose RKM with NLPCA which combines NLPCA with *k*-means clustering to observe the clusters of objects in data including categorical variables in a low-dimensional subspace. The proposed method can basically provide the estimations of component scores, clusters with their centroids, and loadings in the subspace and reproduce the low-dimensional structure well.

We have to investigate the performance in detail, for examples, how the proposed method behaves for more complex date and how to avoid a local minima problem, and compare the proposed method with other methods in previous studies.

References

1. De Soete G, Carroll JD (1994). K-means Clustering in a Low-Dimensional Euclidean Space. In E Diday, Y Lechevallier, M Schader, P Bertrand, B Burts chy (eds.), New Approaches in Classification and Data Analysis, 212–219.
2. Fordellone M., Vichi M. (2017). Multiple Correspondence K-Means: Simultaneous Versus Sequential Approach for Dimension Reduction and Clustering. In: Lauro N., Amaturo E., Grassia M., Aragona B., Marino M. (eds) Data Science and Social Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham.
3. Gifi, A. (1990). Nonlinear multivariate analysis. Wiley, Chichester, England.
4. Hwang, H., Dillon, W. R., Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. Psychometrika, 71, 161-171.
5. Iodice D'Enza, A., Palumbo, F. (2013). Iterative factor clustering of binary data. Computational Statistics, 28(2), 789-807.
6. Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115-129.
7. Mitsuhiro, M., Yadohisa, H. (2015). Reduced k-means clustering with MCA in a low-dimensional space, Computational Statistics, Springer, 30(2), 463-475.
8. Mori. Y., Kuroda, M., Makino, N. (2017). Nonlinear Principal Component Analysis and Its Applications. JSS Research Series in Statistics, Springer.
9. Sano, K. et al. (1977). Statistical studies on evaluation of mind disturbance of consciousness -Abstraction of characteristic clinical pictures by cross-sectional investigation. Sinkei Kenkyu no Shinpo, 21, 1052-1065. (in Japanese)

10. Timmerman. M., Ceulemans. E., Kiers. HAL., Vichi. M.（2010）. Factorial and Reduced K-means Reconsidered. Computational Statistics and Data Analysis, 54(7), 1858–1871.
11. Van Buuren, S., Heiser, W. J. (1989). Clustering n objects in k groups under optimal scaling of variables, Psychometrika, 54, 699-706.
12. van de Velden M., Iodice D' Enza, A., Palumbo, F. (2017). Cluster correspondence analysis. Psychometrika, 82(1), 158-185.
13. Vichi, M., Kiers, H.A.L. (2001). Factorial K-means analysis for two-way data. Computational Statistics and Data Analysis, 37, 49-64.
14. Vidal, R. (2011). Subspace clustering. IEEE Signal Processing Magazine 28(2):52–68.
15. Yamamoto, M., Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41, 115-129.
16. Young, F.W., Takane, Y., de Leeuw, J. (1978). Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 43, 279-281.

Too many zeros? Are two-part models a good choice for the analysis of longitudinal data in health care research?

Iris Reinhard
Biostatistics, Central Institute of Mental Health, Medical Faculty Mannheim / Heidelberg University, Germany

Abstract

In health care research it is common to encounter data characterized by a spike at zero followed by a bcontinuous distribution for the positive values. Examples include health care utilization and health care expenditures, or food consumption in a dietary study. In the first case the point mass at zero represents a population of 'non-users' who therefore have no costs, while the continuous distribution represents the level of costs for those people who use health services. For statistical analyses in order to understand the influence of therapies, programs, demographic and disease-related variables, alternative approaches are needed to accommodate the discrete and continuous features of the data. For the identification of possibly influencing factors on semi-continuous longitudinal data a two-part model is considered which is based on a two-stage design. The first stage involves modelling the risk for the occurrence of a positive outcome and the second stage models the intensity or the amount of nonzero outcomes. Within that model two sets of covariates / factors can be modelled simultaneously that contribute to separate stages. The hierarchical structure of the data is accounted for by including random effects. In a simulation study the performance of this model is evaluated in terms of type I error and the mean squared error (MSE) of the estimates, under different levels of sample size and correlation between covariates as well as correlation between random effects. The data generation process is thereby based on the distribution characteristics of an empirical data set coming from a controlled prospective intervention study which is investigating the cost-effectiveness of an intervention to reduce compulsory admission into inpatient psychiatric treatment. Finally, the results are compared to conventional linear mixed models. The proposed two-stage model performs well for the analysis of semicontinuous health care data which represent the structure of real cost-effectiveness data. With increasing sample size the performance improves. The classical linear mixed model has to be discouraged because it produces inflated type I error and much higher MSEs than the two-part model.

Keywords

Health care expenditures; semicontinuous data; zero modified data; linear mixed model; simulation study

1. Introduction

In health services research it is common to encounter data characterized by an abundance of zeros in combination with a continuous distribution for positive values. Examples can be health care utilization, health care expenditures or food consumption in a dietary study. In the first case the point mass at zero represents a population of 'non-users' who therefore produce no costs, while the continuous distribution represents the level of costs for those people who use health services. For the identification of possibly influencing factors on outcomes of this type, the two-part model is considered to accommodate the discrete and continuous features of the data, and investigated by a simulation study.

2. Methodology

Semicontinuous data can be regarded as a result of two stochastic processes, one for the occurrence of zeros and the second for the observed value given a non-zero response (Neelon et al., 2016). The two - part model seems to be a perfect choice for the analysis of such data, because it is a mixture model involving two basic components:

- Component 1: relates to the risk for the occurrence of a positive outcome (binary outcome model)
- Component 2: regresses the intensity or amount of non-zero outcomes.

A normal distribution can be chosen to model the non-zero values, leading to

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } 1-\phi_{ij} \\ N(\mu_{ij},\sigma), & \text{with probability } \phi_{ij} \end{cases}$$

where Y_{ij} denotes the outcome for subject i at time t_{ij} (i=1, ...n; j=1, ..., m) from a finite mixture (e.g. the cost of inpatient stays). The normal assumption can be relaxed by using alternative distributions for the non-zero outcomes, e.g. lognormal, gamma.

The logistic-normal two-part model can be written as

$$f(y_{ij}) = (1-\phi_{ij})\, I_{(y_{ij}=0)} + (\phi_{ij} \times N(y_{ij},\mu_{ij},\sigma^2))\, I_{(y_{ij}>0)},\ y_{ij} \geq 0,\ 0 \leq \phi_{ij} \leq 1$$

Extending the two-part model to the regression setting, two sets of covariates/predictors can be modelled simultaneously. Let xij be a vector of potential risk factors for the intensity of the outcome and zij another vector of risk factors that are linked with the probability of having a positive response. Then the model for longitudinal semicontinuous data is parameterized as follows

$$\begin{aligned} \text{logit}(\phi_{ij}) &= z_{ij}\gamma + b_i \\ \mu_{ij} &= \mathrm{x}_{ij}\beta + a_i \end{aligned}$$

The β and γ are fixed effects for the covariates xij and zij, respectively, and ai and bi are random effects generating the within-subject correlation and the between-subject heterogeneity. It is assumed that

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$$

Simulation study

In a simulation study the performance of the two-part model is evaluated in terms of type I error and the mean squared error (MSE) of the estimates. The data generation process is thereby based on the distribution characteristics of an empirical data set coming from a controlled prospective intervention

study which is investigating the cost-effectiveness of an intervention to reduce compulsory admission into inpatient psychiatric treatment. So it is possible to deduce more generalized conclusions for health services research. Thereby a mixture distribution framework involving a two-step process is applied. To generate the covariates, random samples from a bivariate normal distribution with different values of correlation (e.g. ρ_{x1x2}=-0.2 or ρ_{x1x2}=0) are drawn. We employ six repeated measurements and manipulate the sample size at four levels: sparse (N=50), small (N=100), medium (N=200) and large (N=500). The true model incorporates two fixed effects for each component (two covariates): β_k (Normal), γ_s (Binary); $k,_s$=1,2, as well as two random effects, each scalar quantities, one for each component (for the binary part *bi* / for the normal part *ai*), correlated or uncorrelated (ρ_{ab}=-0.4 or ρ_{ab}=0). For each scenario 5000 replications are conducted. The nominal α is set to 0.05.

The performance of the mixed two-part model is evaluated under the null hypothesis H_0: β_k=0 resp. γ_s=0, for estimating the significance level and under the H1 for estimating the parameters. Criteria are (i) the type I error and (ii) the mean squared error (MSE) of the estimates. As a control model to examine the consequences of ignoring zero modified data the classical linear mixed model is used.

The programming environment for the implementation of the two-part model is SAS 9.4, where the procedure NLMIXED is employed (see e.g. Liu et al., 2010, Tooze et al., 2002). This procedure directly maximizes an approximate integrated likelihood. Here an adaptive Gaussian quadrature approximation is chosen. As an optimization algorithm to carry out the maximization, a Newton-Raphson optimization is applied which combines a line-search algorithm with ridging. This technique uses the gradient and the Hessian matrix.

3. Results

Results I – Type I Error

Figure 1 illustrates effects of varying sample sizes and scenarios on the type I error rates (in accordance with more results not shown due to limited number of pages). The performance improves with increasing sample size, N=100 seems to be sufficient, while N=50 cannot be recommended. For the impact of the covariate correlation one can find a better performance for uncorrelated covariates.

For the correlation of random effects no obvious effect was found, the level properties seem to be slightly better when uncorrelated. The simulations indicate that the type I error level is kept to a great extent. For the linear mixed model, moderate to strong differences are found, depending on the parameter configuration, partly worse in the normal component (because of "modelling"), see Figure 2. Especially when the effect is located only in the zero component, the type I error for the corresponding effect in the normal component which is the only one in the linear mixed model, is inflated, because it might comprise the fixed effect on the probability for the occurrence of a non-zero outcome. This also holds for large sample sizes.

Results II – Mean Squared Error (MSE)

The simulation results which are summarized in parts in Table 1 demonstrate that the empirical biases for the two-part model are negligible. With increasing sample size the performance regarding the mean squared error improves. The correlation between the covariates does not show an obvious effect. The correlation of the random effects of the two components does not indicate an impact on the MSE (not shown). The evaluation of the two-part model showed a much better performance than the classical linear model, the estimates of which are heavily biased. This also applies for large sample sizes.

Table 1: MSE for varying sample sizes under the two models for the upper left scenario of Figure 1

Parameter N=100	True value	Two-Part Mixed Model		Classical Linear Mixed Model	
		estimate	MSE	estimate	MSE
Normal component					
β_1	-0.020	-0.0199	0.00015	0.0596	0.00715
β_2	0.030	0.0303	0.00022	0.0186	0.00145
Zero component					
γ_1	-0.040	-0.0397	0.00023	-	-
γ_2	0.000	0.0002	0.00033	-	-

Parameter N=200	True value	Two-Part Mixed Model		Classical Linear Mixed Model	
		estimate	MSE	estimate	MSE
Normal component					
β_1	-0.020	-0.0199	0.00007	0.0598	0.00677
β_2	0.030	0.0302	0.00011	0.0190	0.00074
Zero component					
γ_1	-0.040	-0.0400	0.00010	-	-
γ_2	0.000	-0.0002	0.00015	-	-

4. Discussion and Conclusion

In this study a two-part mixed model for semicontinuous longitudinal data is considered and evaluated by simulation. The model performs well for the simulated data which represent the structure of real data from health economy. The estimated effects are very close to their theoretical values. With increasing sample size the MSE decreases and the α-level is maintained from N≈100. Using the classical linear mixed model has to be discouraged because it produces inflated type I error and much higher MSEs for the maximum likelihood parameters, in particular when the effect is located only in the zero component. The estimation can be implemented in SAS using the procedure NLMIXED. A major advantage is the simultaneous modelling and estimation of the two components. An alternative distribution to model the non-zero values can be the lognormal or the gamma distribution for the continuous part. Two-part mixed models encompass a considerable subset of parametric models for semicontinuous longitudinal data previously suggested in the literature.

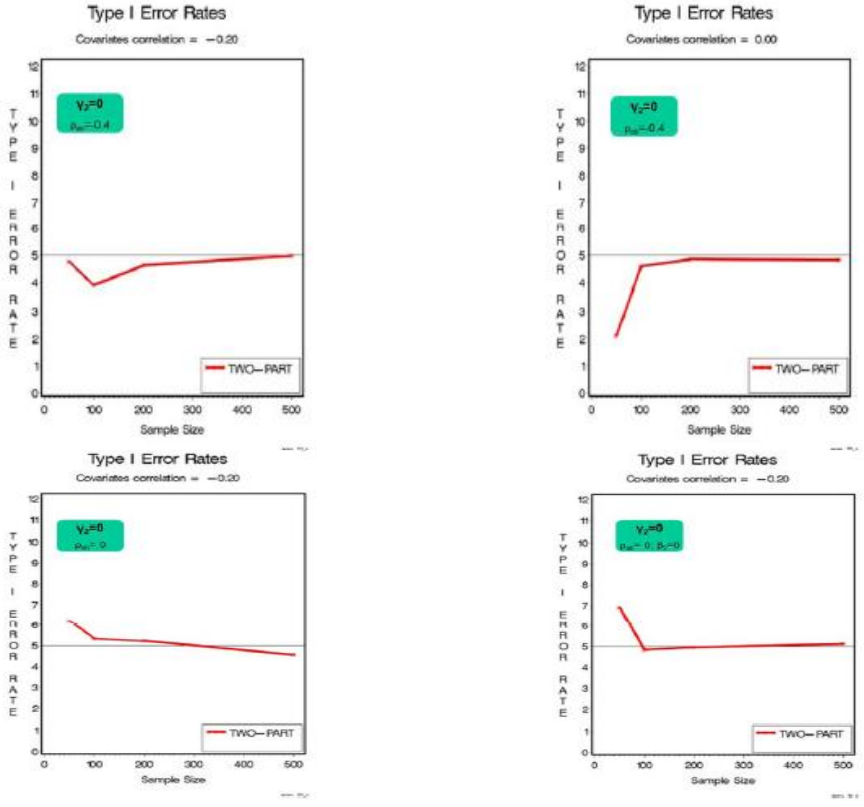


Figure 1: Type I error rates in the zero component (binary) for some scenarios of the simulation study (varying sample sizes, correlation of covariates, correlation of random effects, effect combinations)

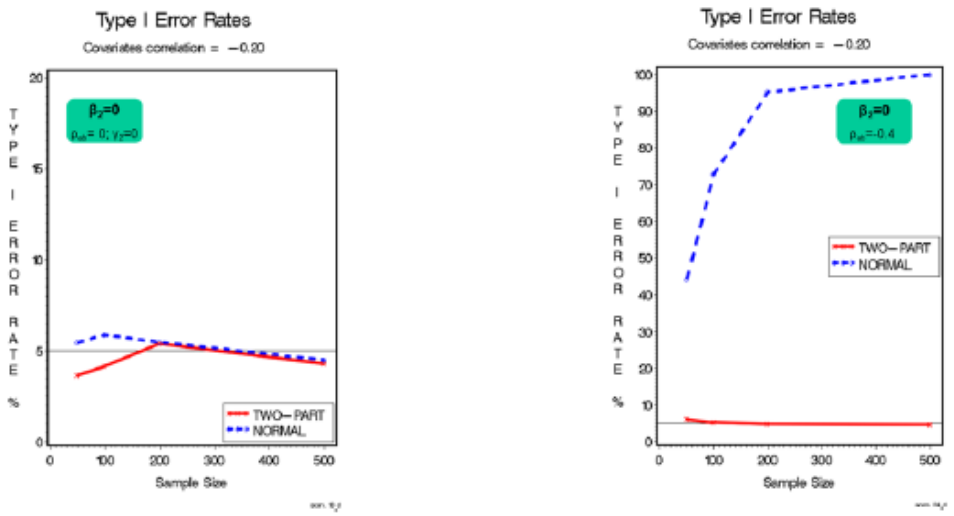


Figure 2: Type I error rates in the normal component for varying sample sizes

References

1. Liu L, Cowen ME, Strawderman RL, Shih YT (2010). A flexible two-part random effects model for correlated medical costs. J Health Econ. 29(1): 110-123.
2. Neelon B, O'Malley AJ, Smith VA (2016). Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview. Stat Med. 35(27): 5070-5093.
3. Neelon B, O'Malley AJ, Smith VA (2016). Modeling zero-modified count and semicontinuous data in health services research Part 2: case studies. Stat Med. 35(27): 5094-5112.
4. Tooze JA, Grunwald GK, Jones RH (2002). Analysis of repeated measures data with clumping at zero. Stat Methods Med Res. 11(4): 341-355.

Assessing trust in official statistics for Southeaster European Countries

Helda Curma , Delina Ibrahimaj
Institute of Statistics Albania

Abstract

To ensure the official statistics quality and reliability, the statistical office shall be guided and follow international agreed principles and methods and try to build trust in the statistics produced. Every statistical office makes its own efforts on following the principles of Professional independence, Impartiality, Objectivity, Reliability, Statistical confidentiality and Cost effectiveness. A very important issue is to evaluate the trust in official statistics in order to undertake actions to improve the positioning of the official statistics in society. This paper has used "OECD - Measuring Trust in Official Statistics Cognitive Testing" as a conceptual framework. This framework has been built to measure trust among users of official statistics, while the same principles have been used to make a self-assessment of the statistical offices regarding trust related factors. The results of this paper are based on a survey distributed to the 10 statistical offices (Greece, Bulgaria, Albania, Croatia, Serbia, Kosovo, Turkey, Montenegro, Macedonia and Bosnia and Hercegovina.

Keywords

Reliability; National Statistical Institutes; Quality; Positioning

Introduction

Trust is a very important aspect of official statistical work. To assess users and statistical offices opinion regarding the trust in the statistics produced the conceptual framework of OECD – "Measuring Trust in Official Statistics Cognitive Testing" is used. This framework has been built to measure trust among users of official statistics, while the same principles have been used to make a self-assessment of the statistical offices regarding trust related factors. The National Statistical Offices Survey was a web-based questionnaire prepared in English and contained seven questions. The questionnaire collected information regarding:

- the perception of statistical offices about public confidence in official statistics from the standpoint of the statistical offices themselves and the perception of users;
- the assistance provided by different national/international organisations received for infrastructure, statistical production and organisational behaviour;

- the quality dimensions, as specified in the European statistics Code of Practice, impacted from the received assistance;
- standardisation initiatives undertaken by NSOs, as defined by the UNECE, High-Level Group for the Modernisation of Official Statistics;
- the extent collaboration with international institutions improved the positioning of the National Statistical Offices and areas where support is needed in the near future.

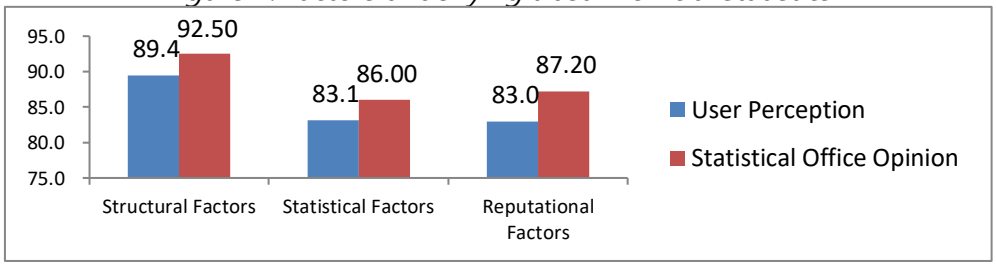
The three sets of factors[1], stated in the European statistics Code of Practice and also on the Albanian law on Official Statistics, underlying trust in official statistics measured in this paper were evaluate from 1 to 5(1 – Strongly disagree,2 – Disagree, 3 – Neither agree or disagree, 4 – Agree and 5 – Strongly agree) and they are as follow:

a. Structural factors, including the extent to which the statistics are, or are perceived as being, objective and independent, impartial, non-partisan and transparent.
b. Statistical factors (including sound statistical processes and quality outputs);
c. Reputational factors which in turn are affected by a number of national practices and considerations: a commitment to informing the public, through the provision of relevant statistics, about major issues of national importance; regular consultation processes; relationship with the media and other key stakeholders; past incidents of erroneous data or unethical behaviour; the preparedness of the agency to openly correct or address misleading or inaccurate media reports; etc.

National Statistical Offices Survey Results

The statistical office opinion has been measured and user perception regarding the three sets of factors (Structural factors, Statistical factors and Reputational factors) underlying trust in official statistics. In the figure 1 are shown the results of the user perception and statistical office opinion for the factors impacting the trust in official statistics. In general Statistical Office evaluates the trust around 88.6 % and the user perceptions around 85.3%.

Figure 1: Factors underlying trust in official statistics

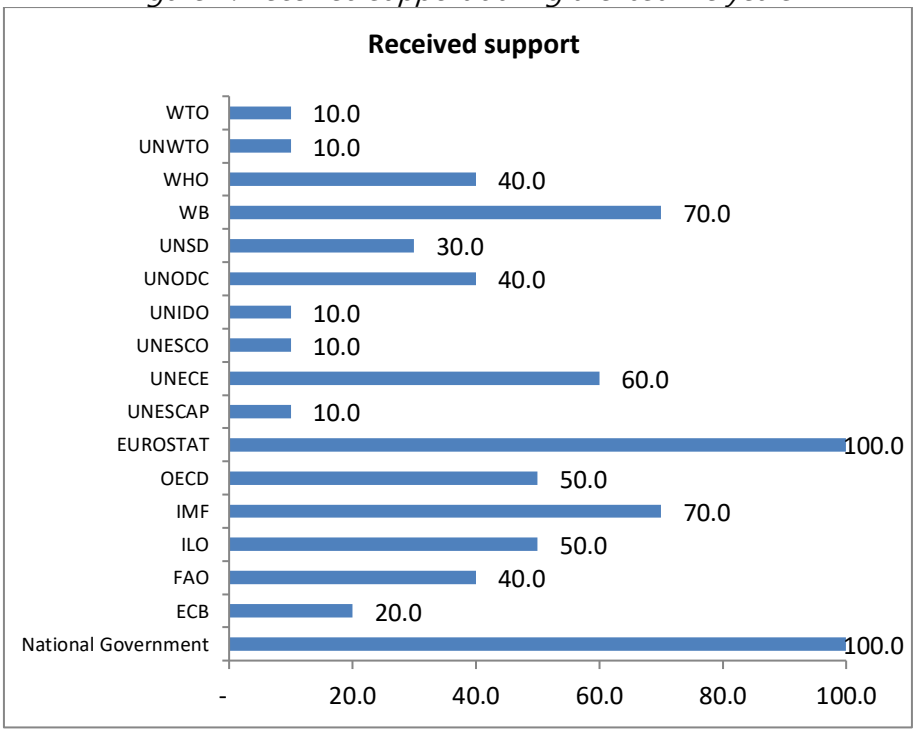


[1] http://www.oecd.org/sdd/50027008.pdf

It is noted that for the three factors of trust, the perception of the statistics office is higher than the perception of users. The difference in percentage between two perceptions is higher for Reputational factors (5.06%) followed by Statistical factors (3.48%) and Structural factors (3.42%).

The NSOs answered also if their institutions have received support and which kind of support from the national /international institutions during the last five years. In figure 2: Received support during the last five years is presented the support given in the last five years from 17 different National/International Institutions out of 23 National/International Institutions listed.

Figure 2: Received support during the last five years



National government and Eurostat have provided support for all statistical offices that fulfilled this questionnaire, followed by WB, IMF, UNECE, OECD and ILO. The kind of support received during the last five years, has been evaluated, as it is considered one important factor in trust in official statistics. Technical Assistance for Statistical production is the main area where support has been given, approximately 47%, while Technical Assistance for Organisational behaviour has been evaluated around 11%.

Official statistics evaluated the impact of the support received and their position in different user groups in order to build and raise awareness regarding trust in statistical products and institution. As Technical Assistance for Statistical production is the main area of intervention from National/International Institutions, NSOs has been asked to evaluate the four

main dimensions impacted (Transparency, Quality, Communication, Statistical production).

Figure 3: Trust related factors impacted by Technical Assistance

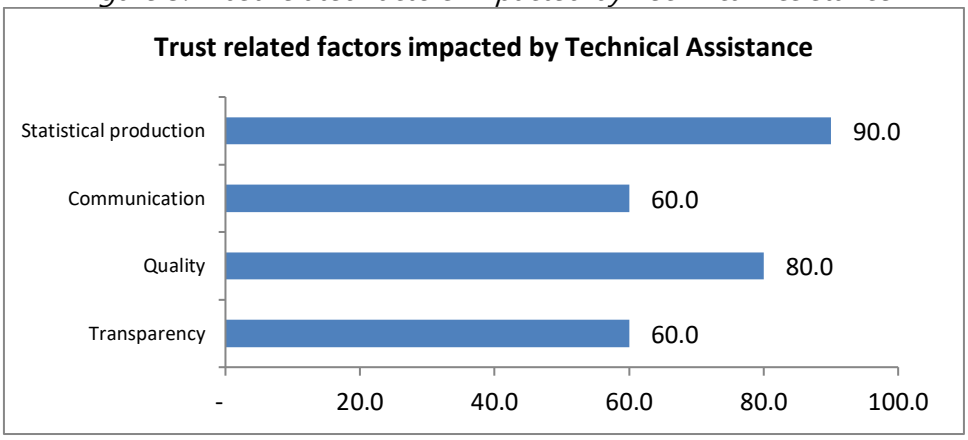


Figure 3 shows that almost all NSOs have estimated that the main areas impacted by the provided technical assistance for statistical production and organizational behaviour are: production of official statistics and overall quality in the organisation.

Trust related factors for statistical products have been measured based on the following dimensions: Accuracy, Reliability, Credibility, Objectivity, Relevance, Timeliness, and Coherence. While the regarding the statistical institution's factors such as Confidentiality, Integrity, Openness/ transparency, Impartiality and Effective stakeholder management has been taken into account.

From the factors directly related to the trust in statistical products, credibility, objectivity, relevance and timeliness have been mostly impacted as a result of the technical assistance received. On the other hand factors directly related to the trust in the statistical institution are confidentiality followed by openness/ transparency.

Figure 4: Dimensions impacted by TA

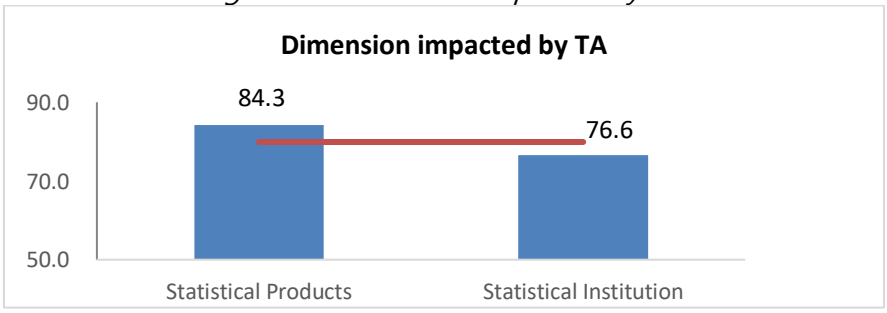
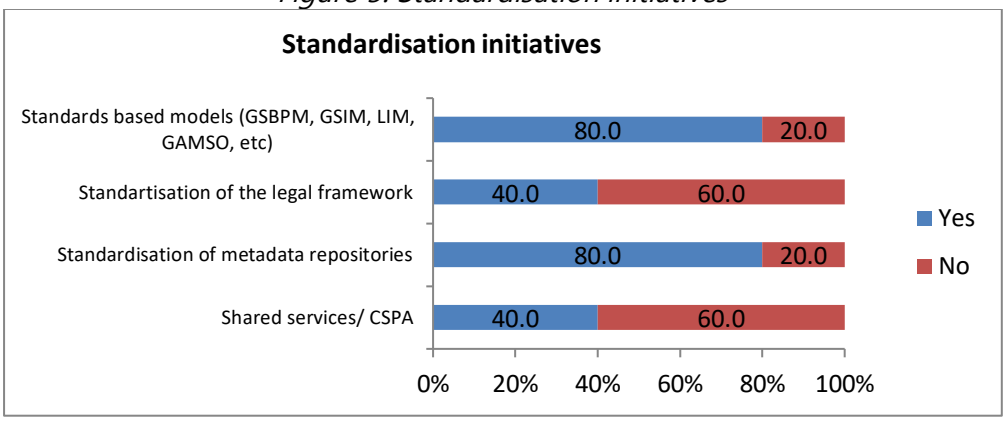


Figure 4 presents the fact that technical assistance received has been impacting trust in statistical products more than the one in statistical institution.

Standardisation initiatives are other elements strongly related to trust factors. Standards2 will consolidate the use of our statistical products in the global information community, improving their accessibility, interpretability and comparability.

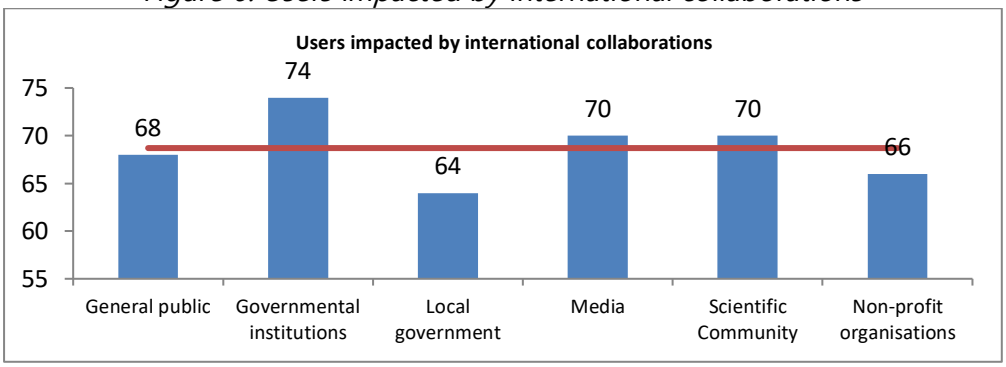
Figure 5: Standardisation initiatives



From the different international standardisation initiatives, standard based models and standard metadata repositories are the two most applied from NSOs.

The impact of collaboration with international institutions in strengthening the role and position of the statistical office has been evaluated as 68.7% by the NSOs.

Figure 6: Users impacted by international collaborations



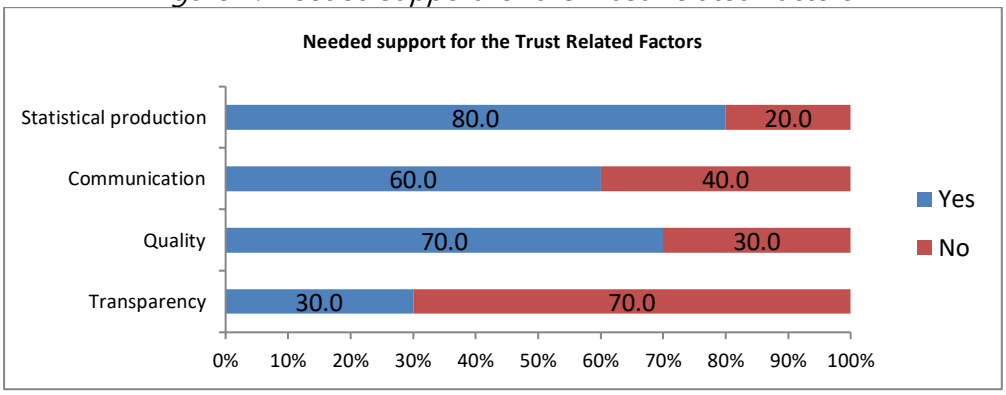
Evaluation has been given based on six different kinds of users of official statistics: General public, Governmental institutions, Local government, Media, Scientific Community and Non-profit organisations. The extent collaboration

[2] Standardisation in the European Statistical System, Meeting on the Management of Statistical Information Systems (MSIS 2013)

with international institutions improved the positioning of the NSOs regarding general public has been evaluated as 60%. Improving the position of the NSOs towards governmental institutions and local government has been evaluated respectively 50% and 60%. Media is one of the key users of official statistics because it facilitates the link between producers of statistical data and general public. Collaboration with international institutions improved the positioning of the NSOs around 90%, which is reflecting the effect of media as a major channel of transmitting official statistics. Non-profit organisations and Scientific Community have been evaluated 50% impacted by the different collaborations.

NSOs have given their opinion also on the areas where support is needed in the near future, related to the building trust initiatives. Figure 8 shows that further support is needed mostly for statistical production (80%), followed by quality and communication (respectively 70& and 60%).

Figure 7: Needed support for the Trust Related Factors



Conclusions

The results of this survey are an important tool to detect user and official opinion regarding the trust in official statistics. This information could be integrated into the planning process of official statistics in order to raise awareness regarding trust in official statistical since the views are given from 10 NSOs. Based on survey results area where NSOs need to further work concerning trust related factors are:

- Clear explanations regarding changes or revisions
- Coherent Statistics
- Measure the quality of statistics
- Increase the quality of statistical products
- Appropriate methods to inform the general public
- Consult key stakeholders regularly
- Collaborate with media to inform the general public and avoid misleading interpretation

The following actions should be undertaken by NSO in order to improve the positioning of the official statistics in society:

- Clear explanations regarding changes or revisions
- Coherent Statistics
- Measure the quality of statistics
- Increase the quality of statistical products
- Appropriate methods to inform the general public
- Produce statistics in line with national needs
- Consult key stakeholders regularly
- Collaborate with media to inform the general public and avoid misleading interpretation

References

1. Report to the OECD of the Electronic Working Group on Measuring Trust in Official Statistics, June 2011;
2. Law No.17/2018 on "Official Statistics" http://instat.gov.al/media/3972/law-no17-2018-on-official-statistics.pdf;
3. European statistics Code of Practice https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice;
4. High-Level Group for the Modernisation of Official Statistics;
5. Standardisation in the European Statistical System, Meeting on the Management of Statistical Information Systems (MSIS 2013).

Comparison of ARIMA, neural network and wavelet models for forecasting Indonesia sharia stock index

Hermansah[1,2]; Dedi Rosadi[3]; Herni Utami[3]; Abdurakhman[3]
[1] Ph. D. Student of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2] Department of Mathematics Education, Universitas Riau Kepulauan, Batam, Indonesia
[3] Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract
Forecasting is an activity that predicts what happens in the future based on the present and past values of a variable. Forecasting is a very important element, especially in planning and decision making. The method that is often used in data forecasting that may occur in the future is the Autoregressive Integrated Moving Average (ARIMA). In the case study, a comparison of forecasting models is made using the ARIMA, Neural Network and Wavelet methods. The Neural Network (NN) method used is Feed-Forward Neural Network (FFNN) or often called Back-propagation Neural Network (BPNN) and Wavelet used is the Daubechies 4 with wavelet type Maximal Overlap Discrete Wavelet Transform (MODWT). Based on the case study on the data close of Indonesia Sharia Stock Index (ISSI), the MSE value obtained of forecasting the ARIMA method is 4,846185 and MAPE is 0,011158. MSE of forecasting the NN method is 2,419994 and MAPE is 0,007553. MSE of forecasting the Wavelet method is 38,620430 and MAPE is 0,032779. Therefore, for forecasting ISSI close data it can be said that the best model is the NN model because the MSE and MAPE values obtained are smaller than the ARIMA and Wavelet models.

Keywords
Forecasting Indonesia Sharia Stock Index; ARIMA; Neural Network; Wavelet

1. Introduction

Forecasting is an activity that predicts what happens in the future based on the present and past values of a variable [1]. Forecasting is a very important element, especially in planning and decision making. The grace period between an event and the upcoming event is the main reason for forecasting and planning. In these situations forecasting is an important tool in effective and efficient planning. The choice of method in forecasting depends on several aspects of research, namely aspects of time, data patterns, types of system models observed, and the level of accuracy of forecasting. The use of these methods in forecasting must fulfill the assumptions used [2].

This study uses three forecasting methods, namely Autoregressive Integrated Moving Average (ARIMA), Neural Network (NN) and Wavelet. ARIMA often called the Box-Jenkins time series method is a method that uses

past and present values of variables to produce accurate short-term forecasting, very well used to look at past patterns and then represent future patterns for forecasting. ARIMA is a stochastic method that is very useful for generating time series data where each event is correlated [3].

Developing of next is the Neural Network model. Neural Network (NN) is an information processing system that has characteristics similar to biological neural networks [4]. The NN forecasting method for its application is the Feed-Forward Neural Network (FFNN) or often called the Back-propagation Neural Network (BPNN). BPNN is the NN method with a network that uses errors to change the value of its weights in the backward direction. To get this error, the forward propagation stage must be done first. This network has an activation function in two layers, namely the hidden layer and output layer. The procedure for establishing BPNN begins with the selection of input variables by looking at a plot of the Autocorrelation Function (ACF) or Partial Autocorrelation Function (PACF) that is significant and determines the target variable. The second stage is data sharing to training data and testing data. Next is the determination of learning parameters. The formed network is selected from the results of training and testing by looking at the smallest Mean Squared Error (MSE) and the smallest Mean Absolute Percentage Error (MAPE) [5].

In the past two decades, other techniques that are widely used are wavelets or wavelet transformations. The use of wavelet transformation as an analytical tool is caused partly because it has advantages in the process of denoising, data compression, and multiresolution [6]. In this study the wavelet method to be used is the Maximal Overlap Discrete Wavelet Transform (MODWT) with the Daubechies 4 wavelet type. MODWT is considered more suitable for time series data because in each decomposition level there are wavelet coefficients and scale coefficients as many data lengths. This advantage reduces the weakness of filtering with Discrete Wavelet Transform (DWT) which cannot be done in any sample size [7]. Determination of decomposition levels and coefficients used as input models using multi-scale decomposition. The results of MODWT will be obtained smooth coefficients and detail coefficients then smooth coefficients and detail coefficients will be used to obtain the value of Multiscale Autoregressive (MAR) that will be used for forecasting.

The purpose of this study is to compare the accuracy of forecasting using the ARIMA, NN and Wavelet models. The case study used close of Indonesia Sharia Stock Index (ISSI) data from September 4, 2017 to September 19, 2018. The selection of the right model is very necessary to predict ISSI, so that an action can be taken or a decision is made. The next three models will be used to forecast the ISSI with the smallest MSE and MAPE values showing the best performance.

2. Method

a. ARMA model

The process $\{X_t, t \in \mathbb{Z}\}$ is said to be an ARMA (p, q) process if $\{X_t\}$ is stationary and if for every t,

$$X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \cdots - \varphi_p X_{t-p} = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \quad (1)$$

where $\{e_t\}$ is a white noise with mean 0 and variance σ^2. We say that $\{X_t\}$ is an ARMA (p, q) process with mean μ if $\{X_t - \mu\}$ is an ARMA (p, q) process.

The Eq. (1) can be written symbolically in the more compact form,

$$\varphi(B)\, X_t = \theta(B)\, e_t, t \in \mathbb{Z} \quad (2)$$

where φ and θ are the pth and qth degree polynomials,

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \quad (3)$$

and

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \quad (4)$$

and B is the backward shift operator [8].

b. Neural Network model

The basic neural network learning approach works by computing the error of the output of the neural network for a given sample, propagating the error backwards through the network while updating the weight vectors in an attempt to reduce the error [9]. The algorithm consists of the following steps.

Step 1: Initialization of the network: The initial values of the weights need to be determined. A neural network is generally initialized with random weights.

Step 2: Feed Forward: Information is passed forward through the network from input to hidden and output layer via node activation functions and weights. The activation function is (usually) a sigmoidal (i.e., bounded above and below, but differentiable) function of a weighted sum of the nodes inputs.

Step 3: Error assessment: Assess whether the error is sufficiently small to satisfy requirements, or whether the number of iterations has reached a predetermined limit. If either condition is met, then the training ends. Otherwise, the iterative learning process continues.

Step 4: Propagate: The error at the output layer is used to re-modify the weights. The algorithm propagates the error backwards through the network and computes the gradient of the change in error with respect to changes in the weight values.

Step 5: Adjust: Make adjustments to the weights using the gradients of change with the goal of reducing the error. The weights and biases of each neuron are adjusted by a factor based on the derivative of the activation function, the differences between the network output and

the actual target outcome and the neuron outputs. Through this process the network "learns".

Forecasting is done based on the best model chosen by fulfilling the above stage. The general form of output from the neural network is as follows [10]:

$$y = f_2\left(w_{01} + \sum_{j=1}^{q} w_{j1}\ f_1\left(v_{0j} + \sum_{i=1}^{p} x_i\, v_{ij}\right)\right) \tag{5}$$

with the sigmoid activation function the following formula is obtained,

$$y_t = f_2\left(w_{01} + \sum_{j=1}^{q} w_{j1}\left(\frac{1-e^{-\left(v_{0j}+\sum_{i=1}^{p} x_i v_{ij}\right)}}{1+e^{-\left(v_{0j}+\sum_{i=1}^{p} x_i v_{ij}\right)}}\right)\right) \tag{6}$$

c. Wavelet model

By using the MODWT, a discrete time series $\{X_t, t = 1, 2, \dots, N\}$ can be written with the following form:

$$X_t = \tilde{S}_{J_0,t} + \sum_{j=1}^{J_0} \tilde{D}_{j,t}, t = 1, 2, \dots, N. \tag{7}$$

The first part $\tilde{S}_{J_0} = \{\tilde{S}_{J_0,t}, t = 1, 2, \dots, N\}$ which presents the tendency of series and which is characterized by slow dynamics, and the second part components $\tilde{D}_j = \{\tilde{D}_{j,t}, t = 1, 2, \dots, N\}$, $j = 1, 2, \dots, J_0$, present the local details of the series X_t and is characterized by fast dynamics especially for low levels.

To compute the predicted value $\hat{X}_{N+h}$ of X, it suffices then to do this for $\tilde{S}_{J_0}$ and the $\tilde{D}_j$, i.e, to evaluate $\hat{\tilde{S}}_{N+h}$ and $\hat{\tilde{D}}_{j,N+h}, j = 1, 2, \dots, J_0$. To do this, let us write:

$$\hat{\tilde{S}}_{J_0,N+h} = f_0\left(\tilde{S}_{J_0,N}, \tilde{S}_{J_0,N-1}, \dots, \tilde{S}_{J_0,N-p_0}\right) \tag{8}$$

and similarly

$$\hat{\tilde{D}}_{j,N+h} = f_j\left(\tilde{D}_{j,N}, \tilde{D}_{j,N-1}, \dots, \tilde{D}_{j,N-p_j}\right), j = 1, 2, \dots, J_0 \tag{9}$$

where $f_j (j = 1, 2, \dots, J_0)$ is the estimator, each estimator f_j may have its proper order p_j. The choice of $f_0, f_1, \dots, f_j$ is related to the dynamic behavior of the series to be predicted.

Based on linear ARMA model theory, the tendency and details can be approximated as following form:

$$\hat{\tilde{S}}_{J_0,N} = \varphi_1 \tilde{S}_{J_0,N-1} + \cdots + \varphi_{p_0} \tilde{S}_{J_0,N-p_0} + e_N + \theta_1 e_{N-1} + \cdots + \theta_{q_0} e_{N-q_0} \tag{10}$$

and

$$\hat{\tilde{D}}_{j,N} = \varphi_{j1} \tilde{D}_{j,N-1} + \cdots + \varphi_{jp_j} \tilde{D}_{j,N-p_j} + e_N + \theta_{j1} e_{N-1} + \cdots + \theta_{jq_j} e_{N-q_j} \tag{11}$$

By using the notation above, Eqs. (3) and (4) can be written as follows:

$$\hat{\tilde{S}}_{J_0,N} = [\varphi(B) - 1]\tilde{S}_{J_0,N} + \theta(B) e_N \tag{12}$$

and

$$\hat{\tilde{D}}_{j,N} = [\varphi_j(B) - 1]\tilde{D}_{j,N} + \theta_j(B) e_N \tag{13}$$

Hence, the MODWT-ARMA prediction model is

$$\hat{X}_{N+h} = \hat{\tilde{S}}_{J_0,N+h} + \sum_{j=0}^{J_0} \hat{\tilde{D}}_{j,N+h} \tag{14}$$

$$\hat{X}_{N+h} = [\varphi(B) - 1]\tilde{S}_{J_0,N+h} + \theta(B)e_{N+h} + \sum_{j=0}^{J_0}\left([\varphi_j(B) - 1]\tilde{D}_{j,N+h} + \theta_j(B)e_{N+h}\right) \quad (15)$$

3. Results

The research data used for modeling is the close of Indonesia Sharia Stock Index (ISSI) data. ISSI, which was launched on 12 May 2011, is a composite index of sharia shares listed on the IDX. ISSI is an indicator of the performance of the Indonesian sharia stock market. ISSI constituents are all sharia shares listed on the IDX and entered into the List of Sharia Securities issued by OJK. ISSI data is periodic data. This data is obtained from IDX, which is daily data from September 4, 2017 to September 19, 2018, with 237 data. The amount of ISSI close data is divided into training data and testing data. The training data is used for the formation of the model as many as 225 data, while the testing data of 12 data is used for checking the model. Plot of movement data from close ISSI as follows:

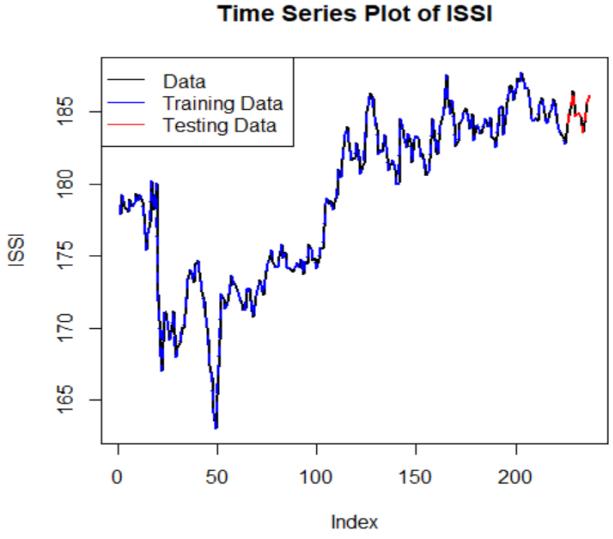


Figure 1: Plot of movement data from close ISSI

a. Modeling with ARIMA

The initial procedure for modeling using ARIMA is checking data stationarity. Using the R program, the Augmented Dickey-Fuller test statistic obtained p-value is 0,1517 greater than α used which is 0,05, it can be concluded that the data is not stationary, so it needs to be stationary both the mean and variance. By performing differencing and log transformations, the data is stationary both in the mean and variance. The best model obtained is ARIMA (3,1,0) with MSE value of 1,5395 and MAPE of 0,0051 or 0,51%. Forecasting results for the next 12 periods obtained MSE values of 4,8462 and MAPE of 0,0112 or 1,12%. Because the MAPE value is below 10%, so it can be concluded that the model has a good performance. The plot of the forecasting results and the actual data are as follows:

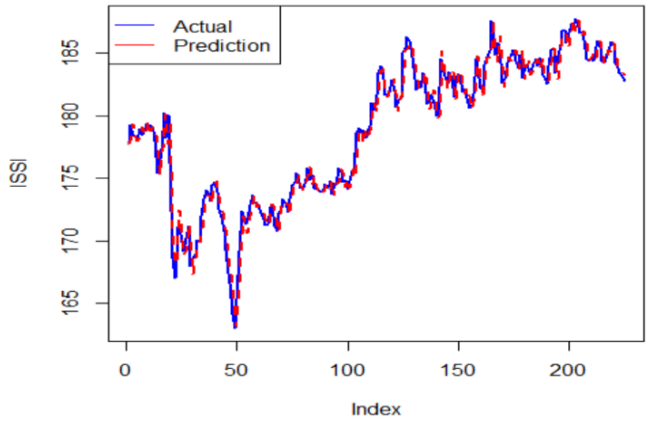


Figure 2: The plot of the actual data and forecasting results of ARIMA models

b. Modeling with Neural Network

The architecture of the Back-propagation Neural Network (BPNN) is determined by trial and error on several types of architecture. The activation function used is the bipolar sigmoid function for the input layer to the hidden layer and the linear function for the hidden layer to the output layer. The training model in BPNN was chosen, namely resilient back-propagation. The parameters chosen are based on the training model selected default with the specified value. Determining the best model is also done by considering the smallest MSE and MAPE values.

With 225 training data and 12 testing data, the best architecture of BPNN by trial and error obtained 4 input layers, 7 hidden layers and 1 output layer with MSE values of 1,2639 and MAPE of 0,0046 or 0,46%. The best model obtained is used for forecasting testing data. Obtained forecasting data with MSE of 2,3735 and MAPE of 0,0075 or 0,75%. Because the MAPE value is below 10%, so it can be concluded that the model has a good performance. Figure forecasting results with actual data are as follows:

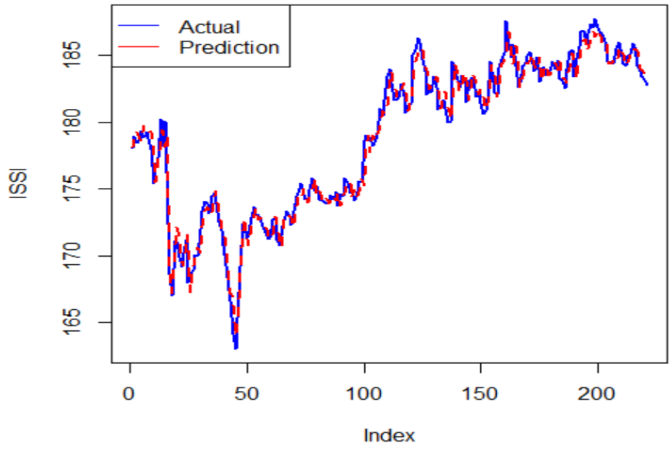


Figure 3: The plot of the actual data and forecasting results of Neural Network models

c. Modeling with Wavelet

Decomposition with MODWT is intended to stationary the detail series so that the results are stationary. In this case, the wavelet family, Daubechies 4, is used with 4 levels. Detailed and smooth values obtained are $\widetilde{D}_1$, $\widetilde{D}_2$, $\widetilde{D}_3$, $\widetilde{D}_4$, and $\widetilde{S}_4$. Furthermore, the forecasting result of close ISSI data is obtained from the sum of forecasting values of each decomposition, visually seen in Figure 4. The results of the training data forecasting obtained MSE values of 0,5743 and MAPE of 0,0032 or 0,32% and testing data were obtained MSE value is 38,6204 and MAPE is 0,0328 or 3,28%. Because the MAPE value is below 10%, so it can be concluded that the model has a good performance.

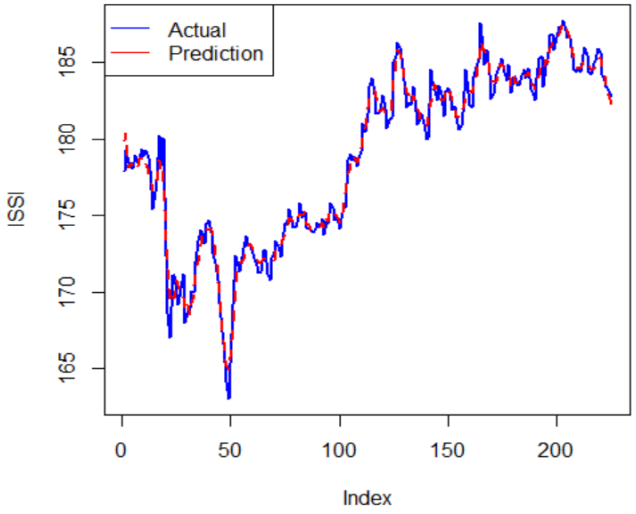


Figure 4: The plot of the actual data and forecasting results of Wavelet models

4. Conclusion

The forecasting accuracy of the ARIMA, Neural Network and Wavelet models for ISSI close data can be compared with the size of MSE and MAPE. Based on the case study of ISSI close data, the best ARIMA model forecasting results were obtained from the results of training data with MSE values of 1,5395 and MAPE of 0,0051 or 0,51%. Whereas the best ARIMA from the results of data testing obtained the MSE value of 4,8462 and MAPE of 0,0112 or 1,12%. The best Neural Network model were obtained from the results of training data with MSE values of 1,2639 and MAPE of 0,0046 or 0,46%. Whereas the best Neural Network from the results of data testing obtained the MSE value of 2,33735 and MAPE of 0,0075 or 0,75%. The Wavelet model were obtained from the results of training data with MSE values of 0,5743 and MAPE of 0,0032 or 0,32%. While the Wavelet model from the results of data

testing obtained MSE values of 38,6204 and MAPE of 0,0328 or 3,28%. Of the three models, forecasting close ISSI data can be said to be the best model is the Neural Network model because the MSE and MAPE values obtained are smaller than the ARIMA and Wavelet models.

References

1. Makridakis, S.; Wheelwright, S.C.; and Hyndman, R.J. (1997). Forecasting: Methods and applications. New York: Wiley.
2. Ord, K.; Fildes, R.; Kourentzes, N. (2017). Principles of Business Forecasting. Second Edition, Wessex Press Publishing Co., Chapter 10.
3. Rosadi, D. (2014). Analisis Runtun Waktu dan Aplikasinya dengan R. Yogyakarta: Gajah Mada University Press.
4. Fausett, L. (1994). Fundamental of Neural Network (Architectures, Algorithms, and Applications). Upper Saddle River, New Jersey: Prentice.
5. Kourentzes, N.; Barrow, B.K.; Crone, S.F. (2014). Neural network ensemble operators for time series forecasting. Expert Systems with Applications, 41(9).
6. Walker, J.S. (2008). A Primer on Wavelets and Their Scientific Applications. Second Edition, Chapman and Hall/CRC, USA.
7. Percival, D.B. and Walden A.T. (2000). Wavelet Methods for Time Series Analysis. Cambridge University Press.
8. Hyndman, R.J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 26(3).
9. Crone, S.F. and Kourentzes, N. (2010). Feature selection for time series prediction – A combined filter and wrapper approach for neural networks. Neurocomputing, 73(10).
10. Wang, X.; Smith, K.A.; and Hyndman, R.J. (2006). Characteristic based clustering for time series data. Data Mining and Knowledge Discovery, 13(3).

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

Determinants of Urban Development in Egypt

Azza Hassan
Central Agency for Public Mobilization and Statistics, Cairo, Egypt

Abstract

This study uses data of the Arab Republic of Egypt covering the eleven years 2006–2016 to examine factors that drive urban development in Egypt. Previous work has examined standard international measures of urban development such as GDP, population density, and average income. In particular, the study asks whether traditionally used economic factors (education, environment, cultural industry, infrastructure, FDI, and government spending) are most closely associated with urban development in Egypt during this time period. Linear regressions are estimated with each of the measures as the dependent variable. The results are presented and discussed with implications for studying Egyptian urban development.

Keywords

Egypt; Urban Development; Urbanization.

1. Introduction

Urbanization is defined as the transformation of the population from rural to urban areas and a gradual increase in the proportion of people living in urban areas, a major cause of urban problems. Urbanization is associated with a range of disciplines such as geography, sociology, economics, urban planning, Modernization, and industrialization. Urbanization leaves enormous social, economic and environmental changes and has the potential to provide sustainability opportunities with the ability to use resources more efficiently and provide sustainable land (McGranahan, Satterthwaite 2014).

In this unprecedented era of growing urbanization, in the context of the 2030 Sustainable Development Plan, the Paris Agreement and other agreements and global frameworks for development, we have reached a critical point in understanding that cities can be the source of solutions, rather than the challenges facing our world today. If urban planning is well planned and well managed, it can be a powerful tool for sustainable development for both developing and developed countries (Clos, 2017).

For the first time in the history of mankind, we are facing a change in the numerical ratio of the population. The proportion of the urban population of the world over the rural population as a result of recent statistics, considering that 60% of the world population will live in urban areas until 2030, Of the

total world population, while some 3.3 billion people live in urban areas. While nearly 180,000 people move daily in urban areas, 60 million people from non-developed countries move annually to urban areas (urbanization is more Developed countries). (Dociu, Dunarintu, 2012).

For the Arab Republic of Egypt, the strategy of urban development in Egyptian cities began to develop general plans for the main cities in 1958, and then urban planning began to take an advanced position in the late sixties by the establishment of Greater Cairo Planning Authority, which includes Cairo and surrounding communities. In Cairo in the identification of land use, road networks and construction systems that took eleven years 1982, and in 1996 began to think of the development of a national strategy at the national level by preparing an urban map of Egypt until 2017, and then began the General Authority for Urban Planning to develop a map of the development In the fields of agriculture, industry, water, population, transportation, transportation, tourism and mineral wealth in order to achieve the country's urban strategy and to move from the narrow valley to the new areas (Ibrahim, 2000).

2. Methodology

The analyses in this paper are based on data from the World Bank. Using the Multiple regression models. The study uses four dependent variables used to represent the development of population density and the level of urbanization to represent urbanization, per capita GDP and average annual wages to represent urban economic development. The variables representing traditional economic factors, human capital development, urban facilities, infrastructure, FDI, Governmental organizations to identify the factors most relevant to urban development in the Arab Republic of Egypt from 2006 to 2016

Table 1 Comparison of dependent variables using Pearson correlations

	Urban population	GDP	Annual Average	population density
Urban population	1			
GDP	.711*	1		
Annual Average	.674*	.975**	1	
population density	.611*	.818**	.824**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 1 shows the results of a Pearson correlation of the possible dependent variables selected for the study in order to indicate the compatibility between the selected dependent variables. All the variables are positively correlated with each other and the ratios in urban population ranged from a low level of 0.611 to a level of 0.975 there is a significant association at 0.05 level, and GDP ranged from a level of 0. 818 to a level of 0.975 there is a significant association at 0.01 level, and Annual Average is high significant in 0. 824 with population density.

With the selection of the dependent variables, the models were selected using the following formula:

$$Y = \beta0 + \beta1x1 + \beta2x2 + \beta3x3 + \beta4x4 + \beta5x5 + \beta6x6 + \varepsilon$$

Where

Y = Urban population (urban population divided by total population).

= Per capita GDP (gross domestic product divided by total population).

= Average annual pay (mean annual remuneration).

= Population density (population in units of 10,000 per square kilometer). and

β0 = Constant.

X1 = Green space (per capita public green space in square meters).

X2 = FDI (per capita foreign direct investment in $USD).

X3 = Public transportation (per capita number of public transportation vehicles).

X4 = Pollution (per capita industrial sulfur dioxide emissions).

X5 = Internet users.

X6 = Electric power generated from stations belonging to production companies.

Table 2 shows the results of a Pearson correlation of the possible independent variables selected for the study in order to indicate the compatibility between the selected independent variables. All of the variables positively correlated with all of the others at a significant level and the relationships ranged from a low of 0. 661 To a high of 0. 951. Per capita GDP appeared to have the strongest link to the other measures of urban development, with correlations of 0.951with Generated Electric Energy of Stations, 0. 941 with Green space, and 0. 935 with Public transport.

Table 2 Comparison of independent variables using Pearson correlations

	GDP	Green space	FDI	Public transport	Pollution	Internet	Generated Electric Energy Of Stations
GDP	1						
Green space	.941**	1					
FDI	.661*	.640*	1				
Public transport	.935**	.840**	.688*	1			
Pollution	.920**	.750**	0.596	.925**	1		
Internet	.856**	.87 4**	.822**	.778**	.692*	1	
Generated Electric	.951**	.883**	.622*	.891**	.930**	.764**	1

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 3 Correlations matrix between dependent variable and independent variable

Independent variables	Urban population	GDP	Annual Average	population density	urbanization level
Green space	.710*	.941**	.974**	.786**	.942**
FDI	.700*	.661*	.709*	.764**	.775**
Public transport	.678*	.935**	.895**	.762**	.903**
Pollution	0.598	.920**	.841**	.712*	.850**
Internet	0.596	.856**	.903**	.853**	.886**
Generated Electric Energy of Stations	.671*	.951**	.915**	.774**	.915**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Correlations were estimated for the four models with the significant results appearing in Table 3 The model for urban population showed significant positive results for Green space, FDI and Generated Electric Energy of Stations at the 0.001 level. On the light of the World Bank there is no significant association between the urban population and Pollution; Internet equals 895.0 and 0.596 respectively. The model with per capita GDP as the dependent variable produced significant positive results for green space and Generated Electric Energy of Stations at the 0.01 level, as well as FDI at the 0.05 level. The model for average annual pay was positive and significant for Green space, Public transport, Pollution; Internet, and Generated Electric Energy of Stations at the 0.001 level.

The signs on most of the significant variables were positive, meaning that they increased in proportion to the dependent variable. The results were generally consistent across the dependent variables, though there were some notable differences. Only Pollution and Internet failed to have a significant positive relationship with urban population.

3. Results

Table 4 Determination of the Stationary Time Series of the Ridge regression

Independent	Urbanization level		Population density		Per capita GDP		Annual Average	
	b	SEb	b	SEb	b	SEb	b	SEb
Cons	45.7408	35.53533	10859.3	7016.895	91.6157	231.7975	-132068.0000	
Green space	-58.8104	37.92653	-13089.8	7489.067	-87.1435	259.1331	0.0131	6.48000
FDI	-0.0742	0.09810	-48.6	19.371	-0.4935	1.2063	52.4720	2.94000
Public transport	0.0000	0.00000	0.0	0.001	0.0000	0.0000	0.0006	6.92000
Pollution	-3.8635	4.44730	-937.4	878.175	-4.6933	33.0808	9011.1000	7.51550
Internet	-0.2056	0.15038	19.9	29.695	0.3387	1.4948	222.3600	5.77760
Generated Electric	0.0000	0.00002	0.0	0.004	-0.0002	0.0002	-0.0039	8.24160

Table 5 Multiple regression model

Independent variable	Urban population	Population density	Per capita GDP	Annual Average
R^2	83.46	88.45	39.89	97.7
R^2 adj	50.38	65.37	80.07	94.25
R	91.35	94.52	63.16	91.35
Se .est	.447	88.34	7.69	1114.7
λ	.15	.15	.15	.15

Table 5 shows R is the multivariate between the dependent variable y and the set of independent variables. R^2 is the ratio of the explanation. Here, the independent variables explain 39% of the changes that occur in y. The rest of the changes are in other variables not taken in the study and are within the error limit. Which were selected within the character regression method to eliminate the problem of linear multiplicity between independent variables.

Figure 1 shows the Autocorrelation Function

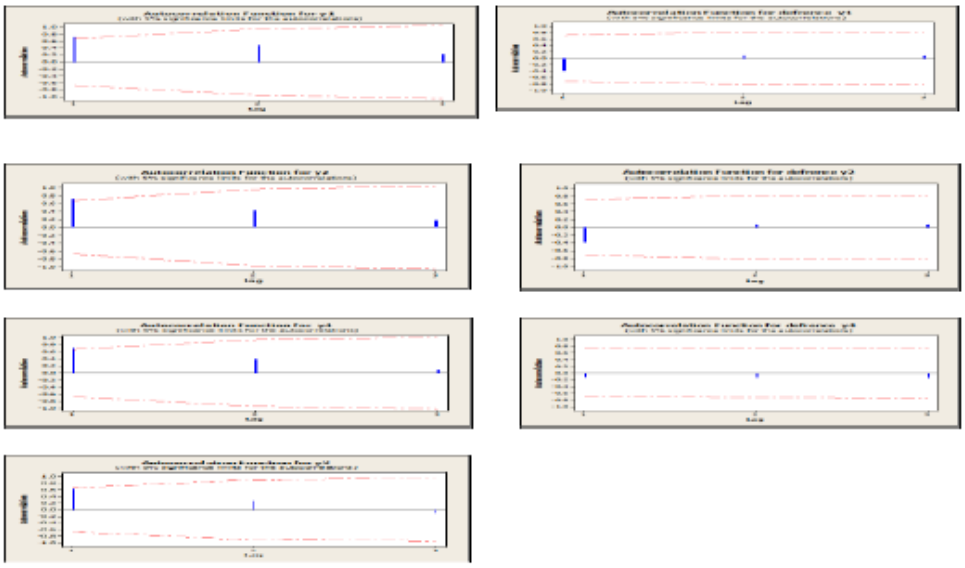


Figure 1 shows the time series of the dependent variable urban population, population density and average annual pay will not Stationary Therefore, the first difference was taken and we obtained a stable series. This can be observed from the self-correlation coefficients of the series before and after the first difference. The stability of the time series is tested using the selfcorrelation coefficient and only capita GDP is Stationary

Table 6 Ridge regression results with models for Urban population, population density, per capita GDP, average annual pay, and urbanization level as dependent variables

Independent variable	Urban population Ridge.015		Population density Ridge parameter .02		Per capita GDP Ridge.015		Annual Average Ridge parameter .015		Urbanization level Ridge parameter .015	
	Coef.	VIF	Coef.	VIF	Coef.	VIF	Coef.	VIF	Coef.	VIF
Cons	46.6036		-2070.51		-132068.		85.1107		0.00000201687	
Green space	-3.80488E-7	6.48	0.000133426	6.48	0.0130517	6.48	-8.8265E-7	6.48	0.00000201687	6.48
FDI	-0.0267094	2.94	-1.6563	2.94	52.4723	2.94	0.347612	2.94	0.076539	2.94
Public transport	-6.74976E-9	6.92	0.000038727	6.92	0.000568636	6.92	0.00000171088	6.92	2.70194E-7	6.92
Pollution	-0.0488167	7.52	228.095	7.52	9011.08	7.52	-1.63665	7.52	1.3351	7.52
Internet	0.0218004	5.78	2.76514	5.78	222.361	5.78	0.601852	5.78	-0.000397407	5.78
Generated	3.18695E-7	8.24	0.0000733826	8.24	-0.00389634	8.24	0.0000765607	8.24	0.00000304765	8.24

Table 6 shows the fitted regression model is

Urban population = 46.6036 - 3.80488E-7*Green space - 0.0267094*FDI - 6.74976E-9*Public transport - 0.0488167*Pollution + 0.0218004*Internet + 3.18695E-7*generated Electric Energy of Stations.

Population density= -2070.51 + 0.000133426*x1 - 1.6563*x2 + 0.000038727*x3 + 228.095*x4 + 2.76514*x5 + 0.0000733826*x6

Per capita GDP= -132068. + 0.0130517*x3 + 52.4723*x6 + 0.000568636*x7 + 9011.08*x11 + 222.361*x13 - 0.00389634*x14

Annual Average = 85.1107 - 8.8265E-7*x3 + 0.347612*x6 + 0.00000171088*x7 - 1.63665*x11 + 0.601852*x13 + 0.0000765607*x14

Urbanization level= -25.3153 + 0.00000201687*x3 + 0.076539*x6 + 2.70194E-7*x7 + 1.3351*x11 - 0.000397407*x13 + 0.00000304765*x14

Table 7 **Goodness of fit**

Independent variable	Urban population Ridge parameter .015	Population density Ridge.02	Per capita GDP Ridge parameter .015	Annual Average Ridge parameter .015	Urbanization level Ridge.015
R-squared	71.49%	98.61%	97.70%	77.18%	95.9%
MSE	0.00670608	124.354	1.2425E6	17.0522	0.0872278
MAE	0.0388028	5.8632	532.713	1.4204	0.151404
MAPE	0.0902767	2.74801	2.0169	1.65978	—
ME	1.03352E-14	-3.64315E-13	-2.08357E-11	-1.42109E-14	-3.21712E-15
MPE	-0.00015136	-0.283025	-0.103432	-0.088759	—
Durbin-Watson	2.22699	2.09844	3.04744	2.8601	2.16959

Table 7 shows The R-Squared statistic indicates that the model as fitted explains 71.4862% of the variability in urban population. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 28.7156%. The standard error of the estimate shows the standard deviation of the residuals to be 0.0818906. The mean absolute error (MAE) of 0.0388028 is the average value of the residuals.

The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file.

The R-Squared statistic indicates that the model as fitted explains 98.609% of the Population density. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 96.5225%. The standard error of the estimate shows the standard deviation of the residuals to be 11.1514. The mean absolute error (MAE) of 5.8632 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 97.701% of the Per capita GDP. The adjusted RSquared statistic, which is more suitable for comparing models with different numbers of ndependent variables, is 94.2526%. .The standard error of the estimate shows the standard deviation of the residuals to be be 1114.67. The mean absolute error (MAE) of 532.713 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 77.1833% of the Annual average. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 42.9583%. The standard error of the estimate shows the standard deviation of the residuals to be 4.12943. The mean absolute error (MAE) of 1.4204 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 95.9305% of the Urbanization level. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 89.8262%. . The standard error of the estimate shows the standard deviation of the residuals to be 0.295344. The mean absolute error (MAE) of 0.151404 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 95.9305% of the Urbanization level. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables is 89.8262%. The standard error of the estimate shows the standard deviation of the residuals to be 0.295344. The mean absolute error (MAE) of 0.151404 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 95.9305% of the Urbanization level. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 89.8262%. The standard error of the estimate shows the standard deviation of the residuals to be 0.295344. The mean absolute error (MAE) of 0.151404 is the average value of the residuals.

The R-Squared statistic indicates that the model as fitted 95.9305% of the Urbanization level. The adjusted R-Squared statistic, which is more suitable for comparing models with different numbers of independent variables is 89.8262%. The standard error of the estimate shows the standard deviation of the residuals to be be 0.295344. The mean absolute error (MAE) of 0.151404 is the average value of the residuals.

4. Discussion and Conclusion

The signs on most of the significant variables were positive, meaning that they increased in proportion to the dependent variable. The results were generally consistent across the dependent variables, though there were some notable differences. Finally, the urban population was the only measure have no significant association with Pollution; Internet.

References

1. Clos, Joan 2017, The New Urban Agenda, United Nations Conference on Housing and Sustainable Urban Development, Habitat III
2. Dociu, Dunarintu 2012." The Socio-Economic Impact of Urbanization" International Journal of Academic Research in Accounting, Finance and Management Sciences Volume 2.
3. Fathi, Muosakazemi, Rostami, Aliakbari 2017. "Analysis of the Citizen's Enjoyment Level of Urban, Services in Kermanshah Province, Iran" Geographical Urban Planning Research, Vol. 5, No. 3, Autumn 2017.
4. http://www.jotscroll.com/forums/3/posts/164/urbanisation-definition-causes-problems-andsolutions.html
5. Ibrahim, Abdel Baki 2000 "Urban Development Strategy in Egyptian Cities" Center for Planning and Architectural Studies.
6. Jong Youl Lee, Chad Anderson and Bo Wang 2018" Urban Development in China: Moving from Urbanization to Quality of Urban Life", Public Policy in the 'Asian Century', International Series on Public Policy.
7. Kladivo , Halás 2012 ."Quality Of Life In An Urban Environment: A Typology Of Urban Units Of Olomouc". Department of Geography, Faculty of Science, Palacký University Olomouc, Czech Republic.
8. McGranahan, Satterthwaite 2014" Urbanisation concepts and trends", Working Paper.
9. Salem, Osman 2016." A new map for urban development in Egypt, depending on mega projects of renewable energy". Conference Paper Sustainable Mega Projects Conference, At British University in Egypt, Volume: 2 https://doi.org/10.1057/978-1-137-60252-7_8

A simulation study on the score test for poison overdispersion under different forms of the variance

Johann Sebastian B. Claveria, Erniel B. Barrios, Joseph Ryan G. Lansangan
University of the Philippines, Diliman, Quezon City, Philippines

Abstract

In the analysis of count data, overdispersion happens when the response variance of the counts exceeds the response mean; and its presence in the count model often leads to underestimated standard errors and erroneous inferences. The existing method for detecting overdispersion is through the regression-based test, where the estimated response variance is modelled with a function of the mean, whose form is established prior to fitting the variance model; a score test statistic is then compared with the quantiles of a t_{n-1} distribution. However, establishing the form of the variance a priori may lead to a mis specified variance, more so that it is not directly observed; consequently, the test statistic is also affected. This study explores the properties of the existing test under different scenarios of specifying the variance of an unknown form. Bootstrap simulations are carried out on different a priori models where the variance is modelled as a constant, a linear function, a quadratic function, and a polynomial function of the mean. Simulations show that while the parametric framework of the score test for overdispersion may seem to yield an improper statistical size yet high statistical power, its nonparametric counterpart yields a better statistical size yet low statistical power. This trend is observed as the assumed form of the variance model becomes more complicated. Moreover, simulations have shown that, as the framework does not consider restrictions in the estimation procedure of the variance model, negative estimates are observed, which may then yield to negative variances.

Keywords

count data; Poisson regression; overdispersion; score test; bootstrap

1. Introduction

The *Poisson distribution* is the canonical model used in the study of count data. It arises naturally from a Poisson counting process which only allows non-negative integers to be observed in a given time *t*, thus making it a more appropriate model for count data than other real-valued distributions in the traditional framework of statistical inference. (Karlin & Taylor, 1975) The Poisson distribution can be characterized by its mass function:

$$P(Y = y|\mu, t) = \frac{e^{-\mu t}(\mu t)^y}{y!}, y = 0,1,2, \ldots \quad (1)$$

The parameter μ is often called the *rate* or *intensity* parameter which is taken to be positive, while t is the measure of space or time where the event of interest is observed. By modelling the rate at how the counts are generated as a function of a set of covariates $\underline{X} = [X_1 \quad X_2 \quad \ldots \quad X_p]'$, the *Poisson regression model* arises, which is immediately derived from the Poisson distribution by conditioning the counts Y_i on the rate parameter as a linear combination of the covariates $\underline{X_i}$ (Cameron & Trivedi, 1998), i.e.:

$$P\left(Y_i = y_i \middle| \underline{X_i}, \underline{\beta}\right) = \frac{e^{-\exp\left\{\underline{X_i'}\underline{\beta}\right\}}\left(\exp\left\{\underline{X_i'}\underline{\beta}\right\}\right)^{y_i}}{y_i!}, y = 0,1,2, \ldots; \underline{\beta} \in \mathbb{R}^p \quad (2)$$

As with the distribution, the Poisson regression model assumes *equidispersion*, where the response variance is equal to its mean. *Underdispersion* happens when the mean exceeds the variance, while *overdispersion* happens when the variance exceeds the mean. *True overdispersion* occurs when the excess variation among the counts is attributed to a non-Poisson data-generating process (DGP), while *apparent overdispersion* occurs as a consequence of a misspecified count model.

The violation of the equidispersion property immediately manifests in the inflated fit statistics, and it can underestimate the standard errors, thereby invalidating the inference coming from the count model. However, while the fit statistics can indicate the presence of overdispersion, the magnitude at which the data is deemed to suffer from overdispersion is relative. (Hilbe, 2011) Formal tests for detecting overdispersion have been developed by shifting the focus to different count distributions or variance specifications. When equidispersion is violated, the response variance can be hypothesized to be some function of the response mean, i.e.:

$$\begin{aligned} Var Y_i|X_i &= E(Y_i - EY_i|X_i)^2|X_i \\ &= EY_i|X_i + h(EY_i|X_i) \\ &= \mu_i + h(\mu_i) \end{aligned} \quad (3)$$

With this form, the variance can be modelled via regression. In practice, $h(\mu i)$ is assumed to have a closed form – at most, an algebraic expression, say, $h(\mu i)=\alpha g(\mu i)$ – and a *regression-based test for overdispersion* can be carried out by testing the null hypothesis H_0:α=0 i.e. equidispersion, using the test statistic (Cameron & Trivedi, 1990):

$$T = \frac{\sum_{i=1}^{n} \frac{g(\hat{\mu}_i)}{2\hat{\mu}_i}((Y_i - \hat{\mu}_i)^2 - Y_i)}{\sqrt{\sum_{i=1}^{n} \frac{g^2(\hat{\mu}_i)}{2\hat{\mu}_i^2}}} \sim t_{n-1} \quad (4)$$

An issue with the regression-based test is how the extra-Poisson variation will be modelled.An overdispersed count distribution is often assumed prior to formally testing for overdispersion because the variance is not directly observed. This makes the variance model in (3) prone to misspecification because $g(\mu i)$ can be made arbitrary for as long as the variance remains positive – a misspecified variance model also has negative effects on the Poisson model regression model where, similar to the consequences of overdispersion, the standard errors might also be incorrect, leading to erroneous inferences.

2. Methodology

Primarily, the variance can be misspecified in the model; consequently, the test statistic may also be affected, as well as the testing procedure itself. Moreover, there are multiple scenarios leading to misspecification of the variance. Thus, to answer the research questions, given these scenarios, this research was designed as a simulation study on the current methodology of Cameron and Trivedi's test for overdispersion (1990) by:

1. Simulating a covariate $X_i \sim N(0,1)$ and computing $\mu_i = \exp\left\{2 + \frac{1}{2}X_i\right\}$
2. Using μ_i, Y_i is simulated from a specified count distribution
3. Fitting a Poisson regression model on (X_i, Y_i) and computing the estimated counts $\hat{\mu}_i$
4. Using $\hat{\mu}_i$, the variance is modeled using $(Y_i - \hat{\mu}_i)^2 - Y_i = g(\hat{\mu}_i) + \varepsilon_i$

Four forms of the variance models will be used:

- *Constant:* $g(\hat{\mu}_i)=\alpha$
- *Linear*: $g(\hat{\mu}_i)=\alpha\hat{\mu}_i$
- *Quadratic*: $g(\hat{\mu}_i)=\alpha\hat{\mu}_i^2$
- *Polynomial*: $g(\hat{\mu}_i) =\alpha\hat{\mu}_i^p$

To facilitate the simulation of overdispersed counts whose variance follows the form in Equation (2), the *Functional Negative Binomial (NB-F) distribution*, proposed by Claveria (2016) as a generalization of the *Polynomial Negative Binomial (NB-P)* model (Greene, 2008) , was used for simulating Y_i:

$$P\left(Y_i = y_i|\underline{X_i}, \underline{\beta}\right) = \frac{\Gamma\left(y_i + \frac{\mu_i^2}{g(\mu_i)}\right)}{y_i!\,\Gamma\left(\frac{\mu_i^2}{g(\mu_i)}\right)}\left(\frac{\mu_i}{\mu_i + g(\mu_i)}\right)^{\frac{\mu_i^2}{g(\mu_i)}}\left(\frac{g(\mu_i)}{\mu_i + g(\mu_i)}\right)^{y_i} \quad (5)$$

$$\mu_i = \exp\left\{\underline{X_i'}\underline{\beta}\right\}$$

For the different scenarios of the variance, the following forms of $g(\mu i$) are used:

- *Equidispersed*: $g(\hat{\mu}_i)$ = 0
- *Weakly overdispersed*: $g(\hat{\mu}_i)$ = 0.25 + 0.2$\hat{\mu}_i^{0.5}$
- *Strongly overdispersed:*$g(\hat{\mu}_i)$= 2.75 + 2.75 $\hat{\mu}_i^{2.5}$
- *Transcendental overdispersion*: $g(\hat{\mu}_i)$ = ln(1 + $\hat{\mu}_i$)

The parametric p-values will be computed from the null distribution of (4), while the nonparametric p-values will be computed based on the bootstrap distribution of (4). For the bootstrap p-value, the following formula is used (Davison & Hinkley, 1997):

$$p = \frac{\#(T^* \geq T)}{B+1} \tag{6}$$

From the estimated p-values, the size of the test will be computed as the proportion of ejections under a true null hypothesis, i.e. when there is truly no overdispersion. Similarly, the power of the test will be computed as the proportion of rejections under a false null hypothesis, i.e. when overdispersion is truly present.

It must also be noted that, by the original method of Cameron and Trivedi (1990), no restrictions on α is made, so that there is a non-zero probability that a negative coefficient may be estimated, so that as a consequence, a negative variance might be attained. For the final part of the simulations, the proportion of negative coefficients will be recorded and compared with the computed size and power as an added measure of reliability of the test.

3. Result

The first table presents the parametric and nonparametric size of the score test for overdispersion, i.e. the proportion of rejections when the null hypothesis of no overdispersion is true:

α	Constant		Linear		Quadratic		Polynomial	
	t	Bootstrap	t	Bootstrap	t	Bootstrap	t	Bootstrap
10%	28%	0%	27%	13%	26%	29%	27%	24%
5%	26%	0%	26%	9%	25%	16%	25%	18%
1%	21%	0%	24%	0%	23%	10%	23%	8%

Table 1: Size of the test

From the table, it can be seen that the parametric t-test for overdispersion tends to reject $H0$ more than the reference level – thus, the test is not properly sized. However, if the bootstrap distribution were to be used instead, the size tends to be smaller than the former, and increases as the assumed model of the variance becomes more complicated. The next table presents the parametric and nonparametric power of the test, i.e. the proportion of rejections when the null hypothesis of no overdispersion is false, under the weak form:

α	Constant		Linear		Quadratic		Polynomial	
	t	Bootstrap	t	Bootstrap	t	Bootstrap	t	Bootstrap
10%	43%	0%	41%	12%	35%	29%	39%	26%
5%	41%	0%	39%	9%	33%	16%	38%	18%
1%	36%	0%	36%	0%	31%	9%	35%	8%

Table 2. Power of the test under a weak form of overdispersion

Here, the parametric test tends to reject H_0 less than half of the time – possibly because of the weak form of the true variance which does not deviate significantly from the variance function of a Poisson distribution. However, the proportion of rejections revealed by using the bootstrap distribution shows that the score test yields a very low power, which increases with the complexity of the assumed model of the variance; but still yields to low statistical power. The next table presents the power of the test under the strong form:

α	Constant		Linear		Quadratic		Polynomial	
	t	Bootstrap	t	Bootstrap	t	Bootstrap	t	Bootstrap
10%	100%	2%	100%	8%	100%	23%	100%	16%
5%	100%	1%	100%	4%	100%	17%	100%	9%
1%	100%	0%	100%	0%	100%	3%	100%	1%

Table 3. Power of the test under a strong form of overdispersion

When the true form of overdispersion is strong, the parametric test tends to almost always reject the false $H0$. However, its bootstrap counterpart reveals that its actual power is very low, which, again, increases with the complexity of the assumed model of the variance, yielding to low statistical power as well. Finally, the last table presents the power of the test under the transcendental form:

α	Constant		Linear		Quadratic		Polynomial	
	t	Bootstrap	t	Bootstrap	t	Bootstrap	t	Bootstrap
10%	60%	1%	56%	14%	46%	31%	54%	25%
5%	58%	0%	54%	10%	45%	18%	52%	18%
1%	54%	0%	51%	0%	43%	11%	49%	8%

Table 4. Power of the test under a transcendental from a overdispersion

In this case, the parametric test tends to reject the false $H0$ only around half of the time, while its bootstrap counterpart remains to suffer from a very low power, which improves with the complexity of the assumed model of the variance. It should be noted, however, that despite the increasing power of the tests along with the complexity of the assumed form of the variance, it may

not be able to reach a very high statistical power as the true form of the variance of the count data is, in fact, a transcendental function, in which a finite combination of polynomials can only give an approximation to an extent of error.

As stated in the previous section, the following table presents the proportion of samples where the estimated coefficient α is negative, which may result in a negative variance of the count data:

Form	Constant	Linear	Quadratic	Polynomial
Equidispersed	64%	65%	69%	69%
Weak	53%	53%	58%	58%
Strong	0%	0%	0%	0%
Transcendental	30%	32%	40%	40%

Table 5. Proportion of negative estimates

From the table, for the forms of weak to no overdispersion, the four assumed variance models estimate a negative coefficient more than half of the time, which diminishes as the true form of overdispersion becomes stronger. For the transcendental form, however, a negative estimate happens only a third of the time. Nonetheless, because the true form of overdispersion is not made known to the researcher, the current method of modelling and testing the variance of count data can be said to be prone to producing an erroneous, negative variance. This has also been observed in the two pioneering studies of Cameron and Trivedi (1986) (1998).

4. Discussion and Conclusion

The assumption of a certain form of the variance and testing its significance using the parametric t-distribution is the commonly used method of testing for overdispersion in Poisson regression. However, because of the dissimilarities of the size and power of the test under the parametric and nonparametric paradigms, it is indicative that the t-distribution may not be the true distribution of the score test statistic, and thus, is an inappropriate choice when a regression-based test for overdispersion is to be executed.

Moreover, assuming the form of the variance is very prone to model misspecification and the poor performance of the model: using less parameters in the variance model will result in a parsimonious model which gives a good statistical size when the appropriate distribution is used but fails to capture the extra-Poisson variation when the null hypothesis is false, thus resulting in low statistical power of the test. On the other hand, adding more parameters to the variance model will results in a model which may be able to capture more of the variation in the data, but fails to be simplified when the

null hypothesis is true, thus resulting in a poorly-sized test. With the advancements in research and computing power, one approach to balance the shortcomings of these closed-form variance models is to let go of the parameters and consider the nonparametric estimation of the variance. This approach can be seen in Claveria (2016).

Finally, an improvement to the shortcomings of the current practice of the regression-based test for overdispersion is to seek out improvements in the test statistic itself. Such is the study of Dean (1992), Baksh et al. (2011), and Novo and Manteiga (2000).

References

1. Baksh, F., Böhning, D., & Lerdsuwansri, R. (2011). An extension of an over-dispersion test for count data. Computational Statistics and Data Analysis 55, 466-474.
2. Cameron, A., & Trivedi, P. K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. Journal of Applied Econometrics 1, 29-53.
3. Cameron, A., & Trivedi, P. K. (1990). Regression-Based Test for Overdispersion in the Poisson Model. Journal of Econometrics 34, 347-364.
4. Cameron, A., & Trivedi, P. K. (1998). Regression Analysis of Count Data. New York: Cambridge University Press.
5. Casella, G., & Berger, R. L. (2002). Statistical Inference, 2nd ed. Duxbury.
6. Claveria, J. B. (2016). A Nonparametric Regression-Based Test for Poisson Overdispersion. Quezon City: University of the Philippines.
7. Davison, A., & Hinkley, D. V. (1997). Bootstrap Methods and their Application. Cambridge University Press.
8. Dean, C. B. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. Journal of the American Statistical Association 87, 451-457.
9. Greene, W. (2008). Functional forms for the negative binomial model for count data. Economics Letters 99, 585-590.
10. Karlin, S., & Taylor, H. M. (1975). A First Course in Stochastic Processes, 2nd ed. New York: Academic Press, Inc.
11. Ramil Novo, L., & Gonzalez Manteiga, W. (2000). F Tests and Regression Analysis Based on Smoothing Spline Estimators. Statistica Sinica 10, 819-837.
12. Wasserman, L. (2006). All of Nonparametric Statistics. Springer: New York.

Automatic assignment of underlying cause of death based on verbal autopsy instrument

Lucia Pereira Barroso, Carmen Saldiva, Paulo Saldiva
University of São Paulo

Abstract

Many of the deaths in low or middle income countries occur without medical attention, especially in rural areas. Verbal Autopsy (VA) are being used to provide information on cause of death. The verbal autopsy questionnaire has a reduced version, which is being validated in Brazil for 22 causes of death in adults. This questionnaire has been applied previously in other countries in a visit to the family of the deceased few months after the death. In Brazil, the application is being made at the Death Verification Service of São Paulo (SVOC), shortly after the death of the individual. The attribution of the underlying cause of death can be made based on the answers given by the family, using an automatic method based on the Tariff Score, applied in SmartVA software, and using weights obtained in a study previously developed in other countries. The Tariff Method was developed by the Population Health Metrics Research Consortium (PHMRC). The SmartVA software was developed by the Institute for Health Metrics and Evaluation. The objectives of this study are to present the automatic method of assigning the underlying cause of death based on a verbal autopsy instrument and to show preliminary results.

Keywords

Verbal autopsy; Tariff Method; SmartVA; cause of death

Introduction

Many of the deaths in low- or middle-income countries occur without medical attention, especially in rural areas. Verbal Autopsy (VA) are being used to provide information on cause of death. The verbal autopsy questionnaire has a reduced version, which is being validated in Brazil for 22 causes of death in adults. This questionnaire has been applied previously in other countries in a visit to the family of the deceased few months after the death. In Brazil, the application is being made at the Death Verification Service of São Paulo (SVOC), shortly after the death of the individual.

The attribution of the underlying cause of death can be made based on the answers given by the family, using an automatic method based on the Tariff Score, applied in SmartVA software, and using weights obtained in a study previously developed in other countries. The Tariff Method was

developed by the Population Health Metrics Research Consortium (PHMRC), (see Murray et al., 2011). The SmartVA software was developed by the Institute for Health Metrics and Evaluation (see Serina et al., 2015).

The objectives of this study are to present the automatic method of assigning the underlying cause of death based on a verbal autopsy instrument and to show preliminary results.

Tariff Score

The Tariff-score was calculated based on information collected from 12,501 individuals in the University of Washington's Population Health Metrics Research Consortium (PHMRC) in 6 cities in 4 countries - India (Andhra Pradesh and Uttar Pradesh), the Philippines (Bohol) , Mexico (Mexico) and Tanzania (Dar es Salaam and Pemba Island).

The questionnaire contains open and closed questions about Closed questions:

- ✓ Symptoms of terminal illness;
- ✓ Diagnosis of chronic diseases - records of previous health services;
- ✓ Risk behavior (alcohol, tobacco);
- ✓ Details of any interaction with health services.

Open questions: open narrative by the relatives through which one can identify words or groups of words associated with the cause of death. The underlying cause of death was known based on medical records and the authopsy made in the SVOC (gold standard). The input worksheet contains values 0 and 1, being 1 when the observed symptom occurred for a specific individual. With this, it is possible to calculate the endorsement rate, given by:

x_{ij} = fraction of VAs for which there is a positive response to deaths from cause i for item j

The Tariff is the endorsement rate standardized by the median and interquartile range of all causes of death, fixed a symptom. For cause *i*, symptom *j* is given by

$$Tariff_{ij} = \frac{x_{ij} - median(x_{ij})}{IIQ(x_{ij})}$$

where median(x_{ij}) is the median fraction with a positive response for item j across all causes, and IIQ(x_{ij}) is the interquartile range of positive response rates averaged across causes. Note that Tariff can assume positive or negative values. As a final step, Tariffs are rounded to the nearest 0.5 to avoid overfitting and to improve predictive validity.

For the calculation of the Tariff-Score of case *k*, cause *i*, the 40 symptoms with the highest absolute value of the Tariff are considered

$$TS_{ki} = \sum_{r=1}^{40} Tariff_{ir} I_{kr}$$

where I_{kr} is the response for death *k* on item *j*, taking on a value of 1 when the response is positive and 0 when the response is negative.

Assignment of underlying cause of death

The assignment of the underlying cause of death to a new individual (k) is made on the basis of the Tariff-ordered score for each cause. The simplest idea would be to assign the cause for which the Tariff-Score was the largest. However, some causes have a naturally higher score and then classification is based on ranks. The Tariff Score of individual k is calculated for each of the possible causes.

The original sample is sorted from lowest to highest value. The Tariff-score of the new individual is included in this order for each possible cause. The new individual is assigned the cause that presents the highest rank.

For this analysis we consider the chapter classification of ICD10 - International Statistical Classification of Diseases (World Health Organization - WHO). Table 1 shows the counts of cases classified according to the gold standard and by the Tariff method.

Table 1: Classification: Gold Standard (row) *versus* Tariff Method (column)

Cause	BC	AIDS	Stroke	DB	IHD	CRD	OD	Total
BC	10	0	0	0	0	0	4	14
AIDS	1	7	0	1	0	0	1	10
Stroke	3	1	44	5	7	3	13	76
DB	0	1	18	30	17	1	11	78
IHD	1	0	98	44	150	20	101	414
CRD	0	0	16	5	10	20	20	71
OD	10	3	90	63	125	12	296	599
Total	25	12	266	148	309	56	446	1262

BC = Breast Cancer
DB = Diabetes
IHD = Ischemic Heart Disease
CRD = Chronic Respiratory Diseases
OD = Other Diseases

Conclusion

- ✓ The most common underlying cause is Ischemic Heart Disease, totaling 414 cases out of 1262 observed, that represents, 32.8%.
- ✓ Considering the 414 cases of Ischemic Heart Disease, 150 were classified correctly by the Tariff method, which means 36.20% accuracy.
- ✓ Considering 633 cases (excluding Other Diseases), 261 cases were correctly classified, 39.4%.

- ✓ The percentages of correct classification are: Breast Cancer (71.4%), AIDS (70.0%), Stroke (57.9%), Diabetes (38.5%), Ischemic Heart Disease (36.2%), Chronic Respiratory Diseases (28.2%).

References

1. Serina P, Riley I, Stewart A, James SL, et al. Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. BMC Medicine. 2015 Dec 8; 13:291.
2. Murray CJL, Lopez AD, Black R, Ahuja, et al. Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. Population Health Metrics. 2011 Aug 4; 9:27.

Acknowledgment

This study is supported by grant #2013/21728-2, São Paulo Research Foundation (FAPESP) and grant 815781/2014, Ministry of Health Brazil.

1) University of São Paulo
2) Sírio-Libanês Hospital
3) Ministry of Health Brazil
4) Secretary of Health, São Paulo
5) Federal University of Minas Gerais
6) Federal University of Rio Grande do Norte

Measuring Real GDP and Changes in the Terms of Trade across Space and Time

Yan Wang
Dongbei University of Finance and Economics

Abstract

Larger international transactions and sharp changes in relative prices have great effect on the estimates of national income and product. If we define the terms of trade as the ratio of export price to import price, an improvement in the terms of trade means that the country can get more for less. This phenomenon is similar to a technological progress (Diewert and Morrison, 1986). Contrary to a technological progress, however, a change in the terms of trade is treated by the national accounts as a price phenomenon, rather than as a real effect. Consequently, the beneficial effect of an improvement in the terms of trade is not taken into account by real GDP (Kohli, 2006). In this paper we are trying to measure the real GDP on both output-side and expenditure-side and the contribution of terms of trade to the real GDP growth across space and time in a more consistent way. Based on production theory and translog GDP methodology introduced by Diewert and Morrison (1986), we further extend Inklaar and Diewert (2016)'s framework to simultaneously calculate real GDP on expenditure-side and output-side and isolate the effect of terms of trade on real GDP.

Keywords

Real GDP; Purchasing Power Parities (PPPs); Terms of trade; Production theory

1. Introduction

Larger international transactions and sharp changes in relative prices have great effect on the estimates of national income and product. These estimates have been more sensitive to the choice of concepts and methodologies since 1970's (Denison, 1981). The economic performance of Switzerland over the long run is paradoxical. In most international comparisons, Switzerland is found to have a growth rate that is significantly lower than that of other industrialized nations. However, in terms of average living standards, Switzerland always ranks among the top nations (Kohli, 2004). If we define the terms of trade as the ratio of export price to import price, an improvement in the terms of trade means that the country can get more for less. This phenomenon is similar to a technological progress (Diewert and Morrison, 1986). Contrary to a technological progress, however, a change in the terms

of trade is treated by the national accounts as a price phenomenon, rather than as a real effect. Consequently, the beneficial effect of an improvement in the terms of trade is not taken into account by real GDP (Kohli, 2006).

SNA (1993) first introduce the concept of real GDP based on production approach and expenditure approach to take into account the terms of trade. There are two main approaches to measuring the Real GDP on output-side and expenditure-side and the terms of trade. The first approach is to focus on temporal changes in the terms of trade from the perspective of national accounts. However, it is difficult to measure the total effect made by the international trade and the consequent price movement on production and consumption based only on an economy's national accounting data. The second approach is to investigate the changes in the terms of trade across time and space from the perspective of international comparison. Latest version of PWT uses two sets of indexes called CGDP and RGDP to measure the real GDP on output-side and expenditure-side, but there are still questions to ask: (1) The implicit PPP based on Fisher quantity index may not be transitive. (2) We can't decompose the Fisher quantity index into meaningful parts to directly measure the effect of terms of trade. In this paper we are trying to measure the real GDP on both output-side and expenditure-side and the contribution of terms of trade to the real GDP growth across space and time in a more consistent way. Based on production theory and translog GDP methodology introduced by Diewert and Morrison (1986), we further extend Inklaar and Diewert (2016)'s framework to simultaneously calculate real GDP on expenditure-side and output-side and isolate the effect of terms of trade on real GDP.

2. Literature Review

2.1 National practice

Since 1981, the U.S. Bureau of Economic Analysis publishes series of what has become known as ''Command-Basis'' GNP. Command-Basis GNP (GDP) is a measure of real GNP (GDP) that tries to take into account the effects of changes in the terms of trade on the purchasing power of a nation.

BEA use import price index to deflate the trade account. Canada and Swiss National Bank use Gross Domestic Final Expenditure (GDFE) to deflate the trade account.

2.2 The SNA approach

SNA (1993) first introduce the concept of real GDP based on production approach and expenditure approach to take into account the terms of trade. According to SNA, the production approach measures the real GDP in a way to measure an economy's production capability and its production possibility frontier; The expenditure approach measures the real GDP in a way to measure an economy's real purchasing power and living standard (SNA, 1993).

The latest SNA(2008) confirms the importance of these two approaches:

$$GDP_t^o = \frac{C_t + I_t + G_t}{P_t} + \frac{X_t}{P_t^x} - \frac{M_t}{P_t^m} \quad (1)$$

$$GDP_t^e = \frac{C_t + I_t + G_t}{P_t} + \frac{X_t - M_t}{P_t^*} \quad (2)$$

As SNA (2008) stated, there is no consensus on the choice of P*. Kohli (2004) suggested that domestic price should be used to deflate the trade account, which enable us to use a superlative index formula to capture the exact contribution of real forces including terms of trade. Reinsdorf (2010) indicated that marginal income arising from trading gains is spent in the same way as average income, so it is appropriate to use GDFE to eliminate a trade imbalance.

2.3 Measuring the terms of trade

There are two main approaches to measuring the Real GDP on output-side and expenditure-side and the terms of trade. **The first approach** is focusing on temporal changes in the terms of trade from the perspective of national accounts. In the pioneer work by Diewert and Morrison (1986), production theory and translog GDP methodology are employed to measure the changes in the terms of trade. The measurement work can be reduced to Tornqvist index. Based on this framework, many authors have attempted to measure the terms of trade for one country over years (Morrison and Diewert, 1990; Kohli, 2004, 2006, 2007; Diewert, 2008; Reinsdorf, 2010; Diewert and Yu, 2012). **The second approach** is to investigate the changes in the terms of trade across time and space from the perspective of international comparison. The traditional Gary-Khamis system has been modified to include differences in the terms of trade between countries, which enable us to measure the real GDP from both the expenditure-side and the output-side (Feenstra et al., 2009). Penn World Table incorporates these technics to provide RGDPe and RGDPo started from version 8.0 (Feenstra et al, 2015).

2.4 Comments on the literature

It is difficult to measure the total effect made by the international trade and the consequent price movement on production and consumption based only on an economy's national accounting data.

Latest version of PWT uses two sets of indexes called CGDP and RGDP to measure the real GDP on output-side and expenditur-side, but there are still questions to ask: (1) The implicit PPP based on Fisher quantity index may not be transitive. (2) We can't decompose the Fisher quantity index into meaningful parts to directly measure the effect of terms of trade.

Our questions are as follows: Can we measure the real GDP on the output-side and expenditure-side across space and time in a more consistent way? Can we measure the contribution of terms of trade to the real GDP growth across space and time in a more consistent way? Based on production theory

and translog GDP methodology introduced by Diewert and Morrison (1986), Inklaar and Diewert (2016) proposed a new method for simultaneously comparing real value added and input quantity, which can be used to compare industry productivity across countries and over time. We further extend Inklaar and Diewert (2016)'s framework to simultaneously calculate real GDP on expenditure-side and output-side and isolate the effect of terms of trade on real GDP.

3. Measurement Framework

3.1 Production theory

In the modern international trade, the majority of trade consists of raw materials and intermediate goods. So-called finished imports can not be treated as final products, because they are not ready to meet final demand. These kinds of imports must still go through unloading, transporting, repackaging, wholesaling, and retailing in the domestic countries. During this process, domestic factor services are added into these products, so that a significant proportion of their final price tag is generally accounted for by domestic activities (Kohli, 2004).

3.2 Translog GDP methodology

Given a vector of output prices P and a vector of available primary inputs x, GDP function can be defined as:

$$g^t(P,x) \equiv \max_y \left\{ P \cdot y \ (y,x) \text{ belong to } \mathrm{S}^t \right\} \quad \text{Where } P = (P^D, P^X, P^M) \qquad (3)$$

Real income generated can be measured by using real output price vector p: $Y_t = g^t(p,x)$ Where $p = (\frac{P^D}{P^*}, \frac{P^X}{P^*}, \frac{P^M}{P^*})$ (4)

Diewert (2008) indicated that if the GDP function $g^t(p,x)$ has the following translog form, we can decompose the change in real income into meaningful parts by index number.

$$Ing^t(p,x) = a_t^0 + \sum_i a_t^i \ln p_t^i + (1/2)\sum_i \sum_j a^{ij} \ln p_t^i \ln p_t^j + \sum_n b_t^n \ln x_t^n + (1/2)\sum_n \sum_m b_{nm} \ln x_t^n \ln x_t^m + \sum_i \sum_n c^{in} \ln p_t^i \ln x_t^n \quad \text{where} \quad i, j \in (D, X, M) \qquad (5)$$

3.3 Decomposition method by Diewert (2008)

Diewert (2008) showed that when using production theory and translog GDP methodology, the change in real income can be decomposed into :

$$\underbrace{Y_t / Y_{t-1}}_{\text{Real income growth}} = \underbrace{\alpha_t^D}_{\text{domestic price contribution factor}} \underbrace{\alpha_t^X \alpha_t^M}_{\text{Terms of trade contribution factor}} \underbrace{[Z_t / Z_{t-1}]}_{\text{real output growth}} \qquad (6)$$

1. domestic real price contribution factor
2. terms of trade contribution factor
3. real output growth

Diewert (2008) further pointed out that if we use GDFE as deflator to calculate real income, the first component can be eliminated from the decomposition.

3.4 PPPs for output-side

Following Inklaar and Diewert (2016), the aggregate price of real value added in country k in period t relative to the aggregate price of real value added in country j in period s can be measured by Törnqvist–Theil output price index:

$$P_{kt/js} = \exp\left[\sum_i \pm\frac{1}{2}(\mathrm{s}^i_{js}+\mathrm{s}^i_{kt})\ln(p^i_{kt}/p^i_{js})\right], \mathrm{i}\in\{D,X,M\} \tag{7}$$

where the value-added output shares s^i_{kt} defined as:

$$\mathrm{s}^i_{kt} = \frac{v^i_{kt}}{v_{kt}},\ \mathrm{i}\in\{D,X,M\} \quad \text{where } v_{kt} = v^D_{kt}+v^X_{kt}-v^M_{kt} \tag{8}$$

To get a base invariant PPPs, we follow the CCD strategy to get the transitive PPPs:

$$P_{kt^*} = \left[\prod_{j=1}^{K}\prod_{s=1}^{T} P_{kt/js}\right]^{\frac{1}{KT}} \tag{9}$$

By rearrangement, we can get a much simple form:

$$\ln P_{kt^*} = \ln P_{kt^{**}} + \alpha$$

$$\ln P_{kt^{**}} = \frac{1}{2}(s_{..D}+s^D_{kt})\ln(p^D_{kt}/p_{..D}) + \frac{1}{2}(s_{..X}+s^X_{kt})\ln(p^X_{kt}/p_{..X}) - \frac{1}{2}(s_{..M}+s^M_{kt})\ln(p^M_{kt}/p_{..M}) \tag{10}$$

Set first country and period 1 as base:

$$P_{kt} = P_{kt^*}/P_{11^*} = P_{kt^{**}}/P_{11^{**}} \tag{11}$$

3.5 PPPs for expenditure-side

The aggregate price of domestic absorption in country k in period t relative to the aggregate price of domestic absorption in country j in period s can be measured by the following price index:

$$P^D_{kt/js} = \frac{p^D_{kt}}{p^D_{js}} \tag{12}$$

To get a base invariant PPPs, we follow the CCD strategy to get the transitive PPPs:

$$P_{kt*}^{D}=\left[\prod_{j=1}^{K}\prod_{s=1}^{T}P_{kt/js}^{D}\right]^{\frac{1}{KT}} \tag{13}$$

By rearrangement, we can get a much simple form:

$$\ln P_{kt*}^{D}=\ln P_{kt*}^{D}-\ln P_{..D} \tag{14}$$

Set the first country and period 1 as base:

$$P_{kt}^{D}=P_{kt*}^{D}/P_{11*}^{D} \tag{15}$$

3.6 Terms of trade contribution factor

Diewert (2008) show that terms of trade contribution factor can be defined as follows:

$$P_{kt/js}^{XM}=\exp\left[\frac{1}{2}(\mathrm{s}_{js}^{X}+\mathrm{s}_{kt}^{X})\ln(\frac{p_{kt}^{X}/p_{kt}^{D}}{p_{js}^{X}/p_{js}^{D}})-\frac{1}{2}(\mathrm{s}_{js}^{M}+\mathrm{s}_{kt}^{M})\ln(\frac{p_{kt}^{M}/p_{kt}^{D}}{p_{js}^{M}/p_{js}^{D}})\right] \tag{16}$$

To get a base invariant PPPs, we follow the CCD strategy to get the transitive PPPs:

$$P_{kt*}^{XM}=\left[\prod_{j=1}^{K}\prod_{s=1}^{T}P_{kt/js}^{XM}\right]^{\frac{1}{KT}} \tag{17}$$

By rearrangement, we can get a much simple form:

$$\ln P_{kt*}^{XM}=\ln P_{kt**}^{XM}+\alpha$$

$$\begin{aligned}\ln P_{kt**}^{XM}=&\left[\frac{1}{2}(s_{..X}+s_{kt}^{X})\ln(p_{kt}^{X}/p_{..X})-\frac{1}{2}(s_{..X}+s_{kt}^{X})\ln(p_{kt}^{D}/p_{..D})\right]\\&-\left[\frac{1}{2}(s_{..M}+s_{kt}^{M})\ln(p_{kt}^{M}/p_{..M})-\frac{1}{2}(s_{..M}+s_{kt}^{M})\ln(p_{kt}^{D}/p_{..D})\right]\end{aligned} \tag{18}$$

Set the first country and period 1 as base, we have $P_{kt}^{XM}=P_{kt*}^{XM}/P_{11*}^{XM}=P_{11**}^{XM}/P_{11**}^{XM}$.

We can show that the following equation holds: $P_{kt}=P_{kt}^{D}*P_{kt}^{XM}$.

Real GDP on output-side is $Y_{kt}=\frac{v_{kt}}{P_{kt}}$, while Real GDP on expenditure-side is $Z_{kt}=\frac{v_{kt}}{P_{kt}^{D}}$.

Using $P_{kt}=P_{kt}^{D}*P_{kt}^{XM}$, we can get $Z_{kt}=Y_{kt}*P_{kt}^{XM}$.

So we can get the year-on-year decomposition:

$$\frac{Z_{kt}}{Z_{k(t-1)}} = \frac{Y_{kt}}{Y_{k(t-1)}} * \frac{P_{kt}^{XM}}{P_{k(t-1)}^{XM}} \tag{19}$$

4. Discussion and Conclusion

Based on production theory and translog GDP methodology introduced by Diewert and Morrison (1986), we further extend Inklaar and Diewert (2016)'s framework to simultaneously calculate real GDP on expenditure-side and output-side and isolate the effect of terms of trade on real GDP ,which enables us to measure the real GDP on both output-side and expenditure-side and the contribution of terms of trade to the real GDP growth across space and time in a more consistent way.

Empirical research should be done in next step. The input data are PPPs data for domestic consumption, export and import. We can get these from PWT 9.0. After estimating the real GDP on expenditure-side and output-side, these results can be compared with datasets such as PWT, WDI, UQICD et al. The estimates for terms of trade contribution factor can be compared with the existing works based on time dimension.

References

1. Denison,E. F. "International transactions in measures of the nation's production."Survey of Current Business,1981,61(5), pp.17-28.
2. Diewert, W. E. "Changes in the Terms of Trade and Canada's Productivity Performance."UBC Working Paper, 2008.
3. Diewert, W. E. and Morrison, C. J. "Adjusting output and productivity indexes for changes in the terms oftrade." Economic Journal, 1986, 96(3), pp.659-679.
4. Diewert, W. E., and Yu, E. "New Estimates of Real Income and Multifactor Productivity Growth for the Canadian Business Sector, 1961-2011." International Productivity Monitor, 2012, pp.27-48.
5. Feenstra, R. C. ; Heston, A. ; Timmer, M. P. and Deng H. Y. "Estimating real production and expenditures across nations: a proposal for improving the Penn World Tables."The Review of Economics and Statistics, 2009, 91(1), pp.201-212.
6. Feenstra, R. C. ; Inklaar, R. and Timmer, M.P."The next generation of the Penn World Table." NBER Working Paper, 2013.
7. Feenstra, R. C. ; Inklaar, R. and Timmer, M.P."PWT 8.0–a user guide."Groningen Growth and Development Centre Working Paper, 2013.
8. Feenstra, R. C. ; Ma, H. ; Neary J. P. and Prasada Rao D. S."Who shrunk China? Puzzles in the measurement of real GDP."The Economic Journal, 2013, 123 (573), pp.1100-1129.

9. Inklaar, R., and Diewert E. "Measuring Industry Productivity and Cross Country Convergence." Journal of Econometrics, 2016, 191(2), pp.426-433.
10. Kohli, U."A gross national product function and the derived demand for imports and supply of exports."Canadian Journal of Economics, 1978, 11(2), pp.167-182.
11. Kohli, U."An implicit Törnqvist index of real GDP."Journal of Productivity Analysis, 2004, 21(3), pp.337-353.
12. Kohli, U."Real GDP, real domestic income, and terms-of-trade changes."Journal of International Economics, 2004, 62(1), pp.83-106.
13. Kohli, U."Real GDP, Real GDI, and Trading Gains: Canada, 1981-2005."International Productivity Monitor, 2006, 13, pp.46-56.
14. Kohli, U."Terms-of-trade changes, real GDP, and real value added in the open economy: reassessing Hong Kong's growth performance."Asia-Pacific Journal of Accounting & Economics, 2007, 14(2), pp.87-109.
15. Reinsdorf, M. B. "Terms of trade effects: theory and measurement." Review of Income and Wealth, 2010, 56(S), pp.177-205.
16. United Nations. System of National Accounts. New York: United Nations, 1993.
17. United Nations. System of National Accounts. New York: United Nations, 2008.

The impact of longevity on a valuation of long-term investments returns: the case of selected european countries

Grażyna Trzpiot, Justyna Majewska
Department of Demography and Economic Statistics, University of Economics in Katowice, Poland, 40-881 Katowice, grazyna.trzpiot@ue.katowice.pl

Abstract
Both individuals and governments are increasingly concerned about effects of aging. Individuals are more concerned about increased longevity, because it affects their own financial and labour market plan, whereas governments are more concerned about old-age dependency as an aspect of population aging. Improvements in longevity and changing structure of population impact economy and financial stability. In this paper, we consider some economic, financial and demographic variables in a context of their impact on longevity. The Principal Component Regression is used in order to construct investment portfolios that are sensitive to risk factors.

Keywords
longevity; risk; PCA; investments; portfolios

1. Introduction

The recent analysis on lower long-term investment returns expectations over the next 20 years than they were in the past three decades is the inspiration for this paper (McKinsey, 2016). Individuals would need to save more for retirement, retire later, or reduce consumption during retirement. The global longevity trend will impact long-term investments returns. We attempt to identify risk factors that could have influence on the long-term investment return. Assessment of impact of each risk factor on portfolio returns regardless of the fixed risk level and scenarios creation is provided. Representative countries with different economy growth level and demographic situation are selected by cluster analysis. The Principal Component Analysis (PCA) is used to specify risk factors. The multifactor regression models (the Principal Component Regression, PCR) were built to describe the return rates of the assets (stock and bond) and risk factors. Three investment portfolios with different risk level (low, medium and high) was proposed as a particularly possibly investments, and they are considered as scenarios for the future level of long-term investment rates of return for selected countries. The paper extends existing analysis on effect of aging on economy and financial markets.

2. Methodology

Our research proceeds with the three main steps. For each task proper method is applied.

1. First step: selection of the European countries to the analysis. The cluster analysis is applied to choose representative countries from each cluster of countries due to the macroeconomic variables. Hierarchical method allows determining the best number of clusters as well as to see the hierarchical relations between obtained groups of countries. Steps 2 and 3 are conducted for each of the selected countries.
2. Second step: identification factors that could have influence on the long-term investment return. Dimension reduction by PCA is used for transformation of highly correlating variables into set of uncorrelated latent variables, and combination of several variables that characterize demographic changes and economic development into uncorrelated factors. Factors are associated with risks related with investments.
3. Third step: Simulation of three investment portfolios with different risk level (low, medium and high) as a particularly possibly investments. The level of the risk for long-term investment is determined by fixed percentage share of stocks and bonds. The investment rates of return were modeled through the PCR: risk factors – obtained in the step 2 – were used as predictors in a regression model fitted using the least squares procedure. There are two main reasons for regressing the investment return on the risk factors rather than directly on the explanatory variables. Firstly, the explanatory variables are often highly correlated (multicollinearity) which may cause inaccurate estimations of the least squares regression coefficients. Secondly, the dimensionality of the regressors is reduced by taking only a subset of PCs for prediction. A method does not require uncorrelated variables or normal distribution of the residuals.

PCR and PCA are both well now techniques for dimensionality reduction when modelling, and are especially useful when the independent variables are highly multicollinear (Jolliffe, 1982).

The selection of variables was preceded by an analysis of literature in the field of research on determinants of macroeconomic and financial implications of ageing. In the process of identification of risk factors the following variables are taken into consideration:

1. Demographic old-age dependency ratio – traditionally seen as an indication of the level of support available to older persons (those aged 65 or over, i.e. age when they are generally economically inactive) by the working age population (those aged between 15 and 64) [expressed per 100 persons of working age (15-64)].

2. Life expectancy at birth – the mean number of years that a newborn child can expect to live if subjected throughout his life to the current mortality conditions (age specific probabilities of dying) [expressed in years].
3. Life expectancy at age 65 – the mean number of years still to be lived by a man or a woman who has reached the age 65, if subjected throughout the rest of his or her life to the current mortality conditions (age-specific probabilities of dying) [expressed in years].
4. Consumer Price Index (CPI) – the change over time in the prices of consumer goods and services acquired, used or paid for by households [measured in an index, 2015 base year].
5. Real GDP per capita – the ratio of real GDP to the average population of a specific year; a measure of economic activity, used as a proxy for the development in a country's material living standards (a limited measure of economic welfare) [per capita, in current prices].
6. Unemployment rate – represents unemployed persons as a percentage of the labour force (the total number of people employed and unemployed) [% of active population].
7. Real effective exchange rates (REER) – aims to assess a country's price or cost competitiveness relative to its principal competitors in international markets; changes in cost and price competitiveness depend not only on exchange rate movements but also on cost and price trends [indices].
8. Gross saving – measures the portion of gross national disposable income that is not used for final consumption expenditure; gross national saving is the sum of the gross savings of the various institutional sectors [current prices].
9. Long-term government bond yields – refer to central government bond yields on the secondary market, gross of tax, with residual maturity of around 10 years; the bond or the bonds of the basket have to be replaced regularly to avoid any maturity drift [%].
10. Long-term care (health) expenditures – expenditures on a range of medical and personal care services that are consumed with the primary goal of alleviating pain and suffering and reducing or managing the deterioration in health status in patients with a degree of long-term dependency [share of current expenditures on health].
11. Currency exchange rates: EUR/USD, EUR/PLN.
12. Stock market a main index: DAX in Germany, IBEX35 in Spain, WIG20 in Poland.
13. Real Estate Funds and Equity/Dividend Funds: Unilmmo Deutchland and Allianz Vermögensbildung Deutschland (Germany), Seguffondo Inversion and Bankia Dividendo España FI (Spain), PZU UFK Investor Nieruchomości i Budownictwa and Investor FIO Subfundusz Akcji Spółek Dywidendowych (Poland).

Economic and demographic variables are derived from Eurostat database (variables 1-9) and OECD (variable 10), stock quotes – from stock exchange (Frankfurt, Madrid, Warsaw) and financial database (the Yahoo Finance) (variables 12 and 13). Time series were obtained for the time period 2010-2016. It is not wide period of time, thus some data were converted to monthly frequency (and then all variables were expressed as indices using a base year of 2010), with maintaining the strength and direction of correlation between variables. The period does not cover years from the financial crisis to avoid unusual observations from financial market.

Relations between the above-mentioned variables and longevity are analyzed in empirical studies. Some relations are clear, while others are still a subject of debate (in particular, the impact of longevity on inflation is unclear). Due to the complexity of these relations and their multidimensionality, it is worth mentioning a few confirmed consequences of longevity (e.g. Bloom et al., 2010; Rachel and Smith, 2015; Maestas at al., 2016; Acemoglu and Restrepo, 2017): reducing investment return, reducing public saving, reducing growth rates, reducing real interest rates, affecting labor supply and returns, reallocation of saving from riskier to safe assets may lead to potential mispricing of risk, running down assets may result in negative wealth effects.

Based on the results of Majewska and Trzpiot (2016) mentioned above variables could be grouped into five clusters: standard of living risk, elderly needs risk, financial risk, longevity risk and long-term investment risk. However, it should be taken into account that the time period in their research covered years of financial crisis 2008-2009. All calculations were made in R software environment.

3. Result

Empirical investigation of relations between longevity phenomenon and selected macroeconomic and financial variables is made for selected European countries with different level of economic growth and life expectancy, i.e. for Germany, Spain and Poland. From longevity perspective, life expectancy (at birth and at aged 65, for both sexes) in Poland is shorter than in Germany, and Spain, while life expectancy is the highest in Spain. Spain is expected to become the world's second oldest country by 2050, behind Japan. According to HDI index Germany – since 2010 – has been in the group of five the most developed countries, Spain – in the second ten, and Poland – in the third ten the most developed countries in world (UNDP, 2018).

Table 1. Risk factor loads of principal components: Germany

Variable	*F1*	*F2*	*F3*
Old Age Dependence Ratio	0.93		
GDP	0.92		
Gross Saving	0.93		
CPI	0.80		
Long-term Care Expenditures	0.95		
Long-term Government Bond Yields	-0.89		
REER			0.92
Unemployment Rate	-0.90		
Dividend Fund	0.75		
Real Estate Fund			0.63
LE65		0.88	
LE birth		0.97	
DAX	0.75		
EUR/PLN	0.74		
EUR/USD	-0.85		
Cumulative Var	*0.59*	*0.73*	*0.83*

Source: own calculations

For Germany (tab. 1) the first principal component explains 59% of the variation, while all components – 83%. The first component is identified as the wealth risk because of the high positive factor loadings on GDP, gross savings, long-term care expenditures combined with a high negative weighting on long-term government bond yields and unemployment rate. All these variables are associated with standard of living risk and elderly needs risk. The second component has been high loadings of variables that reflect longevity. Advancing age due to increased life expectancy itself is a risk factor. The last component explains 10% of total variance and has been loaded only by REER and real estate fund and it would associate with financial market risk.

Table 2. Risk factor loads of principal components: Spain

Variable	*F1*	*F2*	*F3*
Old Age Dependence Ratio	0.93		
GDP		-0.94	
Gross Saving		-0.55	
CPI	0.83		
Long-term Care Expenditures		-0.66	
Long-term Government Bond Yields	-0.78		
REER	-0.61		
Unemployment Rate		0.93	
Dividend Fund	0.69		
Real Estate Fund	-0.91		
LE65	0.85		
LE birth	0.94		
IBEX35			0.96
EUR/PLN			-0.65
EUR/USD	-	-	-
Cumulative Var	*0.44*	*0.68*	*0.82*

Source: own calculations

For Spain (tab. 2) the first component has been the highest loadings with variables that reflect elderly needs and longevity. Thus, this component has been identified as long-term standard of living. This can be supported by noting that the factor loadings associated with long-term care expenditures, REER and real estate fund are negative. The second component has shown to be a strong indicator of longevity risk related with GDP, gross savings, unemployment rate and long-tern expenditures. This component has been clustered with standard of living and long-term investments risk factors. The last component explains 14% of total variance and has been positively loaded with IBEX35 returns and negatively – with EUR to PLN exchange rate. It would be associated with financial market risk.

Table 3. Risk factor loads of principal components: Poland

Variable	*F1*	*F2*	*F3*
Old Age Dependence Ratio	0.86		
GDP	0.89		
Gross Saving	0.77		
CPI	0.72		
Long-term Care Expenditures	-0.84		
Long-term Government Bond Yields	-0.91		
REER		0.80	
Unemployment Rate	-	-	-
Dividend Fund		0.92	
Real Estate Fund			-0.97
LE65	0.95		
LE birth	0.95		
WIG20		0.88	
EUR/PLN		-0.81	
EUR/USD	-	-	-
Cumulative Var	*0.54*	*0.77*	*0.87*

Source: own calculations

For Poland (tab. 3) the first component has been loaded with variables related with increasing life expectancy as well as economic well-being and explains 54% of total variance. Therefore, the component has been identified as demo-economic risk. The second component has been heavily loaded with dividend fund, WIG20 returns, REER and negatively with EUR to PLN exchange rate. It would be identified as financial market risk and explained 23% of total variance. The last component was loaded negatively only by a real estate fund and would be associated with individual wealth risk.

Construction of portfolios of weight average of stock and bonds, then return rate of main index on stock exchange and relative change of monthly return rate of long-term government bond yields 10-year. We construct different portfolio with weights: 40/60, 50/50 and 60/40 (proportion of stock and bond respectively). The return rates of built portfolios were calculated by means of the multifactor model.

Scenario #1: Portfolio return rate 40/60:

$R_{pGERMANY} = -0.72F1 \quad R^2 = 0.53$

The interpretation for this result is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.72%.

$R_{pSPAIN} = -0.25F1 + 0.94F3 \quad R^2 = 0.95$

The interpretation of this equation is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.25%, if risk represented by *F3* increase by 1, then R_p will increase by 0.94%.

$R_{pPOLAND} = -0.84F1 + 0.4F2 - 0.16F3 \quad R^2 = 0.91$

The interpretation of this equation is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.84%, if risk represented by *F2* increase by 1, then *Rp* will increase by 0.4%, if risk represented by *F3* increase by 1, then *Rp* will decrease by 0.16%.

Scenario #2: Portfolio return rate 50/50:

$R_{pGERMANY} = -0.39F1 + 0.33F2 \quad R^2 = 0.27$

The interpretation for this result is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.39%, if risk represented by *F2* increase by 1, then *Rp* will increase by 0.33%.

$R_{pSPAIN} = -0.22F1 + 0.049F2 + 0.95F3 \quad R^2 = 0.95$

The interpretation of this equation is as follows: if risk represented by *F1* increase by 1, then *Rp* will increase by 0.02%, if risk represented by *F2* increase by 1, then R_p will decrease by 0.045%.

$R_{pPOLAND} = -0.81F1 + 0.47F2 - 0.17F3 \quad R^2 = 0.92$

The interpretation of this equation for Poland is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.81%, if risk represented by *F2* increase by 1, then R_p will increase by 0.47%, if risk represented by *F3* increase by 1, then *Rp* will decrease by 0.17%.

Scenario #3: Portfolio return rate 60/40:

$R_{pGERMANY} = 0.51F2 \quad R^2 = 0.3$

The interpretation of this equation for Germany is as follows: if risk represented by *F2* increase by 1, then *Rp* will increase by 0.51%.

$R_{pSPAIN} = -0.21F1 + 0.063F2 + 0.95F3 \quad R^2 = 0.94$

The interpretation for this result is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.21% , if risk represented by *F2* increase by 1, then *Rp* will increase by 0.063%, if risk represented by *F3* increase by 1, then *Rp* will increase by 0.95%.

$R_{pPOLAND} = -0.76F1 + 0.56F2 - 0.18F3 \quad R^2 = 0.93$

The interpretation for this result is as follows: if risk represented by *F1* increase by 1, then *Rp* will decrease by 0.76%, if risk represented by *F2* increase by 1, then *Rp* will increase by 0.56%, if risk represented by *F3* increase by 1, then *Rp* will decrease by 0.18%.

4. Discussion and Conclusion

There is statistically significant effect extracted by PCA risk factors on investment returns. In Germany we can point out three factors: wealth risk, systematic longevity risk and financial market risk that impact statistically significant effect on each portfolio returns – czy to jest potrzebne bo w każdym zdaniu tak jest, I nie widać różnic między krajami. In Spain we have three different factors: long-term standard of living risk, local economy risk, financial market risk which impact statistically significant effect on each portfolio returns. At the end, for Poland we receive: demo-economic risk, financial market risk, individual wealth risk which have statistically significant effect on each portfolio returns. Calibrated models are acceptable, statistically significant so we can use these models for prediction return of portfolio. The best quality of estimated models we obtain for Spain and Poland, the worst for Germany.

The effect of large long-term investors on both their investments and on the markets generally has prompted a key debate in academic literature. Presented results, we hope are in this important flow in age of aging.

References

1. Acemoglu, D., & Restrepo, P. (2017). Secular Stagnation? The Effect of Aging on Economic Growth in the Age of Automation. *American Economic Review, 107(5),* 174–179.
2. Bloom, D. E., Canning, D., & Fink, G. (2010). Implications of population ageing for economic growth. *Oxford Review of Economic Policy, 26(4).* 583-612.
3. Bosworth, B., Bryant, R., & Burtless, G. (2004). The Impact of Aging on Financial Markets and the Economy: A Survey (July 1, 2004). *The Brookings Institution*. Available at SSRN: https://ssrn.com/abstract=1147668.
4. Jolliffe, Ian T. (1982). A Note on the Use of Principal Components in Regression, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3). 300-303.
5. Maestas, N., Mullen, K., & Powell, D. (2016). The Effect of Population Aging on Economic Growth, the Labor Force and Productivity. Working Papers. Rand Corporations.
6. McKinsey Global Institute. (2016). Diminishing Returns: Why Investors May Need to Lower Their Expectations.
7. Rachel, L., & Smith, T. (2015). Secular Drivers of the Global Real Interest Rate. *Bank of England Staff Working 571*. J.P. Morgan Asset Management Multi-Asset Solutions.
8. UNDP. (2018). Human Development Indices and Indicators 2018: Statistical update, UN, New York
9. Trzpiot, G., & Majewska, J. (2016). The Impact of Longevity on Long-term Investments Returns: Scenarios for Europe. Available at https://www.cass.city.ac.uk/__data/assets/pdf_file/0020/334082/L12-32-TRZPIOT-and-MAJEWSKA.pdf

Profiling filipino senior high school students' performance in statistics and probability

Armi S. Lantano[1], Kevin Carl P. Santos[2]
[1] Center for Educational Measurement, Inc.; Makati City, Philippines
[2] School of Statistics, University of the Philippines

Abstract
One of the major reforms brought about by the new K to 12 Basic Education Program in the Philippines is the implementation of the Senior High School (SHS) curriculum for Grades 11 and 12. The SHS curriculum was launched in School Year (SY) 2016-2017 and the first batch of completers graduated in SY 2017-2018. To date, reports on the academic performance of those completing the program are very limited. Because Statistics and Probability, as one of the core curriculum subjects, is an area where skills necessary to critical consumers of information are taught, this study aims to examine the performance profile of the SHS students in this subject. This study employs the Higher-order Item Response Theory model framework to identify the specific strengths and weakness of the students in the different content areas of the said subject by administering the Center of Educational Measurement Achievement Test in Statistics and Probability. Results of the study may serve as a springboard in evaluating the success of the program.

Keywords
Statistics Education; Statistics and Probability; Philippine K to 12 Basic Education Program; Senior High School Performance; Student Performance

1. Introduction

In the Philippines, the K to 12 Basic Education Program was implemented starting School Year (SY) 2012-2013 to hone basic education level completers to be globally competitive. From the previous curriculum covering 10 years of basic education, the new curriculum is now extended to 13 years of study (i.e., one year of pre-elementary level, 6 years of elementary level, 4 years of junior and 2 years of senior high school). One of the salient features of the program and in fact the biggest reform brought of the K to 12 programs was the carrying out of Senior High School (SHS) curriculum that added two more years of specialized study in the secondary level. Students in the upper secondary level (i.e., Grades 11 and 12) may opt specializations based on their aptitudes, interests and the school's capacity. Depending on their choice of specializations, students take subjects related to their specific SHS Track and those falling under the Core Curriculum. There are seven Learning Areas under the Core Curriculum namely, Communication, Languages, Literature,

Philosophy, Natural Sciences, Social Sciences and Mathematics. Embedded in these Learning Areas are the content and performance standards of the 21st Century Skills. Such skills include problem-solving, information literacy and critical thinking.

In SY 2016-2017, the SHS curriculum commenced and produced its first batch of graduates in SY 2017-2018. Reports have been published on the success of the SHS program exceeding expectations, using the enrolment rate and those completing the program as indicators (Montemayor, 2018). Several studies have also been published and presented in international conferences on the effectiveness of the capacity building of teachers in the SHS teachers (Ocampo & Ocampo, 2018; Reston & Locquas, 2018; Candelario-Aplaon, 2017). In SY 2017-2018, mandated by the Department of Education (DepEd), the Grade 12 National Achievement Test (NAT12) for SY 2017-2018 was administered to selected SHS students in both public and private schools nationwide in March 2018 (Hernando-Malipot, 2018). This being the first national examination given to K to 12 completers was geared towards determining whether the students are meeting the learning standards, providing information to improve instructional practices, assessing/evaluating effectiveness of education service delivery using learning outcomes as indicators, and providing empirical information as bases for curriculum, learning delivery, assessment and policy reviews, and policy formation. Up to moment of writing this research, details as well as results of NAT12 have not been made available to the public. There have also been reports, but very limited, on the academic performance of those completing the program.

Thus, this study examined the performance of the SHS students in one of the core subjects under Mathematics – Statistics and Probability. Statistics and Probability was particularly considered in this study due to current demand of individuals having skills on data management and analysis in this Age of Information and Technology. Furthermore, the strengths and weaknesses of students in terms of content areas of Statistics and Probability were identified. Results of the study may serve as springboard in evaluating success of the program, particularly in Statistics Education.

2. Methodology

CEM K to 12 Achievement Test in Statistics and Probability. In order to keep abreast with the curricular changes in the country, the Center for Educational Measurement, Inc. (CEM) – the pioneering testing and research institution in the Philippines, developed a series of achievement tests, the CEM K to 12 Achievement Tests. The CEM K to 12 Achievement Tests are standardized tests designed to measure knowledge and skills learned in school based on the national curriculum. These include tests in English, Mathematics and Science from Kindergarten to Grades 11/12. By SY 2017-

2018, CEM has already released five achievement tests for SHS level including Statistics and Probability. These tests may be taken at end of any semester of Grades 11 and 12 depending on the semester when the subjects are taught.

The CEM K to 12 Achievement Test in Statistics and Probability, in particular, is composed of 60 multiple choice items partitioned in five content areas, namely: (1) *Random Variables and Probability Distribution* (CA01) (2) *Sampling and Sampling Distribution* (CA02), (3) *Estimation of Parameters* (CA03), (4) *Test of Hypothesis* (CA04) and (5) *Correlation and Regression Analyses* (CA05). The reliability of the test is 0.83, whereas the concurrent validity of the test yields coefficients ranging from 0.38 to 0.65, indicating that the achievement test is reliable and valid.

Participants of the Study. A total of 2,536 Filipino SHS students coming from 11 private schools nationwide and who took the CEM K to 12 Achievement Test in Statistics and Probability for Grades 11/12 in SY 2017-2018 were considered in the study. The distribution of the students by gender, location, school type, and grade level is summarized in Table 1.

Table 1. Demographic Profile of the Students

	N	Percent (%)
Gender		
Male	1,113	55.8
Female	1,416	43.9
Missing	7	0.3
Total	**2,536**	**100.0**
Location		
National Capital Region (NCR)	1,895	74.7
Luzon	235	9.3
Visayas	295	10.2
Mindanao	147	5.8
Total	**2,536**	**100.0**
School Type		
Private Non-Sectarian	1,924	75.9
Private Sectarian	612	24.1
Total	**2,536**	**100.0**
Grade Level		
Grade 11	510	20.1
Grade 12	2,026	79.9
Total	**2,536**	**100.0**

Higher Order Item Response Theory. Due to the nature of educational assessments that measure multiple abilities or construct (Reckase, 1985; Ackerman, et.al., 2003), this study considered using a Higher-Order Item Response Theory (HO-IRT) model approach in analyzing the data. HO-IRT is

a framework integrating one general and several domain-specific abilities in the same model and used to analyze assessment data consisting of different domains (e.g., content areas, objectives). Particularly, one-factor HO-IRT model formulation (de la Torre & Song, 2009) was employed in obtaining the ability estimates for the overall and five content areas of the Statistics and Probability test. The HO-IRT model can be illustrated as that of Figure 1. Based on Figure 1, this approach assumes that the overall domain ability is associated with subdomain abilities, where λs represent the slope of the overall domain ability ϑ when regressed with each subdomain ability $\vartheta^{(k)}$. $\boldsymbol{\beta}$s represent the item parameters based on an IRT model. In this case, the three- parameter logistic model was fitted to the data. Both the $\boldsymbol{\beta}$s and $\vartheta^{(k)}$ influence the item responses of the examinees.

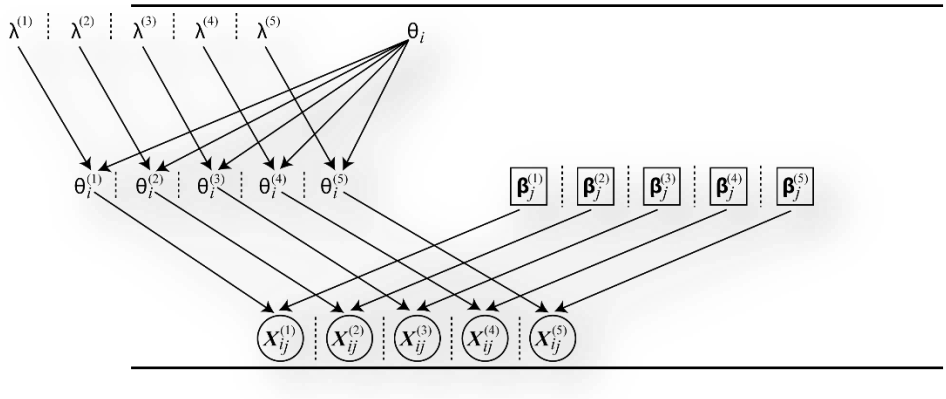


Figure 1. HO-IRT Model of the Statistics and Probability Test

3. Results

Correlations between Content Areas. The estimate of the correlational structure using HO-IRT approach of the five content areas in the Statistics and Probability test is summarized in Table 2. It can be seen that the correlations among the content areas yielded coefficients ranging 0.73 to 0.90, with association between *Estimation of Parameters* and *Test of Hypothesis* as the strongest whereas association between *Sampling and Sampling Distribution* and *Correlation and Regression Analyses* as the weakest. The high correlation values indicate that as student's mastery in one content area becomes apparent, his/her mastery in the other content area becomes evident as well.

Students' Profile. Table 3 presents the descriptive statistics of ability estimates for Overall and the five content areas. The mean ability estimates were found to be consistently greater than 0.00, indicating that the students performed above average. In terms of content areas, the mean ability estimates ranged from 0.13 to 0.17, with *Random Variables and Probability Distributions* as the lowest and *Estimation of Parameters* as the highest. This

means that the students found *Random Variables and Probability Distributions* as the hardest content area while *Estimation of Parameter* as the easiest among the five. Moreover, the distribution of the ability estimates for the Overall and across the five content areas were found to be slightly skewed to the left, indicating that most of the students have non-negative ability estimates. This further means that most of the students performed above average.

Table 2. Estimated Correlation Between the Five Content Areas

Domains	CA02	CA03	CA04	CA05
CA01	0.79	0.87	0.87	0.78
CA02		0.82	0.81	0.73
CA03			0.90	0.81
CA04				0.80

Table 3. Descriptive Statistics of Students' Performance (N=2,536)

Overall/Content Area	Mean	Standard Deviation	Minimum	Maximum	Skewness
Overall	0.16	0.82	-1.99	2.94	0.21
CA01	0.13	0.81	-2.01	2.98	0.18
CA02	0.15	0.75	-1.79	2.68	0.29
CA03	0.17	0.80	-1.90	2.83	0.25
CA04	0.16	0.80	-1.91	2.84	0.27
CA05	0.15	0.77	-1.77	2.78	0.25

Proportions of Students with Above Average Proficiency Level. The proportions of students with above average proficiency level overall and by subgroup are presented in Figures 2 and 3, respectively. Figure 2 shows that in general, there were slightly more students with above proficiency level in terms of overall and the five content areas. However, this also implies that roughly half of the students had below average proficiency level across the content areas. Hence, teachers might need to perform some remediation measures with those students to improve their performances.

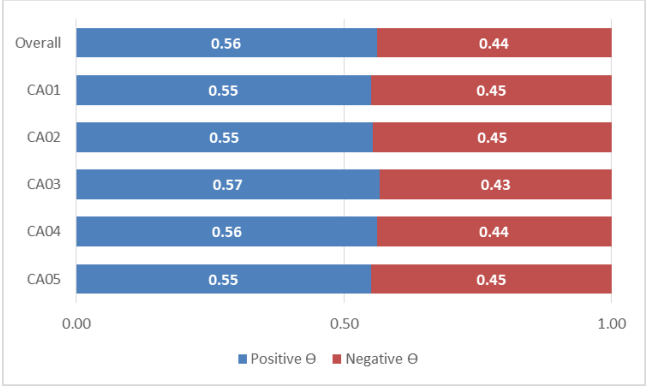
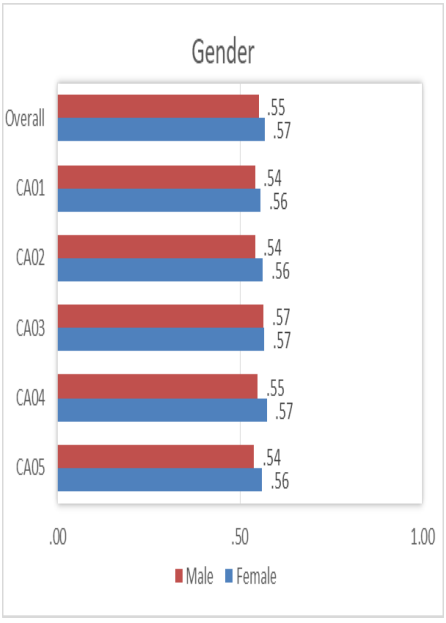
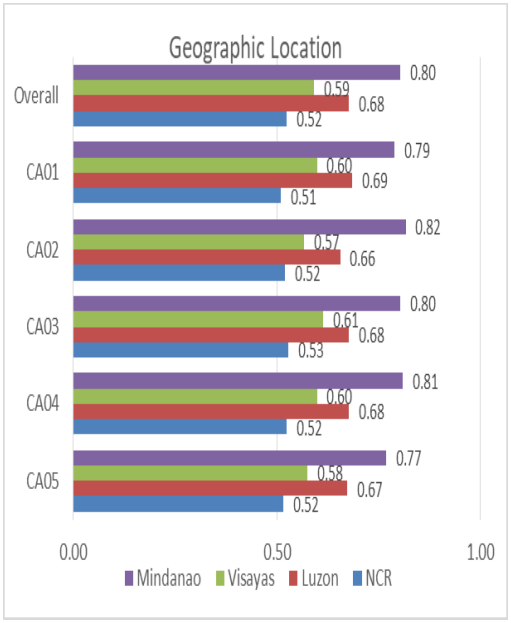


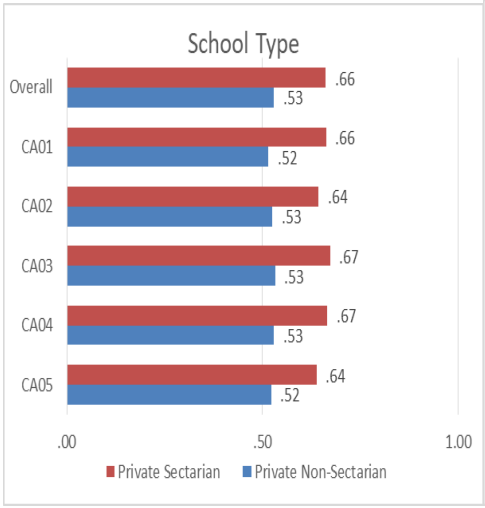
Figure 2. Proportions of Students with Above and Below Average Proficiency Level

Note: Positive ϑ = above average; Negative ϑ = below average

Analyzing the proportions by group, Figure 3 reveals that the proportions of students with beyond average proficiency exceeded 50% for all subgroups of gender, geographic location, school type and grade level. Figure 3 also shows that the proportions of both male and female students having above average proficiency level were almost the same. In terms of geographic location, results showed that Mindanao had consistently the highest proportions whereas NCR has consistently the lowest proportions. Private Sectarian schools consistently outperformed Private Non-Sectarian schools. The same pattern could observe between Grades 11 and 12, respectively.







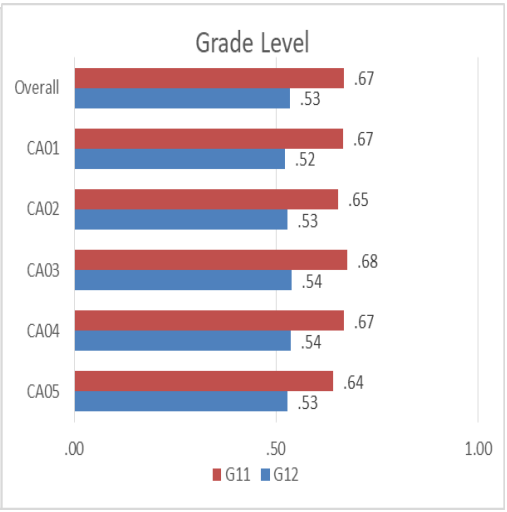


Figure 3. Proportions of Students with Above Average Proficiency Level by Group

4. Discussion and Conclusion

Results of this study revealed that the performance of the SHS students in the CEM K to 12 Achievement Test in Statistics and Probability was slight above average. This was further strengthened with the findings based on the proportions of students with above average proficiency level. The proportions also showed that the students' performance was slightly above average in all the five content areas, be at any subgroups of the grouping variables. The difference in proportions recorded in terms of school type, with private sectarians outperforming private non-sectarian schools, is consistent with other studies on academic performance of Filipino students. The advantage of Grade 11 over Grade 12 warrants further researches on the possible causes of such results. Contrary to the pattern of proportions observed on the subgroups of geographic location, most studies showed that students in NCR tends to perform better non-NCR (i.e., Luzon, Visayas and Mindanao) students mainly due to NCR being the center of the country's socio-economic welfare (Franco & Lantano, 2009; Gatchalian & Lantano, 2010). It is interesting to determine the possible reasons why NCR students did not perform superior to other students in this achievement test.

As a start, the K to 12 SHS program in Statistics and Probability seemed to work for the schools included in this study although remediation measures and, probably, improvement in classroom instruction must be done to improve the performance of the students from just slightly above average proficiency. It must be noted, however, that only 11 private schools were included in this study. This count of schools comprises a very small portion of the Philippine school population. A large-scale study using the same method and considering greater number and variety of schools as well as researches on other core curriculum is strongly recommended to conclude that the K to 12 SHS program is indeed effective in raising the standards of basic education completers. Exploring the profile of students grouped by their chosen tracks could also be a future direction for research.

References

1. Auckerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*, 37-51.
2. Candelario-Aplaon, Z. (2017). Needs Assessment of Senior High School Mathematics Teachers in Teaching Statistics and Probability. *International Forum, 20(2)*, 143-159.
3. De la Torre, J. and Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model Approach. *Applied Psychological Measurement, 33(8)*, 620-639.
4. Franco, M.L. & Lantano, A. (2009). Study of fairness in the Philippine Aptitude Classification Test. Paper present at the 35th International Association of Educational Assessments (IAEA) Conference, Brisbane, Australia.
5. Gatchalian, C. & Lantano, A. (2010). Revisiting the Philippine Aptitude Classification Test: An analysis of potentially biased items. Paper presented at the 36th International Association of Educational Assessments (IAEA) Conference, Bangkok, Thailand.
6. Hernando-Malipot, M. (2018). Grade 12 students to take assessment test in March. Manila Bulletin. Retrieved from https://news.mb.com.ph/2018/02/13/grade-12-students-to-take-assessment-test-in-march/
7. Montemayor, M.T. (2018). 2018 Senior High School implementation exceeds DepEd outlook. *PTV News*. Retrived from https://ptvnews.ph/2018-senior-high-school-implementation-exceeds-deped-outlook/
8. Ocampo, S. and Ocampo, B. (2018). Capacity Building of Statistics Teachers' through Mentoring and Innovative Ways. A paper presented at the 10th International Conference on Teaching Statistics (ICOTS10), Kyoto, Japan.
9. Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.
10. Reston, E. and Locquias, C. (2018). Improving Statistical Pedagogy among K to 12 Mathematics Teachers in the Philippines. A paper presented at the 10th International Conference on Teaching Statistics (ICOTS10), Kyoto, Japan.

Small area estimation for linear parameter under a spatial unit-level lognormal model

Dian Handayani[1,2,4], Henk Folmer[2,3], Khairil Anwar Notodiputro[4], Anang Kurnia[4], Asep Saefuddin[4], Arno J. Van der Vlist[2], I Wayan Mangku[5]

[1.] Department of Statistics, Faculty of Mathematics and Natural Sciences, State University of Jakarta, Indonesia.

[2.] Department of Economic Geography, Faculty of Spatial Sciences, University of Groningen, The Netherlands.

[3.] College of Economics and Management, Northwest Agriculture and Forestry University, Yangling, China.

[4.] Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia.

[5.] Department of Mathematics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia.

Abstract

In this paper, we propose the extended Spatial Empirical Best Prediction (SEBP) to estimate linear parameter, i.e. small area mean, whenever variable of interest has positively skewed distribution and spatial dependence among small areas are taken into account. The extended SEBP improve the SEBP (Handayani, 2018) by estimating values of variable of interest with the conditional expectation of variable of interest given the data and random area effects. A parametric bootstrap is proposed for the estimation of mean square error of estimates of linear parameter. Our simulation studies indicate that the relative performance of the SEBP that we propose is outperform in terms of bias and mean square error.

Keywords

spatial empirical best predictor; skewed data; spatial dependence; parametric bootstrap.

1. Introduction

Indirect estimation method in Small Area Estimation (SAE) is usually model-based method. The standard SAE method is developed under linear mixed model which assumes normality on variable of interest and independence among random area effect. Berg and Chandra (2014) proposed Empirical Best Predictor (EBP) to estimate linear parameter, i.e population mean, for variable of interest which has positively skewed distribution but among small areas are still assumed to be independent. Handayani et al (2018) developed the Spatial Empirical Best Predictor (SEBP) to estimate population mean, for positively skewed variable of interest and the spatial dependence among random area effect are taken into account. By using the SEBP,

Handayani et al (2018) estimate the values of variable of interest for non-sampled units using the expectation of the values of variable of interest. In this paper, we extent the SEBP which provides the estimates of values of variable of interest using conditional expectation the values of variable of interest given the data and random area effects.

2. EBLUP under Nested Regression Mode:

In this section, we describe the EBLUP of population mean in small area i (denoted by μ_i) under unit level model. Suppose there are M small areas and N_i units within small area i (i = 1,2 … M). The EBLUP of μ_i (denoted by $\hat{\mu}_i^{EBLUP}$) based on variable of interest y_{ij} and auxiliary information x_{ij} which are available in units' level is derived under nested error regression model as follows:

$$y_{ij} = x_{ij}{}^T \beta + z_{ij} v_i + e_{ij} \quad j = 1,2 \dots N_i; i = 1,2 \dots M \tag{1}$$

where β is the parameter of fixed effect xij, zij is matrix of known positive constant, vi is random area effect i which is assumed to be independently normally distributed $vi \sim iid\ N(0, \sigma v2)$ and eij is sampling error unit- j in small area which is also assumed to be independently normally distributed

$$e_{ij} \sim iid\ N(0, \sigma_e{}^2)$$

The estimation of μi will be based on the selected sample with size is ni, i = 1,2 … M. The mean

μi of the yij in small area i can be written by :

$$\mu_i = \frac{1}{} \left[\sum_{j \in Ni\ si} y_{ij} + \sum_{j \in ri} y_{ij} \right] ; i = 1,2 \dots M \tag{2}$$

where s denotes sampled observations and r non-sampled observations. Under model (1) for small area i, the best linear unbiased predictor (BLUP) for μi is given by :

$$\hat{\mu}_i^{BLUP} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{BLUP} \right] ; i = 1,2 \dots M$$

$$\text{where } \hat{y}_{ij}^{BLUP} = x_{ij}^T \hat{\beta} + z_{ij} \hat{v}_i = x_{ij}^T \hat{\beta} + z_{ij} \gamma_i \left(\bar{y}_{is} - \bar{x}_{is}^T \hat{\beta} \right) \ ; \ \gamma_i = \frac{(z_{ij}^2 \sigma_v^2)}{(z_{ij}^2 \sigma_v^2 + \sigma_e^2 / n_i)} \tag{3}$$

is a shrinkage factor where $\hat{y}_{ij}$ (ratio between the model variance relative to the total variance); $\hat{\beta} = (X^T V^{-1} X)^{-1} (X V^{-1} Y)$ is a weighted least squares estimator of β, $\bar{y}_{is}$ and $\bar{x}_{is}$ are the sample means of the interested variable Y and auxiliary variable X in small area i (Rao and Molina, 2015).

In practice, the parameters $\sigma_v{}^2$ and $\sigma_e{}^2$ are usually unknown. By replacing ($\sigma_v{}^2$, $\sigma_e{}^2$) in (3) by their estimates ($\hat{\sigma}_v{}^2$, $\hat{\sigma}_e{}^2$), the empirical best linear unbiased predictor (EBLUP) of μ_i is obtained:

$$\hat{\mu}_i^{EBLUP} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{EBLUP}\right];\ i = 1,2 \ldots M \tag{4}$$

where : $\hat{y}_{ij}^{EBLUP} = x_{ij}^T\hat{\beta} + z_{ij}\hat{v}_i = x_{ij}^T\hat{\beta} + z_{ij}\hat{\gamma}_i(\bar{y}_{is} - \bar{x}_{is}^T\hat{\beta})$; $\hat{\gamma}_i = \frac{(z_{ij}^2\hat{\sigma}_v^2)}{(z_{ij}^2\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i)}$

See Rao and Molina (2015) for further details of second-order approximation to the Mean Square error (MSE) of $\hat{\mu}_i^{EBLUP}$and the estimator of MSE approximation.

3. EBP under Unit Level Lognormal Model:

Berg and Chandra (2014) developed Empirical Best Prediction (EBP) method to estimate linear parameter, i.e. population mean, for variable of interest y_{ij} have log normal distribution. The EBP is derived under unit level lognormal model as follows:

$$\log(y_{ij}) := l_{ij} = \beta_0 + x_{ij}'\beta_1 + v_i + e_{ij} \tag{5}$$

where $v_i \sim iid\ N(0, \sigma_v^2)$, $e_{ij} \sim iid\ N(0, \sigma_e^2)$ and v_i and e_{ij} are independent. Based on (5), y_{ij} is given by :

$$y_{ij} = exp(\beta_0 + x_{ij}'\beta_1)exp(v_i + e_{ij}) \tag{6}$$

The EBP for population mean in small area- i (denoted by μ_i) is given by :

$$\hat{\mu}_i^{EBP} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{EBP}\right];\ i = 1,2 \ldots M$$

where

$$\hat{y}_{ij}^{EBP} = E\left(y_{ij}|(y,x)\right) = exp(\beta_0 + x_{ij}'\beta_1)\, E\left(exp(v_i + e_{ij})|(y,x)\right) \tag{7}$$

By properties of the normal distribution, for $j \in \bar{s}_i$,

$$(u_i, e_{ij})\ |\ (y,x) \sim N\left[\{\gamma_i\,(\bar{l}_{is} - \beta_0 - \bar{x}_{is}'\beta_1), 0\}, diag(\gamma_i n_i^{-1}\sigma_e^2), \sigma_e^2\right] \tag{8}$$

By (7) and the moment generating function of the lognormal distribution, for $j \in \bar{s}_i$

$$E\left(exp(v_i + e_{ij})|(y,x)\right) = exp\left[\gamma_i(\bar{l}_{is} - \beta_0 - \bar{x}_{is}'\beta_1) + \frac{1}{2}(\gamma_i n_i^{-1}\sigma_e^2 + \sigma_e^2)\right] \tag{9}$$

By (9) and (7), the Best Predictor (BP) for y_{ij} is given by:

$$\hat{y}_{ij}^{BP} = exp\left(\beta_0 + x_{ij}'\beta_1 + \gamma_i(\bar{l}_{is} - \beta_0 - \bar{x}_{is}'\beta_1) + \frac{1}{2}\hat{\sigma}_e^2(\gamma_i n_i^{-1}\sigma_e^2 + \sigma_e^2)\right) \tag{10}$$

Since $\theta = (\beta_0, \beta_1, \sigma_v^2, \sigma_e^2)^T$ is usually unknown , the θ is replaced with $\hat{\theta}$ to obtain the Empirical Best Prediction (EBP) $\hat{y}_{ij}^{EBP}$ as follows :

$$\hat{y}_{ij}^{EBP} = exp\left(\hat{\beta}_0 + x'_{ij}\hat{\beta}_1 + \gamma_i(\bar{l}_{is} - \hat{\beta}_0 - \bar{x}'_{is}\hat{\beta}_1) + \frac{1}{2}\hat{\sigma}_e^2(\hat{\gamma}_i n_i^{-1} + 1)\right) \quad (11)$$

where: $\bar{l}_{is} = n_i^{-1}\sum_{j\in s_i} l_{ij}$; $\bar{x}'_{is} = n_i^{-1}\sum_{j\in s_i} x'_{ij}$; $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n}}$

See Berg and Chandra (2014) for further details of derivation of the closed form of Mean Square Error (MSE) of $\hat{y}ij$ $EBP$and the estimates of MSE.

4. SEBP under Spatial Unit Level Lognormal Model

In this section, we extend the SEBP (Handayani et al, 2018) by estimating y_{ij} with the conditional expectation of y_{ij} given the data and random area effects. The SEBP of population mean μi is derived under spatial unit level lognormal model as follows:

$$Y^* = \log Y = X\beta + Zu + e = X\beta + Z(I_M - \rho W)^{-1}v + e \quad (12)$$

where autoregressive coefficient u is vector of random area effect which is assumed to follow a SAR process with spatial ρ and weighted matrix W; $u = \rho Wu + v = (I_M - \rho W)^{-1}v$; $v \sim N(0, G = \sigma_v^2 I_M)$; $e \sim N(0, R_1 = \sigma_v^2 I_N)$; $u \sim N(0, D = \sigma_v^2[(I_M - \rho W)(I_M - \rho W^T)]^{-1})$; $Y^* \sim N(X\beta, \Sigma_1), \Sigma_1 = ZDZ^T + R_1$; $(Y^*|u) \sim N(X\beta + Zu, R_1)$.

The spatial best predictor (SBP) for Y is given by:

$$\hat{Y}^{SBP} = E(Y|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) = exp(X\beta)\, exp\left(1'DZ'\Sigma_1^{-1}(Y - X\beta) + \frac{1}{2}1'\Sigma_2 1\right) \quad (13)$$

where: $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) = \{y_{ij}; i = 1,2 \ldots M, j \in s_i\} \cup \{x_{ij}; i = 1,2 \ldots M, j \in U_i\} \cup \{v_i; i = 1,2 \ldots M\}$; $\Sigma_2 = (Z'R_2^{-1}Z + D^{-1})^{-1} + R_2$; $R_2 = \sigma_e^2 I_M$. Detailed derivation of this result is provided in Handayani (2019).

The spatial best predictor (SBP) for $\mu_i = \frac{1}{}[\sum_{j\in Nisi} y_{ij} + \sum_{j\in ri} y_{ij}]$ is given by:

$$\hat{\mu}_i^{SBP} = \frac{1}{N_i}[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{SBP}]; i = 1,2 \ldots M, j = 1,2 \ldots N_i \quad (14)$$

where $\hat{y}_{ij}^{SBP} = exp\left(x_{ij}^T\beta + z_{ij}u_i + \frac{1}{2}(z_{ij}^2\tau_i^2 + \sigma_e^2)\right)$; $u_i = b_i^T DZ'\Sigma_1^{-1}(Y - X\beta)b_i$; $\tau_i^2 = b_i^T\Sigma_2 b_i$

b_i^T is an M vector (0,0,...0,1,0,0...) with 1 for the i^{th} area. The spatial empirical best predictor (SEBP) of μ_i , $\hat{\mu}_i^{SEBP}$, is derived by replacing (σ_v^2, σ_e^2, ρ) by their estimates ($\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$, $\hat{\rho}$).

If the log transformed of variable of interest does not exactly follow a normal distribution, then $\hat{\mu}_i^{SEBP}$ will be biased. In this case, the bias correction factor can be applied such that $\hat{\mu}_i^{SEBP}$is given by:

$$\hat{\mu}_i^{SEBP} = \frac{1}{N_i}[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{SEBP\_c}]; i = 1,2 \ldots M, j = 1,2 \ldots n_i \quad (15)$$

where:
$\hat{y}_{ij}^{SEBP\_c} = \left(c_{ij}^{SEBP}\right)^{-1} e^{x_{ij}^T\hat{\beta}+z_{ij}\hat{u}_i+\frac{1}{2}\left(z_{ij}^2\hat{\tau}_i^2+\hat{\sigma}_e^2\right)}$; c_{ij}^{SEBP} is a bias correction factor which is obtained by second-order Taylor approximation.

References

1. Berg, E. and Chandra, H. (2014). Small Area Prediction for a Unit-Level Lognormal. Computational Statistics and Data Analysis, Vol.78, 159-175.
2. Handayani, D., Folmer, H., Kurnia, A., and Notodiputro, K.A. (2018), "The Spatial Empirical Bayes Predictor of The Small Area Mean for a Lognormal Variable of Interest and Spatially Correlated Random Effects", Empirical Economics, Vol.55 No. 1, 147-167.
3. Handayani, D. (2019), "Small Area Estimation Based on A Spatial Unit-Level Lognormal Model", PhD Dissertation (in Progress).
4. Rao, J.N.K and Molina, I. (2015). Small Area Estimation. New York: John Wiley and Sons.

Comparative performance of estimation maximisation and other known methods of residual estimators in structural equation models

Abdul-Aziz Abdul-Rahaman[1], Albert Luguterah[2], Bashiru Imoro Ibn Saeed[1]
[1] Kumasi Technical University
[2] University for Development Studies

Abstract
As the field of methodology has advanced, alternative estimation methods of residuals have been developed including regression method, Bartlett's method and Anderson-Rubin method. Somehow, their performance have experienced some level of challenges. Therefore, this study incorporated the estimation maximization approach and compared it with the other methods to identify the efficient method in estimating residuals under the structural equation model framework. The results showed that the strength of the existing methods are the weaknesses of EM method, and vice versa. It was therefore found from the comparative model fits information that the Bartlett's based method gave better residual parameter estimates over the regression-based method and the Anderson Rubin based method. However, the EM method gave better residual parameter estimates than the other three existing methods (i.e. the regression, Bartlett's and the Anderson Rubin based methods).

Keywords
Estimation maximization, Estimators, Structural equation modelling, Maximum likelihood

1. Introduction

Structural equation models (SEM) have been successfully utilised in different research areas, including educational studies (Miranda & Russell, 2011; Saçkes, 2014), clinical psychology (Little, 2013; Löfholm et al., 2014), developmental psychology (Geiser et al., 2010), organizational studies (Binnewies et al., 2010; Kiersch & Byrne, 2015; Mahlke et al., 2016), and multi-trait multimethod (MTMM) analysis (CarreteroDios et al., 2011). Approaches to SEM estimation may be described as covariance-based (e.g., ML) and component-based (e.g., PLS, GSCA), or as frequentist (e.g., ML, PLS, GSCA) and Bayesian (e.g., MCMC). Simply put, the primary distinction between covariance- and component-based estimation is that the former is suited to model testing and the latter is better suited to explaining variance and making predictions (Hulland et al., 2010; Tenenhaus, 2008). Although it is difficult to know whether or not theoretical models are specified correctly in applied research, simulation-based research has illustrated the impact of

misspecification on parameter recovery across estimation methods (Asparouhov & Muthén, 2010; Hwang, et al., 2010). The extent to which estimates are impacted by the misspecification of the model depends on design features and overall complexity of the model (Henseler, 2010). In the context of SEM, latent variables can be modeled as the cause of those observed values (Bollen & Lennox, 1991; Curtis & Jackson, 1962).

Until recent years, it was held that SEMs including measurement models were inappropriate for traditional ML approaches altogether (Chin, 1998; Hwang & Takane, 2004; Ringle et al., 2009). Several estimation methods and variations of those methods have been developed and applied to SEMs, including maximum likelihood (ML), and ML with robust standard errors (Muthén & Muthén, 1998-2010), generalized least squares (GLS), and weighted least squares (WLS). However, all of these methods are known to perform poorly under some conditions. Specifically, ML and WLS typically fail to produce accurate parameter estimates when applied to small samples (Hu et al., 1992 & Olsson et al., 2000); the more precise estimates produced by MLR are generally restricted to estimates of standard errors instead of path coefficients. In response to the limitations of these, additional estimation approaches have been applied to the estimation of SEMs (Wold, 1975; Hwang & Takane, 2004; Kline, 2011). The most common estimation method used with SEM is maximum likelihood (Hoyle, 2000). ML has been studied across myriad contexts and data conditions, and its limitations are well documented. One context in which ML does not perform well is in the presence of small samples (Kline, 2011). As the field of methodology has advanced, alternative estimation methods have developed and include generalized least squares, weighted least squares, PLS, GSCA, and MCMC approaches (Henseler, 2012; Hwang et al., 2010; Hwang Malhotra et al., 2010). Although estimation methods other than those described here have been developed for use with SEMs when the assumptions of ML are violated (robust ML, weighted least squares), it is not feasible to compare and evaluate the performance of all such alternatives in a single study. Thus, the present study will focus on the comparative performance of regression method, Bartlett's, Anderson-Rubin and EM methods because they represent diverse and promising approaches for addressing the problem of estimating residuals in SEM.

2. Methodology

In order to apply residual estimators in estimating the residuals of both measurement and latent variables, a recursive model with a mediation component was adopted from Hildreth (2013)

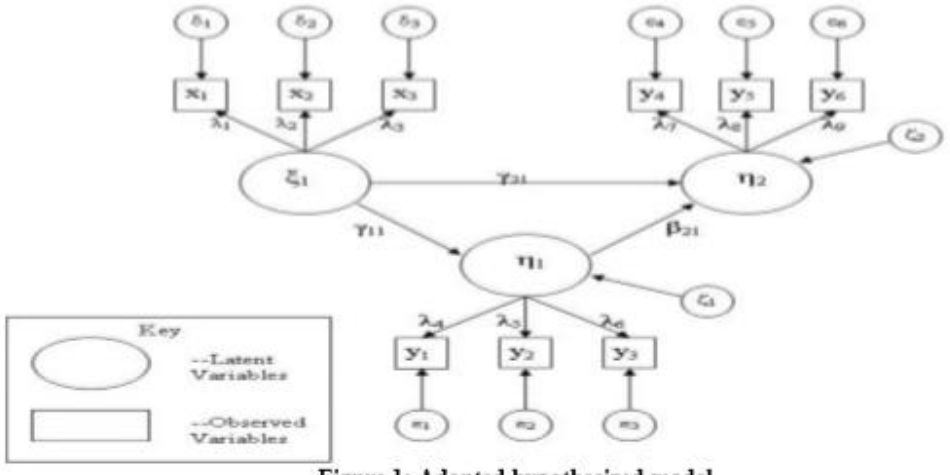


Figure 1: Adopted hypothesized model

2.1 Residual Estimators

Three residual estimators, comprising regression, Bartlett's and the Anderson-Rubin methods, in SEM have been proposed in the past. This study incorporates the EM method within the SEM framework and seeks to compare it against the other known residual estimators.

2.1.1 Regression Method

The most popular choice to use for the weight matrix W is based on the work of Thurstone (1935) who used the principles of least squares to derive W. Consequently, this method is frequently referred to as the regression method. Under this method, W is chosen such that ϵ

$$Tr\left[E\left\{\left(L_i - L_i\right)\left(L_i - L_i\right)'\right\}\right]$$

$$W_r = \sum_{LL} \Lambda' \sum_{zz}^{-1}$$

Where $\sum_{LL}$ is the (m+n)×(m+n) population covariance matrix of Li such that

$$\sum_{LL} = \begin{bmatrix} \sum_{\eta\eta} & \left(I - B\right)^{-1} \Gamma\Phi \\ \Phi\Gamma'\left(I - B\right)^{-\mathrm{T}} & \Phi \end{bmatrix}$$

And $\sum_{zz}^{-1}$ is the inverse of the population covariance matrix of z_i. Bollen and Arminger (1991) and Sanchez et al. (2009) use this weight matrix in the construction of their residual estimators.

2.2.2 Bartlett's Method

Another popular choice to use for the weight matrix is referred to as Bartlett's method due to Bartlett (1937) who derived the weight matrix using the principles of weighted least squares. Under this method, **W** is chosen such that

$$Tr\left[E\left\{\left[\sum_{vv}^{-1/2}\Lambda\left(L_i-L_i\right)\right]\left[\sum_{vv}^{-1/2}\Lambda\left(L_i-L_i\right)\right]'\right\}\right]$$

is minimized (McDonald and Burr, 1967). This leads to the weight matrix

$$W_b=\left(\Lambda'\sum_{vv}^{-1}\Lambda\right)^{-1}\Lambda'\sum_{vv}^{-1} \quad (2)$$

The estimator was also employed by Bollen and Arminger (1991).

2.2.3 Anderson-Rubin Method

The third, and perhaps least popular, choice for **W** was developed by Anderson and Rubin (1956) through an extension of Bartlett's method. This method is also derived using the principles of weighted least squares under the constraint of an orthogonal factor model. Under this method Equation (2) is minimized subject to the condition that

$$E[L_iL_i']=I$$

This leads to the weight matrix

$$W_{ar}=A^{-1}\Lambda'\sum_{vv}^{-1} \quad (3)$$

Where $A^2=(\Lambda'\Sigma_{vv}^{-1}\Sigma_{zz}\Sigma_{vv}^{-1}\Lambda)$. In practice, an orthogonal factor model is not realistic for SEM as the factors are expected to be correlated to one another. However, for completeness, this estimator is considered in this dissertation. Only one of the previous studies on residuals in SEM have examined the use of the Anderson-Rubin method-based estimator.

In practice the sample weight matrices W_r, W_b, and W_{ar} are used to obtain the estimated (unstandardized) residuals (Hildreth, 2013).

i. The EM Algorithm

In contrast to the aforementioned residual estimators, the EM algorithm, which utilizes a two-step iterative procedure, provides a ML estimate of the covariance matrix and mean vector that can, in turn, be used as input for further modelling. Suppose we have a model for the complete data Y, with associated density $f\,Y(/\theta)$, where $\theta=(\theta_1,\ldots,\theta_d)$ is the unknown parameter. We write $Y=(Y_{obs},Y_{mis})$, where Y_{obs} represents the observed part of Y and Y_{mis} denotes the missing values. The EM algorithm finds the value of θ,θ^* that maximizes $f(Y_{obs}/\theta)$, that is, the MLE for θ based on the observed data Y_{obs}. The EM algorithm starts with an initial value $\theta^{(0)}$. Letting $\theta^{(t)}$ be the estimate θ at the ith iteration, iteration (t +1) of EM is as follows; E step: Find the expected complete-data log-likelihood if θ were $\theta^{(t)}$:

$$Q\left(\theta/\theta^{(t)}\right)=\int L(\theta/Y)f\left(Y_{mis}/Y_{obs},\theta=\theta^{(t)}\right)dY_{mis} \quad (4)$$

Where $L(\theta/Y)=\log f(Y/\theta)$

M step: Determine $\theta^{(t+1)}$ by maximizing this expected log-likelihood:

$$Q\left(\theta^{(t+1)} / \theta^{(t)}\right) \geq Q\left(\theta / \theta^{(t)}\right) \text{ for all } \theta \quad (5)$$

The M step of EM algorithm is easy to implement in broad classes of problems, such as in exponential families, since it uses the identical computational method as ML estimation from $L(\theta / Y)$.

3. Result

For the purposes of simplifying and organizing the presentation of results, the residual parameters presented and discussed here are within the context of the four categories of estimators being compared.

Table 1. Parameter Estimates and Standard Errors of Residual Estimators

Parameter	Regression method	Bartlett's method	Anderson Rubin method	EM method
δ_1	0.598 (0.12)	0.599 (0.10)	0.600 (0.10)	0.600 (0.09)
δ_2	0.648 (0.14)	0.648 (0.13)	0.649 (0.14)	0.650 (0.14)
δ_3	0.699 (0.15)	0.701 (0.13)	0.703 (0.12)	0.700 (0.15)
ϵ_1	0.636 (0.11)	0.637 (0.10)	0.639 (0.10)	0.641 (0.09)
ϵ_2	0.572 (0.09)	0.573 (0.09)	0.574 (0.09)	0.578 (0.09)
ϵ_3	0.504 (0.09)	0.505 (0.09)	0.507 (0.08)	0.510 (0.08)
ζ_1	0.407 (0.24)	0.418 (0.19)	0.429 (0.16)	0.432 (0.23)
Test				
$\chi 2$	28.29	25.29	26.82	57.80
RMSEA	0.040	0.026	0.034	0.023
p	0.205	0.335	0.264	0.001
SRMR	0.041	0.037	0.038	0.020
CFI	0.989	0.994	0.992	0.984
AIC	268.466	259.516	264.692	246.317

From Table 1, the Regression method yielded fit indices of χ^2 (df =23, N=145) = 28.29, p=0.205, RMSEA=0.040, CFI=0.989, SRMR = 0.041 for the model in terms of the residual parameter estimates. Using the Bartlett's regression-based method, it was observed that χ^2 (df =23, N=145) = 25.29, p=0.335, RMSEA=0.026, CFI=0.994, SRMR=0.037 which were close to the values with regression-based method. SRMR is not yet obtainable as it does not directly depend on χ^2. Using the Anderson Rubin based method as implemented with the same CFA model, the following were obtained: χ^2 (df =23, N=145) = 26.82, p=0.264, RMSEA=0.034, CFI=0.992, SRMR=0.038. The AIC preferred the Bartlett's method over the Regression and Anderson Rubin with differences of 259.516, 268.466 and 264.692 respectively, indicating some evidence for slightly heavier tails in the sample distributions even without EM method. However, the fit information and parameter estimates (as shown in Table 1) under all three methods were similar, so the choice could be deeming to be trivial among these three existing residual estimator methods. However, the EM method was subsequently applied to modify the

estimation method adopted by the first three residual estimators which yielded different residual parameter estimates. The EM method gave χ^2 (df =23, N=145) = 57.80, p < 0.001, RMSEA=0.023, CFI=0.984, SRMR=0.020. Although both χ^2 and RMSEA indicated worse model fit, the impact on CFI was small and SRMR actually indicated better model fit compared to the other three existing methods. Therefore, it implies that with these methods, the model fit indices gave ambiguous fit information. As a result, the standard errors associated with the parameters and the comparative fit information would be preferable in understanding which specific residual estimator method gave a better parameter estimate. Thus, the fit information using EM method was comparable with the other three existing methods. Hence the AIC, BIC, and CAIC all strongly favoured the EM method over these other methods. Moreover, the parameter estimates in Table 1 above, indicated that whereas estimates were strongly close for these existing methods, they were more robust with Bartlett's and in particular the EM method.

4. Conclusion

In conclusion the strength of the existing methods are the weaknesses of EM method, and vice versa. It was therefore found from the comparative model fits information, by comparing among the three existing residual estimators, that the Bartlett's based method gave better residual parameter estimates over the regression-based method and the Anderson Rubin based method. However, the EM method gave better residual parameter estimates than the other three existing methods (i.e. the regression, Bartlett's and the Anderson Rubin based methods).

References

1. Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. Proceedings of the third Berkeley Symposium on mathematical statistics and probability, 5, 111-150.
2. Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Retrieved December 18, 2012 from http://www.statmodel.com
3. Bartlett, M. S. (1937). The statistical conception of mental factors. British Journal of Psychology, 28, 97-104.
4. Binnewies, C., Sonnentag, S., & Mojza, E. J. (2010). Recovery during the weekend and fluctuations in weekly job performance: A week-level study examining intra-individual relationships. Journal of Occupational and Organizational Psychology, 83(2), 419–441.
5. Bollen, K. A. & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. Sociological Methodology, 21, 235-262.
6. Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110(2), 305-314.
7. Carretero-Dios, H., Eid, M., & Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: An application to the validation of the state trait cheerfulness inventory. Journal of Research in Personality, 45(2), 153–164
8. Chin, W. W. (1998). Issues and opinion on structural equation modeling. Management Information Systems Quarterly, 22(1).
9. Curtis, R. F., & Jackson, E. F. (1962). Multiple indicators in survey research. American Journal of Sociology, 68, 195-204.
10. Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. Developmental Psychology, 46(1), 29–45. doi: 10.1037/a0017888
11. Henseler, J. (2012). Why generalized structured component analysis is not universally preferable to structural equation modeling. Journal of the Academy of Marketing Science, 40, 402-413.
12. Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm. Computational Statistics, 25, 107-120.
13. Hildreth L. (2013). Residual analysis for structural equation modeling. Graduate theses and dissertation. Paper 13400
14. Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling. New York: Academic Press.

15. Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? Psychological Bulletin, 112(2), 351-362.
16. Hulland, J., Ryan, M. J., & Rayner, R. K. (2010). Modeling customer satisfaction: A comparative performance evaluation of covariance structure analysis versus partial least squares. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), Handbook of Partial Least Squares. Berlin: SpringerVerlag.
17. Hwang, H., Malhotra, N. K., Kim, Y., Tomiuk, & Hong (2010). A comparative study on parameter recovery of three approaches to structural equation modeling. Journal of Marketing Research, 47, 699-712.
18. Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. Psychometrika, 69(1), 81-99.
19. Kiersch, C. E., & Byrne, Z. S. (2015). Is being authentic being fair? Multilevel examination of authentic leadership, justice, and employee outcomes. Journal of Leadership & Organizational Studies, 22(3), 292–303. doi: 10.1177/1548051815570035
20. Kline, R. B. (2011). Principles and Practice of Structural Equation Modeling (3rd Edition). New York, NY: The Guilford Press.
21. Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the life skills profile–16. Psychological Assessment, 25(3), 810–825. doi: 10.1037/a0032574
22. Löfholm, C. A., Eichas, K., & Sundell, K. (2014). The Swedish implementation of multisystemic therapy for adolescents: Does treatment experience predict treatment adherence? Journal of Clinical Child & Adolescent Psychology, 43(4), 643–655. doi: 10.1080/15374416.2014.883926
23. Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R., & Brodbeck, F. (2016). A multilevel CFAMTMM approach for multisource feedback instruments: Presentation and application of a new statistical model. Structural Equation Modeling: A Multidisciplinary Journal, 23, 91–110. doi: 10.1080/10705511.2014.9901533
24. Miranda, H., & Russell, M. (2011). Predictors of teacher directed student use of technology in elementary classrooms: A multilevel SEM approach using data from the USEIT study. Journal of Research on Technology in Education, 43(4), 301–323. doi: 10.1080/15391523.2011.10782574
25. McDonald, R. P. & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. Psychometrika, 32(4), 381-401.
26. Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus user's guide. (7th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.
27. Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000) The performance of ML, GLS, and WLS estimation in structural equation

modeling under conditions of misspecification and nonnormality. Structural Equation Modeling, 7(4), 557-595.
28. Ringle, C. M., Götz, O., Wetzels, M., & Wilson, B. (2009). On the use of formative measurement specifications in structural equation modeling: A Monte Carlo simulation study to compare covariance-based and partial least squares model estimation methodologies. In METEOR Research Memoranda (RM/09/014): Maastricht University.
29. Saçkes, M. (2014). How often do early childhood teachers teach science concepts? determinants of the frequency of science teaching in kindergarten. European Early Childhood Education Research Journal, 22(2), 169–184. doi: 10.1080/1350293X.2012.704305
30. Sanchez, B. N., Houseman, E. A. & Ryan, L. M. (2009). Residual-based diagnostics for structural equation models. Biometrics, 65, 104-115.
31. Tenenhaus, M. (2008). Component-based structural equation modelling. Total Quality Management, 19, 871-886.
32. Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), Quantitative sociology: International perspectives on mathematical and statistical modeling (pp. 307-357). New York: Academic.

Space dependent ordinal pattern probabilities in time series

Ines Muenker, Alexander Schnurr
Siegen University

Abstract

We consider the ordinal information of n consecutive points in a given data set. This information is encoded in a permutation which is called ordinal pattern. Applications of ordinal patterns include statistical estimation of the Hurst parameter in long-range dependent time series, calculation of the Kolmogorov-Sinai entropy in dynamical systems as well as tests for structural breaks. Recently it has been shown that the probabilities of ordinal patterns in stationary time series are different, if – instead of the whole data set – only extremal events are considered. Here, we analyze in an empirical study, whether (and how) the current area of the data set influences the appearance of certain patterns. It turns out that in the discharge data we consider, we indeed find different pattern frequencies for different level sets.

Keywords

Order structure; model free data analysis; permutation; long-range dependence; hydrology

1. Introduction

Ordinal patterns were invented by Bandt and Pompe (2002) in order to analyze the chaotic behavior of dynamical systems. They have also been used to analyze – mostly model-free – data from biology, medicine, finance and hydrology (cf. Bandt and Shiha (2007), Keller et al. (2007) and Keller and Sinn (2005)).

The concept works as follows: for n consecutive data points, there are $n!$ possibilities how they can be ordered (if ties are excluded). These possibilities are called ordinal patterns. We encode them by writing down a permutation as follows: first the index of the data point with the highest value, then the index of the data point with the second highest value and so on. The mathematical definition can be found in Section 2. In Figure 1 some of the 24 patterns of length 4 are showcased.

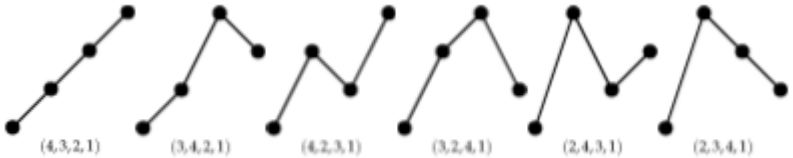


Figure 1: some patterns of length 4 – structure and corresponding permutation

Over the last decade ordinal patterns have been used in different areas of statistics: Sinn and Keller (2011) have used their relative frequencies in order to estimate the Hurst parameter of a fractional Brownian motion. Sinn et al. (2012) have described a method to detect change points in a given data set. Schnurr (2014) and Schnurr and Dehling (2017) have analyzed the dependence structure between two time series by means of ordinal patterns. The relationship between ordinal pattern dependence and other kinds of dependence has been analyzed in Schnurr and Muenker (2017).

Recently Oesting and Schnurr (2019) have tackled the question which patterns can be found in clusters of extremal events. This is the starting point of the present study. We ask ourselves the question whether the probabilities of the patterns which appear depend on the current state of the process. To this end we separate the state space of a stationary time series into blocks which are analyzed one-by-one.

The paper is organized as follows: the ordinal pattern analysis is described in Section 2 along with some statistical issues related to our data example. The latter one is described in Section 3, where the reader finds in addition the results of our study. The discussion and outlook in Section 4 round out the paper.

2. Methodology

Our method of choice is the so called *ordinal pattern analysis*. Let us begin with the formal definition of ordinal patterns (in this section we follow closely Schnurr and Muenker (2017)): let $n \in \mathbb{N} = \{1,2,3, \dots\}$ be a positive integer and $x = (x_{/}, x_0, \dots, x_1) \in \mathbb{R}^1$ a vector. Furthermore, let S_1 denote the space of permutations of length n, that is,

$S_1 := \{\pi \in \mathbb{N}^1: 1 \leq \pi_9 \leq n$ and $\pi_9 \neq \pi_;$ whenever $i \neq k\}$.

The *ordinal pattern* of x is the unique permutation

$$\prod(\boldsymbol{x}) = (\pi_{1,}, \dots, \pi_{\pi}) \in S_n$$

such that

(i) $x_{\pi_1} \geq x_{\pi_2} \dots \geq x_{\pi_n}$ and
(ii) $\pi_{j-1} > \pi_j$ if $x_{\pi_{j-1}} = x_{\pi_j}$ for $j \in \{2,3, \dots, n\}$.

The latter is to make a decision in the case of ties. Dealing with real world data it is sometimes more appropriate to put a small noise on the data set in order to get rid of the ties. Otherwise, following (ii) one overestimates the probabilities of certain patterns. In the data set we consider in Section 3 it is most likely that the ties we encounter are due to rounding.

One advantage of ordinal patterns is that the whole ordinal information, that is, the up-and-down behaviour is kept, while the metrical information is not considered. Hence, the method is stable under monotone transformations

and robust under measurement errors. In particular in the context of hydrological data it is an additional advantage that the method is robust with respect to some kinds of structural breaks like 'shift of mean'. Some other methods (autocorrelation) are sensitive to these shifts which appear e.g. in discharge data, if a dam is built. The variety of applications we have mentioned above shows that the ordinal information is sufficient for several situations.

We analyze the probabilities of ordinal patterns respectively their relative frequencies in given data sets of length m. To this end we use a moving window approach. That is, for every $\pi \in S_n$ we use the estimator:

$$\widehat{p_\pi} = \frac{1}{m-n+1} \sum_{i=1}^{m-n+1} 1_{\{\Pi(X_i,\dots,X_{i+n-1})=\pi\}}$$

For the probability

$$p_\pi := P(\Pi(X_1, \dots, X_n) = \pi).$$

Here, and in the following we tacitly assume that the model we are considering in the background satisfies standard assumptions like being stationary as well as some ergodicity condition (cf. Section 4) in order to ensure that $\widehat{p\pi}$ converges to p_π.

In our considerations m always denotes the length of the data set while n is the length of the pattern. In our empirical study, we have chosen n = 3. There is always a trade-off between the information contained in the pattern and the number of parameters under consideration: since for n data points we have to consider $n!$ parameters (the probability of each pattern), it is better to stick with some small number although a larger n results in a more detailed knowledge of the ordinal structure of the process. In theoretical papers the authors even let n tend to infinity. This is mostly done in the context of dynamical systems (cf. Bandt et al. (2002)).

3. Result

We are considering daily discharge data from the river Rhine measured in Cologne from November 1st 1816 to October 30th 2013 which was extracted from the webpage of GRDC (Global Runoff Data Center).

Since the winter months December, January and February are known to fulfil the demanded property of stationarity better than the all year long data we will consider them additionally. So after extracting the winter months for each of the 197 years, we have transformed the two data sets (all-year data and only-winter data) to standard-normal distributions. It is a standing assumption in the literature that afterwards we can assume that the process $(\widetilde{x_t})$, t = 1,2, … is Gaussian (Chilès and Delfiner(1999)). This yields the nice property that the increment process we will consider later on is also Gaussian. Since ordinal patterns are not affected by monotone transformations this

transformation does not affect our results. In the given data set, one often finds consecutive data points having an equal value. This yields a certain ordinal pattern, for example Π(810,810, 810) = (2,1,0) in the case n = 3 by the convention (ii) above. In our case this would distort the results. We hence disturbed the data with a small white noise process with standard deviation 0.01 to avoid this case. In a way this is natural, since the data has not been measured accurately and the ties are a result of rounding rather than being real ties.

First we ran a Dickey-Fuller test for stationarity implemented in GNU R to justify our assumption of stationarity for both data sets (all-year and only-winter). After that we estimated the Hurst parameter, which describes the degree of dependence within a time series, with an estimator implemented also in GNU R ('hurstexp'). If $H \in (0.5,1)$ the time series exhibits long-range dependent, if $H < 0.5$ it is short-range dependent. In the all-year data, we obtained H = 0.72 and in the winter data, for all 197 years, we got values between 0.7528 and 0.7979 with mean 0.7528, so the conjecture of long-range dependence of the given data is satisfied.

Let us first focus on the distribution of ordinal patterns in the winter data set. In order to do so we have divided the values of the data into disjoint intervals, namely $(-\infty,-1.8),[-1.8,0),[0,1.8)$ and $[1.8,\infty)$(displayed in Figure 2) and analyzed the frequency of the patterns which appear in each of these intervals.

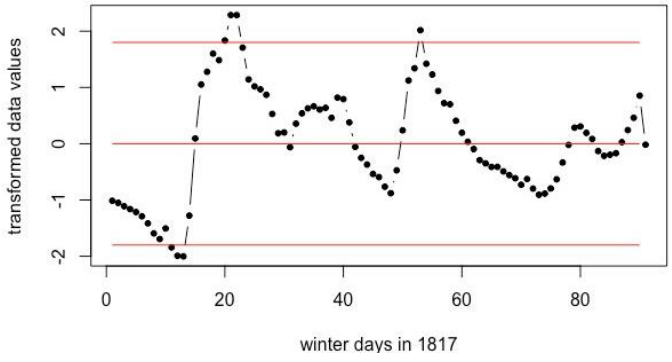


Figure 2 : winter data of the first year

Due to the larger range of data values we chose a slightly different division dealing with the all-year data, namely $(-\infty, -3),[-3, -1),[-1,1),[1,3)$ and $[3, \infty)$. Our results are shown in the following tables, where in both tables the first column describes the relative frequency of each ordinal pattern in the entire data set.

	all data	< -1.8	[-1.8,0)	[0,1.8)	> 1.8
(3,2,1)	25.96%	10.17%	18.59%	23.22%	23.26%
(1,2,3)	53.91%	29.67%	58.59%	57.60%	26.57%
(2,1,3)	4.97%	2.54%	4.92%	5.37%	20.30%
(3,1,2)	5.51%	27.12%	7.12%	4.26%	2.95%
(2,3,1)	5.15%	4.24%	4.10%	6.46%	25.09%
(1,3,2)	4.50%	26.27%	6.68%	3.10%	1.48%

Table 1 : relative frequencies of the ordinal patterns in the winter data

	all data	< -3	[-3,-1)	[-1,1)	[1,3)	> 3
(3,2,1)	25.76%	28.30%	20.85%	24.08%	26.04%	27.08%
(1,2,3)	50.54%	20.00%	46.71%	52.33%	50.00%	29.17%
(2,1,3)	6.06%	8.30%	8.20%	5.70%	7.91%	20.83%
(3,1,2)	5.95%	15.00%	8.50%	6.24%	3.87%	0.00%
(2,3,1)	5.78%	13.30%	5.62%	5.62%	8.50%	18.75%
(1,3,2)	5.89%	15.00%	10.00%	6.03%	3.67%	4.17%

Table 2 : relative frequencies of the ordinal patterns in the all-year data

We restrict ourselves here to the interpretation of the ordinal pattern distributions of the winter data because they exhibit a better illustration of the effects we want to describe. Therefore, we focus on Table 1. First, we observe that there is a large difference between the relative frequencies of the two patterns that describe a monotone behaviour of the time series, more precisely, that the relative frequency of the pattern (1,2,3) is nearly twice the value of (3,2,1), while the values for the four remaining patterns are close to being equal. A heuristic explanation considering that we are dealing with discharge data might be that, especially in the winter months, if we have melting snow we might have a fast increase but a slow decrease. If we now look at the distributions of the patterns in the different intervals, we first observe that in the middle regions namely $[-1.8,0)$ and $[0,1.8)$ we get a very similar distribution as in the entire data set. This is easy to explain because within these areas we find most of the data points as one can easily see in Figure 2.

However, in the extremes, meaning the values smaller than -1.8 and larger than 1.8 we obtain a different distribution. The first point to notice is that the frequencies of the patterns (3,2,1) and (1,2,3) are getting closer to each other, in particular in the case where the data values are larger than 1.8. In the last four rows we get the interesting phenomenon that the almost equal frequencies from before now change to two almost equally large values and two almost equally small values. In the case where the values are smaller than

-1.8 the patterns (3,1,2) and (1,3,2) are getting a relative frequency of 0.2712 resp. 0.2627, while (2,1,3) and (2,3,1) only get 0.0254 resp. 0.0454. When the smoke has settled this phenomenon can easily be explained in the geometric interpretation of these ordinal pattern. In the case where the data values are below a certain threshold one would expect the smallest entry of a data vector with length n=3 to be the second entry, and this is exactly what the patterns (3,1,2) and (1,3,2) describe. Hence it is only natural that in the other extremal event, namely when one only considers data values above a certain threshold (in our case 1.8) then one observes the same phenomenon but only for the respective space reversion of the patterns discussed above. This can be seen in the last column of Table 1 concerning the winter data above.

4. Discussion and Conclusion

The most important result – which opens a new perspective on ordinal patterns – is, that we indeed find the phenomenon of state-space-dependence of ordinal pattern frequencies in a real-world data set. One advantage of the method provided here is that it is (at least in the first step) model free. In order to prove limit theorems, one has to make some assumptions on the theoretical model in the background, but at first one does not have to care for the model. However, let us shortly comment on the theoretical background of the pilot study described above:

Since ordinal patterns of $(X_i)_{i\in\mathbb{N}_0}$ are uniquely determined by the increment process $(Y_i)_{i\in\mathbb{N}}$ where $Y_i = X_i - X_{i-1}, i = 1,2,\dots$ we can now define the mapping (cf. Keller and Sinn (2011))

$$\widetilde{\Pi}(Y_2, \dots, Y_n) := \Pi(0, Y_2, Y_2 + Y_3, \dots, Y_2 + \dots + Y_n) = \Pi(X_1, \dots, X_n)$$

and hence rewrite the estimator above to

$$\widehat{p_\pi} = \frac{1}{m-n+1}\sum_{i=1}^{m-n+1} 1_{\{\Pi(X_i,\dots,X_{i+n-1})=\pi\}} = \frac{1}{m-n+1}\sum_{i=1}^{m-n+1} 1_{\{\widetilde{\Pi}(Y_{i+1},\dots,Y_{i+n-1})=\pi\}}$$

Betken and Wendler (2019+) have shown that there is a large class of long-range dependent processes which satisfy a slightly technical condition such that the increment process of these processes is shortrange dependent. This basically means that the autocorrelations of this process are summable. The estimator for the Hurst parameter yields H=0.35 for the increment process of the all-year data which confirms the theoretical result above. We can now apply the results of Arcones (1994), Theorem 4 and hence get asymptotic normality of our estimator. Let us mention that Keller and Sinn (2011) provide an estimator for probabilities of ordinal patterns that has better statistical properties than the one we have used here. However, in order to use this estimator, one needs that the probability of each pattern remains unchanged

if the pattern is reversed in space and/or time. In our case that means we should get almost equal relative frequencies for all patterns in $P/$={(1,2,3), (3,2,1)} as well as in $P0$={(1,3,2), (2,3,1), (3,1,2), (2,1,3)}. Unfortunately, it is easily seen in the tables above that this assumption is not fulfilled in the empirical results here. Anyway, it is interesting to mention that if we study the relative frequencies of ordinal patterns in the integrated process of the all-year data this assumption is satisfied (as one can see in Table 3).

	Add data
{3,2,1}	48,35%
{1,2,3}	47,73%
{2,1,3}	0,96%
{3,1,2}	1,05%
{2,3,1}	1,00%
{1,3,2}	0,92%

Table 3 : relative frequencies of ordinal patterns in the integrated all-year data

So if one is interested in the distributions of the ordinal patterns of the integrated process (which are not the same distributions as in the integrated process of the original data set before transformation), one could apply the improved estimator mentioned above

$$q_\pi = \frac{1}{2}\sum_{\pi\in P_1}\widehat{p_\pi} \quad \text{for } \pi \in P_1 \qquad \text{and} \qquad q_\pi = \frac{1}{4}\sum_{\pi\in P_2}\widehat{p_\pi} \quad \text{for } \pi \in P_2$$

and would get asymptotic normality here, too, since the estimated Hurst parameter of the all-year data (which plays the role of the increment process here) is smaller than 0.75. Concerning the integrated process we could not conclude asymptotic normality of the estimator $\widehat{p_\pi}$ by now, because in this case the increment process would be long-range dependent and the asymptotic behaviour in this case is ongoing research.

Finally, let us sketch some applications: In the context of model selection ordinal patterns can be very helpful. A good model should (at least) match the ordinal structure of the data sets under consideration. In a continuous-time Markovian setting, our state-space dependent analysis can be used to decide whether Levy processes are a useful model (homogeneous in space) or more complicated models are appropriate (like Feller processes). In the future we would like to understand the theoretical background better. Another long time goal of the method presented here is to predict the length of extremal events or the remaining time within one 'regime' by analyzing the encountered patterns.

Acknowledgement

Financial support by the DFG (German Science Council) for the project, Ordinal-Pattern-Dependence: Grenzwertsaetze und Strukturbrueche im langzeitabhaengigen Fall mit Anwendungen in Hydrologie, Medizin und Finanzmathematik' (SCHN 1231/3-1) is gratefully acknowledged. In addition we would like to thank Albert Piek (Luebeck) for providing us with Figure 1.

References

1. Arcones, M.A. (1994): Limit Theorems for Nonlinear Functionals of a Stationary Gaussian Sequence of Vectors. Ann. Prob. 22(4), 2242-2274.
2. Bandt, C., Keller, G. and Pompe, B. (2002): Entropy of interval maps via permutations. Nonlinearity 15, 1595-1602.
3. Bandt, C. and Pompe, B. (2002): Permutation entropy: A natural complexity measure for time series. Phys. Rev. Lett., 88 174102 (4 pages).
4. Bandt, C. and Shiha, F. (2007): Order Patterns in Time Series. J. Time Ser. Anal., 28, 646-665.
5. Betken, A., Wendler, M. (2019+): Subsampling for General Statistics under Long Range Dependence. To appear in Statistica Sinica.
6. Brockwell, J. and Davis, R.A. (1991): Time Series: Theory and Methods. Springer, New York.
7. Chilès, J.P., Delfiner, P. (1999): Geostatistics - Modeling Spatial Uncertainty. Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.
8. S. Fischer, Schumann, A. and Schnurr, A. (2017): Ordinal Pattern Dependence between Hydrological Time Series, Journal of Hydrology 548, 536-551.
9. Ibragimov, I.A. and Linnik, Yu.V. (1971): Independent and stationary sequences of random variables. Wolters-Noordhoff Groningen.
10. Keller, K., Sinn, M. and Emonds, J. (2007): Time Series from the Ordinal Viewpoint. Stochastics and Dynamics, 2, 247-272.
11. Keller, K. and M. Sinn (2011): Estimation of ordinal pattern probabilities in Gaussian processes with stationary increments, Comp. Stat. Data Anal., 55, 1781-1790.
12. Keller, K. and Sinn, M. (2005): Ordinal Analysis of Time Series. Physica A, 356, 114-120.
13. Oesting, M. and Schnurr, A. (2019): Ordinal Patterns in Clusters of Subsequent Extremes of Regularly Varying Time Series. Preprint.
14. Schnurr, A. (2014): An Ordinal Pattern Approach to Detect and to Model Leverage E_ects and Dependence Structures Between Financial Time Series. Stat. Papers 55(4) (2014), 919-931.
15. Schnurr, A. and Dehling, H. (2017): Testing for Structural Breaks via Ordinal Pattern Dependence. JASA, 112(518), 706-720.

16. Schnurr, A. and Muenker, I. (2017): Ordinal Pattern Dependence in Contrast to Classical Concepts of Dependence. Proceedings of the 61st World Statistics Congress 2017.
17. Sinn, M., Ghodsi, A. and Keller, K. (2012): Detecting Change-Points in Time Series by Maximum Mean Discrepancy of Ordinal Pattern Distributions. In: Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI), 786-794.

Detecting outlier in a circular regression model – A review

Intan Mastura Ramlee, Safwati Ibrahim
Institute of Engineering Mathematics, University Malaysia Perlis, Pauh Putra Main Campus, 02600 Arau, Perlis

Abstract

Presently, circular data is very relevant and important application technique in many fields such as Biology, Medicine and others. Whereas one type of data direction is a data circular. In this case, author have a tendency to study and explore in detail about circular regression model. In this paper aim to review the outlier detection methodologies in circular regression model based on 11 articles. In general, this paper performs a survey of circular regression from 2011 to 2018 in order to see the trend of current study. Here, we concentrate the attention on the methodologies of identifying outlier in this model. This survey of circular regression model in which many interesting properties and is good enough to detect the occurrence of outlier. Through the survey may highlight the significant of methodologies to detect outliers in circular regression model and provide guideline for future work to look into the research gap.

Keywords

Circular regression model, outlier detection, statistical analysis

1. Introduction

Statistical analysis is the study the relationship between the independent variable and dependent variable in regression analysis. Research on circular variable in regression model has long tradition since four decade ago. The field of the circular regression is referred of relationship when both explanatory and response variable are circular. Circular data arise in many different fields such as biology, physics, geology, medical science and others. Thus, a practical need for circular regression can been in real-life problem such the wind direction and the direction of movement of clouds, the arrival of patient (24 hours) in the emergency room in a hospital and others [1].

Data circular is types of data direction whereas the data measurement take range in degree direction or unit time [2]. The analysis of circular data is the measurement value of data are repeated periodically. Therefore, the circular data inaccurate to analyse with linear statistic, so that circular data needs to be analysed by using circular statistics. As a result, the statistical software have been provided including Axis, DDSTP, Oriana, MATLAB and R/S language for circular data analysis.

A data will be easily to analyse when illustrated in a graph. According to [3], the presentation of circular data in the graphics is important in the analysis of circular data. The graphical form used for circular data is

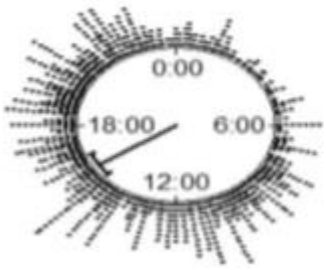


Figure 1: Transmit diagram

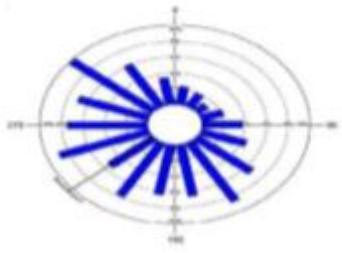
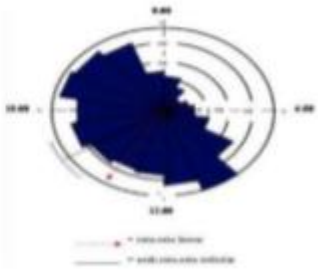


Figure 2: Cycle diagram



3: Rose diagram Figure

In the analysis of circular data, these focused on the descriptive and investigation to develop of descriptive measure and special characteristics of circular data [4, 5, 6, 7]. The circular descriptive measures are namely; the mean direction, the median direction and the sample circular standard deviation.

One of the common problem in circular regression modelling is an outlier. Outlier is defined as extreme values that deviate from other observation on set of data [8]. In other words, an outlier is an observation that diverge from overall pattern on a sample. Thus, it is important to detect and access the observation and estimate its impact on the proposed model [9].

2. Circular Regression

Analysis of circular regression have been proposed by a number of authors starting four decades ago. A circular regression equation distribution for the data is divided into three types, namely [10]

a) Circular – Linear Regression: The circular variable is an independent variable while the linear variable is a dependent variable. The model is given [11]

$$Y_i = M + \sum_i \beta_i x_i + g(t_i) + \varepsilon_i ,$$

where $g(t_i) = A_1 \cos(\omega t - \phi_1) + A_2 \cos(2\omega t - \phi_2) + \ldots + A_k \cos(k\omega t - \phi_k)$.

b) Linear – Circular Regression: The linear variable is an independent variable while the circular variable is a dependent variable. The model linear – circular regression can written as [12]

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}$$

c) Circular – Circular Regression: The circular variable is an independent variable while the circular variable is a dependent variable. The regression curve of the proposed regression model is defined by [12]

$$y = \beta_0 \frac{x + \beta_1}{1 + \overline{\beta_1} x} \in,$$

where β_0 and β_1 are complex parameters with $\beta_0 \in \Omega$ and $\beta_1 \in \mathbb{C}$.
Circular statistical used in data measurement in the form of direction and not the magnitude of the vector, where is expressed in angular size. Both of statistical technique and statistical distribution is to analyze random variable in form cycle there using trigonometric function.
[11] proposed a regression model to predict the mean direction of a circular response variable from a vector of linear covariates $X = (x_1, \dots, x_k)$.The proposed model is given by

$$\mu = \mu_0 + \sum_{j=1}^{p} \beta_j x_j$$

Where μ and β_j are unknown parameters and x_j is a linear covariate with $j = 1,\dots,p$.

3. Outlier Detection in Circular Regression Model

At this time, mostly authors are performed a research based on due to the bounded property of circular observation. Most of the paper publish today was concentrate on detecting outlier in circular data and circular regression model with one independent circular variable. The main aim here is to develop an outlier detection procedure in circular regression based on 11 papers that has been published in 2011 to 2018.

One of the methods to detecting outliers is the row deletion method. It investigates how the deletion of any row affect the residuals, the estimated coefficient, the estimate covariance structure of the coefficient as well as the predicted value such DFBETAs, DFFITs and COVRATIO. In this paper, we review some method for deleting outliers in circular regression method. The methods are listed in table 1 with name of researcher propose and solving in short.

Author/s	Ref	Objective Function	Proposed Method	Optimization
Alkasadi et al., 2018	[9]	DFBETAs statistic	Circular Regression Model	Multiple Circular Regression Model (MCRM)
Jayant Jha and Atanu Biswas, 2017	[14]	MCR 1 and MCR 2	Multiple Circular – Circular Regression Model	DM Circular Regression Model (MCRM)
Di et al., 2017	[15]	Single – linkage method		Down and Mardia Circular – Circular Regression Model
Alkasadi et al., 2016	[16]	COVRATIO statistic	Circular Regression Model	Multiple Circular Regression

				Model (MCRM)
Rambli et al., 2016	[17]	DMCEs statistic	Circular Regression Model	DM Circular Regression Model
Kim and SenGupta, 2016	[18]	Least circular mean – square estimation (LCMSE) and asymptotic properties of the LCMSE estimation	Multivariate – Multiple Circular Regression	Arc – tangent link model 1 and Arc – tangent link model 2
Abuzaid and Allahham, 2015	[19]	Wrapped Cauchy Error	Circular Regression	JS Circular Regression Model
Rambli et al., 2015	[20]	COVRATIO statistic	Circular Regression	DM Circular Regression Model
Ibrahim et al., 2013	[7]	COVRATIO statistic	Circular Regression	JS Circular Regression Model
Abuzaid et al., 2013	[21]	Mean circular error (MCE) statistic	Circular Regression	DM Circular Regression Model
Jurgen A. Doornik, 2011	[22]	Robust estimation	Circular Regression	Least trimmed Square

4. Discussion and Conclusion

At this instant, the studies of circular regression model have become a familiar area among authors to be explore. This study looks at outlier detection methodologies in circular regression model. There are many outlier detection methods have been used to detect and remove the rest of data in the literature and in practice. Also, most of them are attracted to solve this problem and intentions to achieve some objectives in their studies. This paper presents a survey of research method to detect outlier in circular regression model that cover up 11 paper from 2011 until 2018. The finding displays this study has contribute to the survey methodology development of statistic to detect outlier in circular regression model.

References

1. Mardia, K. V. & Jupp, P. E. (2000). Directional Statistics. New York: John Wiley & Sons.
2. Jammalamadaka, S. R. & A. SenGupta. (2001). Topic in Circular Statistics. London: World Scientifics Publishing.
3. Fisher, N. I. (1993). Statistical Analysis of Circular Data. Cambridge University Press.

4. Jammalamadaka, S. R. & Sarma, Y. R. (1993). Circular Regression. In: Matsusita. Statistical Science and Data Analysis. Utrecht: VSP.
5. Kato, S., Shimizu, K., & Shieh, G. S. (2008). A Circular – Circular Regression Model. Statistica Sinica, 18(2), 633 – 643.
6. Abuzaid, A. H., Mohamed. I. B., (2012). Boxplot for Circular Variable. Computational Statistics, 27(3), 381 – 392.
7. Ibrahim, S. (2013). Some Outlier Problem in a Circular Regression Model/PhD Thesis, University of Malaya.
8. Freeman, P. R. (1980). On the number of outliers in data from linear model. Trabajos de estadistica y de investigacion operative. 31(1), 349 – 365.
9. Alkasadi, A. N., Abuzaid, A. H., Ibrahim, S., & Yusoff, M. I. (2018). Outlier detection in multiple circulr regression model via DFBETAc statistic. International Journal of Applied Engineering. 13(11), 9083 – 9090.
10. Nurhab, M. I., Kurnia, A. & Sumertajaya, I. M. (2014). Circular Circular – Linear Regression Analysis of Order m in Circular Variable · and · against Linear Variable (Y). IOSR Journal of Mathematics. 10(4), 49 – 54.
11. SenGupta, A. & Ugwuowo, F. I. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data. Environ Ecol Stat. 13:299–309 DOI 10.1007/s10651-005-0013-1.
12. Sikaroudi, A. E. & Park. C. (2016). A Mixture of Linear-Linear Regression Models for LinearCircular Regression. Department of Industrial and Manufacturing Engineering, Florida State University, Tallahassee FL 32310.
13. Gould A. L. (1969). A Regression Technique for Angular Response. Biometrics, 25, 683 – 700. [14] Jaya, J. & Biswas, A. (2017). Multiple circular – circular regression. Statistical Modelling. 17(3), 1 – 30.
14. Di, N. F. M., Satari, S. Z., & Zakaria, R. (2017). Detection of different outlier scenarios in circular regression model using single – linkage method. IOP Conference Series: Journal of Physics: Conference Series 890.
15. Alkasadi, N. A., Ibrahim, S., Ramli, M. F., & Yusoff, M. I. (2016). A comparative study of outlier detection procedures in multiple circular regression. *AIP Conference Proceedings*, *1775*.
16. https://doi.org/10.1063/1.4965152.
17. Rambli, A., Abuzaid A. H. M., Mohamed, I. B., & Hussin, A. G. (2016). Procedure for detecting outlier in a circular regression model. PLoS ONE 11(4): e0153074. doi:10.1371/journal.pone.0153074.
18. Kim, S. & A. SenGupta. (2016). Multivariate – multiple circular regression. Journal of Statistical Computation and Simulation.

19. Abuzaid, A. H. & Allahham, N. R. (2015). Simple circular regression model assuming wrapped cauchy error. Pakistan Journal Statistics, 31(4), 385 – 398.
20. Rambli, A., Yunus, R. M., Mohamed, I. B., & Hussin, A. G. (2015). Outlier Detection in a Circular Regression Model. Sains Malaysiana, 44(7).
21. Abuzaid, A. H., Hussin, A. G., & Mohamed. I. B. (2013). Detection of outlier in simple circular regression models using the mean circular error statistic. Journal of Statistical Computation and Simulation, 83(2), 269 – 277.
22. Doornik, J. A. (2011). Robust estimation using least trimmed square.

Quantitative comparisonson pioneer – family resilience index Indonesia 2015

Rida Agustina, Amiek Chamami, Sapta Hastho Ponco
Statistics Indonesia

Abstract

In 2015, Statistics Indonesia and the Ministry of Women's Empowerment and Child Protection (MWECP) jointly prepared a measure of family resilience, which was later referred to as the Pioneer Family Resilience Index (P-FRI). P-FRI is structured based on five dimensions including the foundation of family legality and wholeness, physical resilience, economic resilience, social psychology resilience, and socio-cultural resilience. In the calculation, the P-FRI is measured composite by using the Analytic Hierarchy Process (AHP) method based on the assessment of family resilience experts on 24 P-FRI constituent indicators. There needs to be a statistically valid and reliable measurement to measure the level of family resilience in Indonesia. The aim of this study was to determine the quantitative comparison of family resilience index calculations. The analysis used is Structural Equation Modelling (SEM) with Latent Variable Score. The data used is the same as the data in calculating P-FRI using AHP. Samples are 34 provinces in Indonesia. The results of SEM analysis show that the valid and reliable P-FRI dimension is the basis of family legality and wholeness, physical resilience, and psychological social security. Of the three dimensions, only 6 indicators are valid and reliable and have a good fit model. There is no significant difference in the calculation of family resilience index using 6 indicators or 24 indicators.

Keywords

Quantitative comparisons; family resilience; structural equation modelling; latent variable score

1. Introduction

Family is the main foundation for building a social order system so that it can be said that family resilience is the basis of national security. Several ministries and community agencies stated that a measure of family resilience is needed to measure the strength of family resilience. In fact, there is no definite concept and measurement that is methodologically and generally applies to determine the level of family resilience in Indonesia and internationally.

In 2015, Statistics Indonesia and the Ministry of Women's Empowerment and Child Protection (MWECP) jointly prepared a measure of family resilience, which was then referred to as the Pioneer Family Resilience Index (P-FRI), as a

material for reviewing and assessing the level of family resilience in Indonesia (MWECP, 2016). The compilation of a measure of family resilience refers to the Minister of Women's Empowerment and Child Protection Regulation No. 6 of 2013 concerning the Implementation of Family Development which explains the concept of family resilience and welfare including the foundation of family legality and wholeness, physical resilience, economic resilience, social psychology resilience, and socio-cultural resilience. The five coverage forms dimensions in measuring the level of family resilience.

P-FRI is measured in a composite manner including multidimensional, multivariable, and multiindicators that require weighting dimensions, variables, and indicators. The method used to calculate the dimensions, variables and indicators of P-FRI is Analytic Hierarchy Process (AHP), which is based on the assessment of experts.

Sixbey's (2005) and Euis (2003) research show that significant dimensions in calculating family resilience are: systemic belief, organizational pattern, communication/problem solving, physical endurance, psychological endurance, and social resilience. While the level of family resilience compiled by MWECP and Statistics Indonesia is calculated based on the dimensions of the foundation of legality and family integrity, physical resilience, economic resilience, social psychology resilience, and socio-cultural resilience. The method of determining the weighting of dimensions/variables/indicators from AHP is based on evaluations from experts who are subjective. Other than that, there needs to be a statistically valid and reliable measurement to measure the level of family resilience in Indonesia. Therefore, the aim of this study is to determine the quantitative comparison of family resilience index calculations.

2. Methodology

The dimensions and variables of P-FRI are latent variables that cannot be measured directly. The dimensions and variables P-FRI are measured based on 24 manifest/independent variables.

Table 1. Dimentions, Component Variables, and Indicators

Dimention (D)/Component Variables (V)/Indicator (X)			
D1		the foundation of legality and family integrity	
X1	:	the legality of marriage	the foundation of legality (V1)
X2	:	birth legality	
X3	:	family unity	family unity (V2)
X4	:	togetherness in the family	gender partnership (V3)
X5	:	partnership husband and wife	
X6	:	the openness of financial management	
X7	:	family decision making	
D2		Physical resilience	

X8	:	food adequacy	food dan nutritional adequacy (V4)
X9	:	nutritional adequacy	
X10	:	limitations of chronic diseases and disabilities	family health (V5)
X11	:	availability of fixed locations for sleep	availability of fixed locations for sleep (V6)
D3		Economic resilience	
X12	:	home ownership	family residence (V7)
X13	:	income per family	family income (V8)
X14	:	the adequacy of family income	
X15	:	ability to finance children's education	financing for children's education (V9)
X16	:	sustainability of children's education	
X17	:	family savings	family financial guarantee (V10)
X18	:	family health insurance	
D4		Social psychological resilience	
X19	:	anti-violence attitude towards women	family harmony (V11)
X20	:	the behavior of anti-violence against children	
X21	:	respect for the law	compliance with the law (V12)
D5		Social culture resilience	
X22	:	respect for the elderly	social care (V13)
X23	:	participation in social activities in the environment	social closeness (V14)
X24	:	participation in religious activities in the environment	religious obedience (V15)

Data Sources:

1. Integrated Database Update (PBDT Indonesia) 2015 by Statistics Indonesia.
2. National Socio-Economic Survey Characteristics of Household (Susenas Kor Indonesia) 2015 by Statistics Indonesia.
3. National Socio-Economic Survey of Socio-Cultural and Education Module (Susenas MSBP Indonesia) 2015 by Statistics Indonesia.
4. National Labor Force Survey (Sakernas Indonesia) 2015 by Statistics Indonesia.
5. National Socio-Economic Survey of Social Resilience Module (Susenas Module Hansos Indonesia) 2014 by Statistics Indonesia.
6. Happiness Measurement Survey (SPTK Indonesia) 2014 by Statistics Indonesia.
7. Publication of Basic Health Research (Riskesdas) 2013 by Indonesian Health Ministry.
8. Indonesia Demographic and Health Survey (SDKI) 2012 by Statistics Indonesia.

The analysis used is Structural Equation Modelling (SEM). Samples in this study are 34 provinces in Indonesia, therefore researchers used the Latent Variable Score (LVS) method as suggested by Wijanto (2015). The study was conducted by analysing each dimension to check the validity and reliability of each independent variable with Robust Estimation. After that every dimension and indicator that is valid and reliable is combined then analysed and seen how the overall suitability of each model is based on goodness of fit values.

The software program used is LISREL 9.1. Analysis is carried out for parameter estimation, testing the overall suitability of the model, and evaluating the overall suitability between the data and the model. Independent variables or dimensions that does not have a significant effect will be excluded from the model so that the valid and reliable dimensions and indicators are obtained.

3. Result

Based on AHP calculation, there are 5 provinces in Indonesia that have very high P-FRI values including Central Java (75.33), Bali (75.37), South Kalimantan (75.57), East Kalimantan (76.10), Riau Islands (77.56), and DI Yogyakarta (79.56). In addition, 3 provinces have low and very low PFRI values are: West Nusa Tenggara (64.99), East Nusa Tenggara (62.12), and Papua (56.56).

The following are the results of the initial data processing of the Structural Equation Modelling (SEM) of each dimension.

Table 2. Initial Data Results of Structural Equation Modeling (SEM)

	the foundation of legality and family integrity	Physical resilience	Economic resilience	Social psychological resilience*)	Social culture resilience *)
Standardized Solution					
Construct Reliability (CR)	0,67	1,04	0,79	0,78	0,67
Variance Extracted (VE)	1,01	2,13	1,51	0,90	0,84
Goodness of Fit					
degree of freedom	11	2	10	0	0
Chi Square p-value	83,967 0,000	1,070 0,5856	29,881 0,0009	0,00 1,00	0,00 1,00
RMSEA (Root Mean Square Error of Approximation) p-value	0,442 0,000	0,00 0,61	0,242 0,00206	0,00 NA	0,00 NA
GFI (Goodness of Fit Index)	0,653	0,985	0,811	NA	NA
AGFI (Adjusted Goodness of Fit	0,115	0,923	0,472	NA	NA

Index)					
NNFI (Nonnormed Fit Index)	-1,267	1,109	0,365	NA	NA
NFI (Normed Fit Index)	-0,018	0,966	0,656	NA	NA
AIC (Akaike Information Criterion) (saturated)	1334,255 (1272,288)	NA	1197,784 (1187,903)	424,466 (424,466)	500,946 (500,946)

Source: Results of data processing, 2018
NA: Not Available
*) after respesification

Four of the five dimensions have a positive influence on P-FRI. The four dimensions are: the foundation of the legality and integrity of the family (with a total influence of 0.994); physical endurance (with a total influence of 0.949); economic resilience (with a total influence of 0.051); and social psychology resilience (with a total influence of 0.751). While the dimensions of socio-cultural resilience have a negative influence on P-FRI with a total influence of-0.969.

The five dimensions are selected based on the validity and reliability test and the overall suitability of the model. The dimensions of the foundation of legality and family integrity have two valid and reliable independent variables (the openness of financial management and family decision making) and the suitability of the entire model to data that is good fit. The physical resilience dimension has two valid and reliable independent variables (nutritional adequacy and freedom from chronic illness and disability) and the overall suitability of the model to data that is good fit. The economic resilience dimension does not have any valid and reliable independent variable and the overall suitability of the model to the data is not good. The dimensions of social psychology resilience have two valid and reliable independent variables (the attitude of anti-violence against women and the behaviour of anti-violence against children) and the overall suitability of the whole model with data is good fit. And finally, the dimensions of socio-cultural resilience do not have any valid and reliable independent variables and the overall suitability of the model to the data is not good. Therefore, valid and reliable dimensions for calculating P-FRI are dimensions of the foundation of family legality and wholeness, physical resilience, and social psychological resilience. The results of SEM analysis with Latent Variable Score obtained 3 dimensions and 6 indicators that are valid and reliable and have a good level of suitability for calculating the level of family resilience.

After obtaining the valid and reliable dimensions of the level of family resilience, the family resilience index calculation is carried out again based on

those valid and reliable dimensions and indicators. There are 5 provinces in Indonesia that have high FRI values including Jambi (73.74), DI Yogyakarta (73.69), Riau Islands (72.30), Bali (72.07), Central Kalimantan (72.01), and West Sumatra (71.27). While 3 provinces that have very low FRI values are: Gorontalo (59.82), East Nusa Tenggara (59.17).and Papua (56.03).

4. Discussion and Conclusion

Dimensions of family legality and integrity

Based on the validity and reliability and overall suitability of the entire model, it is necessary to be respected by issuing invalid independent variables. From the magnitude of the effect of each independent variable and component of the variable, it can be seen that the birth legality has a positive effect on the foundation of legality of 0.523. As stated by MWECP (2016), birth legality influences the foundation of family legality where a positive influence indicates the higher the number of birth legalities, the stronger the foundation of legality. This value is greater than the legality of marriage on the basis of legality. This can be caused by data on marital legality that only includes poor households with the lowest level of welfare of 40 percent. In addition, husband and wife partnerships have very little influence on gender partnerships, which are only 0.001 compared to the other three variables in the gender partnership component. This shows that the partnership variable between husband and wife has not been able to describe the components of the gender partnership variable. At least the influence of gender partnership variables on the dimensions of legality and family integrity is not in line with the MWECP statement (2016) where family togetherness through managing the household together will affect family resilience/family integrity. The final results show, the dimensions of family legality and integrity have the best model with three valid independent variables namely togetherness in the family, openness of financial management, and family decision making.

Physical Resilience Dimensions

The four independent dimensions of physical resilience have SFLs below 0.50, which means that there are no valid independent variables. The respecification is done and finally a saturated model is obtained and has a perfect fit where the variable component remains removed. The final results show that the dimensions of physical resilience have the best model with two valid independent variables namely nutritional adequacy and freedom from chronic illness and disability. The SFL value of the two independent variables is greater than 0.05. The dimensions of physical resilience after respecification have a value of CR= 0.95(>0.70) and VE=0.90(>0.50) which means their liability is good. This is in line with the MWECP statement (2016) where the variable

nutritional adequacy and freedom from chronic diseases and disability can affect family physical resilience.

Dimensions of Economic Resilience

From the standardized solution results of SEM analysis, the dimensions of economic resilience are known that the independent variables of family income adequacy and family health insurance have SFLs below 0.50 which means they are invalid and can be excluded from the model. Based on the results of calculations, the initial CR value of the economic resilience dimension is 0.79 (> 0.70) and VE is 1.51 (> 0.50). So, it can be concluded that the reliability is good. The GOF value indicates that the overall suitability of the initial model dimension of economic resilience is not good. Therefore, based on the validity and reliability and suitability of the whole model, it can be concluded that the model is not good, so it needs to be respected by issuing invalid independent variables.

From the magnitude of the effects of each independent variable and component of the variable, it can be concluded that the adequacy of family income has a positive effect on the variable component of family income but is much smaller than the family income per capita. In addition, family health insurance has a smaller influence on family financial guarantees than family savings variables. This shows that the continuity variable of children's education and family health insurance can be excluded from the model because the variable is invalid and has little effect on the component variables and dimensions of economic resilience.

The respecification is done and a saturated model is obtained and has a perfect fit where the variable component is omitted. The final results show that the economic resilience dimension has the best model with three valid independent variables namely home ownership, the ability to finance children's education, and the continuity of children's education.

Dimensions of Social Psychological Resilience

The results of SEM analysis show that the independent variable of anti-violence attitude towards women and anti-violence behaviour towards children has an SFL greater than 0.50 which means the two independent variables are valid. While the independent variable of respect for the law has an SFL below 0.50 which means invalid. Based on the results of calculations, the dimensions of social psychology resilience have a value of CR = 0.78 (> 0.70) and VE = 0.90 (> 0.50). So, it can be concluded that the reliability is good. The invalid variable of respect for the law that is invalid is not in line with the statement of MWECP (2016) regarding the effect of respect for law that affects family resilience. From the suitability of the whole model, it seen that the value of Chi Square (0,00) is very small with a p-value of 1.00 which means good fit.

In addition, the RMSEA value is also very small, which is 0.00 (<0.08) which means good fit. The smallest AIC value (424,466) compared to other respecification models is equal to saturated value. So, it can be concluded that the model of respectiveness of the dimensions of psychological social resilience with the independent variables of anti-violence against women and anti-violence behaviour against children is the best model.

Dimensions of Socio-cultural Resilience

The results of SEM analysis showed that the three independent variables were respect for the elderly, participation in social activities in the environment, as well as participation in religious activities in the environment having SFLs below 0.50 which meant that none of the independent variables were valid. Based on the results of calculations, the dimensions of socio-cultural resilience have a value of CR = 0.67 (<0.70) and VE = 0.84 (> 0.50). So, it can be concluded that the reliability is not good. From the suitability of the whole model, it can be seen that the value of Chi Square (0,00) is very small with a p-value of 1.00 which means good fit. In addition, the RMSEA value is also very small, which is 0.00 (<0.08) which means good fit. The AIC value is quite small, which is 500.946 equal to the saturated value. So, it can be concluded that the model of respectiveness of the dimensions of socio-cultural resilience with the independent variable of respect for the elderly, participation in social activities in the environment, and participation in religious activities in the environment are the best models.

Latent Variable Score for Pioneer-Family Resilience Index(P-FRI) in 2015

Comparison of the values of P-FRI and FRI with valid and reliable dimensions and indicators is done to find out whether there are differences between the two index values. Based on analysis, there are differences in the results of calculating the P-FRI with the Family Resilience Index with the valid and reliable dimensions. The value of the Family Resilience Index with the valid and reliable dimensions tends to be lower compared to P-FRI. The difference between the two index values is between 0.24 to 11.35. This can be happened because the independent variables used in the calculations are different. But the values of the two indices are still in a range that is not much different, between 56.00 and 79.00. It can be concluded that there is no significant difference in the calculation of family resilience index using 6 indicators (SEM) or 24 indicators (AHP).

References

1. Minister of Women's Empowerment and Child Protection of the Republic of Indonesia Regulation Number 6 of 2013 Implementation of Family Development. February7, 2014. Jakarta.
2. The Ministry of Women's Empowerment and Child Protection. (2016). Family Resilience Development. Jakarta: KPPPA.
3. Sixbey, M.T. (2005). Dissertation: Development of the Family Resilience Assessment Scale To Identify Family Resilience Constructs. Florida: University of Florida.
4. Sunarti, Euis et al. (2003). Formulation of Measures of Family Resilience. Nutrition and Family Media27(1):1-11.
5. Wijanto, Setyo Hari. (2015). Research Methods Using Structural Equation Modeling with Lisrel 9. Jakarta: Institute for Publishers of the Faculty of Economics, University of Indonesia.

Multidimensional poverty among women in Morocco - Overview and analysis of the dynamics between 2004 and 2014

Hicham El Marizgui, Khalid Soudi
Observatory of Population's Living Conditions; High Commission of Planning of Morocco

Abstract
OPHI's approach for the measurement and the analysis of multidimensional poverty offers the advantage of breaking down this form of social deprivation. Indeed, since the multidimensional poverty index MPI can be decomposed into social groups and dimensions, it allows us to highlight the main determinants of multidimensional poverty. Thus, this approach implemented into women using exhaustively the 2004 and 2014 censuses data has shown that more than two million women are in a situation of multidimensional poverty, representing 18.1% of women population in Morocco compared to 40,4% of women who were concerned in 2004. Although declining at national, regional and provincial level, this form of poverty remains in the same configuration. Education is still the main scourge depriving women from attaining a situation of well-being, since years of schooling and illiteracy indicators explain respectively 34,1% and 31,5% of multidimensional poverty of women. In addition, the analysis by poverty typology found that 2% of women combine both monetary and multidimensional forms of poverty; these constitute the hard core of poverty.

Keywords
Poverty; Multidimensional; Monetary; OPHI; Women

1. Introduction

From a human development perspective, women's well-being is a prerequisite for sustainable development and a key objective of human progress. It implies the fight against female poverty, equal opportunities and empowerment. If women represent half of humanity, then the construction of a fairer world will necessarily require their social, political and economic integration and not their exclusion. From this perspective, exclusion and poverty are central concerns that must be addressed if the sustainability aspect of any development project is to be ensured.

The 1990s was a significant turnaround in the warning and fight against poverty. In 1992, in Rio de Janeiro, it was agreed that environmental protection involves reducing the masses of the poor who find their only resources in nature. In 1994, the Cairo conference considered poverty as a major obstacle to solving population problems. At the Fourth United Nations Conference on

Women, held in Beijing in 1995, poverty was classified as a problem of particular gravity for women.

It is currently recognized that the status of women is lower, and their poverty is higher than that of men. Women living in poor households are doubly poor or even more if the multidimensional aspect of poverty is taken into account. Because poverty cannot be reduced to insufficient income but also to the absence of choices and capacities (the case of unequal opportunities for access to public goods or services, to the labour market, to the exercise of power, etc.), the poor female population can only be poorer than traditional measures of poverty reveal.

Women's vulnerability to poverty, in the context of developing countries, is often reinforced by the prevailing socio-cultural value system. Among these values are stereotypes and prejudices according to which the role of women, because they are less empowered than men to perform other decision-making tasks and responsibilities, is that of mothers.

From the outset, the patriarchal regime, which continues to imbue the behaviour of traditional social strata with prejudice, combined with the predominant macho ideology, relegates the issue of women and delays any change in the relationship between the two sexes in order to create the foundations for improving the status of women, and thus their escape from their state of poverty. This seems to be perpetuated through the process of socialization, which differentiates children according to their gender and instils in everyone rules and behaviours specific to their gender.

This paper attempts to highlight what preceded, namely the status of poverty of women in Morocco. We will establish the first multidimensional poverty mapping of women based on the completeness of the 2014 population and housing census data, a typology of women's poverty in 2014 by combining the results of this mapping with those related to poverty from the income poverty mapping.

2. Methodology

The Oxford Poverty and Human Development Initiative (OPHI) approach bases the measurement of multidimensional poverty on a wide range of needs whose lack of satisfaction is a factor in the prevalence or manifestation of poverty or in its social reproduction. These needs include access to basic social services - water, electricity and sanitation, housing conditions, education, health and communication. These are the main objectives of the 2030 sustainable development agenda. By aggregating a series of unidimensional indicators of well-being, this approach provides information on the complex reality of well-being and identifies segments of the population in situations of multiple deprivation or multidimensional poverty. Thus, a person (a woman in our case) is considered multidimensionally poor if he/she accumulates a

number of deprivations above the poverty line, set by this approach at "at least 30% of the basic deprivations due to unmet needs in the above-mentioned areas".

The MPI we retained after discussing the context of women in morocco and considering data availability is structured using 12 indicators covering four dimensions: education, health, economic activity and living conditions. Each dimension includes one or more items of well-being that express a situation of deprivation of women in our case. These dimensions are also weighted with a quarter each (1/4) and each item in a dimension is proportionally weighted to the number of items in the same dimension (see table below). In this sense, health education and economic activity indicators each have a weight of 1/4 and living standards indicators a weight of 1/28 (see table below) The measurement approach consists of:

i Measure basic deprivation in relation to each poverty factor
ii Establish a composite deprivation index summarizing all elementary deprivations;
iii Calculate multidimensional poverty indices. (MPI)

Implemented on the exhaustiveness of the 2014 and 2004 censuses data, the retained national MPI of women has the following structure:

Dimension	Indicators	Cut-off	Weight	
Education	Years of schooling	If the woman is under 40 years old and has not completed five years of schooling	1/4	
	Illiteracy	If the woman is over 40 years old and is illiterate	1/4	
Health	Disability	If the woman is unable to perform any of the following organic functions: vision, hearing, walking, cognitive ability (remembering or concentrating), self-care and communication	1/4	
Economic Activity	Unemployment	If the woman is unemployed	1/8	
	Unpaid work	If the woman is employed as a family worker	1/8	
	Drinking water	If the household does not have access to clean water within 30 minutes of walking from home	1/28	1/4

Living Conditions	Electricity	If the household has no electricity	1/28	
	Sanitation	If the household does not have private toilets or a clean sanitation system	1/28	
	Floor	If the dwelling floor is dirty, sand or clay	1/28	
	Cooking fuel	If the household cooks using wood, coal or manure	1/28	
	Overcrowding	If, on average, there are more than 3 people per room	1/28	
	Assets	The household does not own a car, tractor or truck **and** does not own at least two of the following: telephone, television, radio, motorcycle, bicycle and refrigerator	1/28	

3. Results

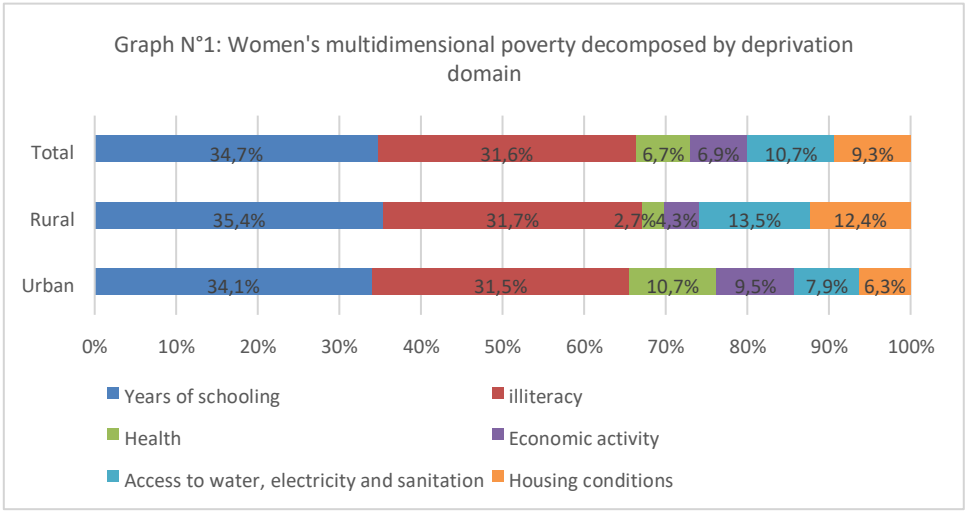
National level

In 2014, the incidence of multidimensional poverty among women aged 18 and over at the national level is 18.1%, which corresponds to 2.05 million poor women. With an incidence of 37.9%, nearly 1.58 million women are multidimensionally poor in rural areas. In cities, this phenomenon remains less pronounced (6.5%), i.e. nearly 470,000 women living in multidimensional poverty. Thus, in 2014, 77.2% of multidimensionally poor women in Morocco live in rural areas.

Between 2004 and 2014, multidimensional female poverty fell sharply. The incidence rose from 40.4% in 2004 to 18.1% in 2014 at the national level, from 30.5% to 6.5% in urban areas, and from 69.5% to 37.9% in rural areas. In 2004, the number of women who were multidimensionally poor was 3.8 million, representing an average annual reduction of 6.2%.

Furthermore, the decomposition of multidimensional female poverty by deprivation domain provides information on the sources behind this phenomenon. Deprivation in terms of school enrolment explains 34.7% of multidimensional female poverty at the national level; illiteracy contributes up to 31.6%. As for deprivation in terms of access to basic infrastructure, it explains 10.7% of this form of poverty, those in terms of housing conditions 9.3%. As for deprivation in terms of economic activity, it contributes 6.9% to

women's multidimensional poverty. This contribution reaches 6.7% for deprivations in terms of health.

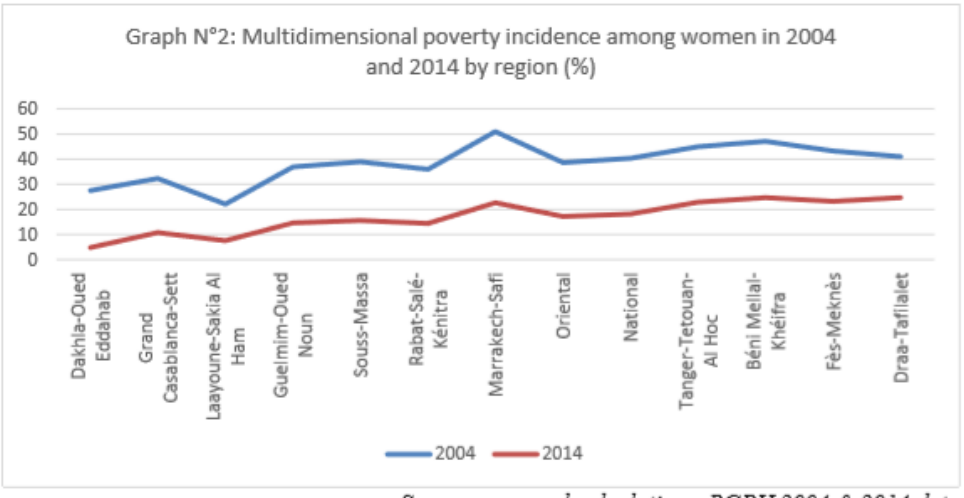


Source: personal calculations, RGPH 2004 & 2014 data

At the regional level:

In In 2014, the ranking of regions by the incidence of multidimensional poverty among women indicates that five regions have a poverty rate higher than the national average (18.1%). The most affected regions are Beni Mellal-Khenifra and Draa-Tafilalet with an incidence rate of 24.7% each, followed by FezMeknes (23.2%), Tangier-Tetouan (22.8%) and Marrakech-Safi (22.6%). Conversely, the regions with an incidence below the national average are Dakhla-Wadi Eddahab (4.8%), Laayoune-Sakia Al Hamra (7.5%), Casablanca-Settat (10.8%), Rabat-Salé-Kenitra (14.3%), Guelmim-Oued-Noun (14.6%), Souss-Massa (15.5%) and Oriental (17.2%).

Between 2004 and 2014, the incidence of multidimensional poverty among women declined in all regions of the Kingdom. The Dakhla Oued Eddahab region experienced the largest decline in terms of the incidence of multidimensional poverty among women, from 27.5% to 4.8%, followed by Casablanca Settat, from 32.3% to 10.8% and Laayoune Sakia Al Hamra, from 22.1% to 7.5%.



Source: personal calculations, RGPH 2004 & 2014 data

Typology of poor women in Morocco:

Based on the combined results of multidimensional mapping and monetary poverty mapping, a typology of Moroccan women according to the type of poverty they experience has been established. In this respect, three categories have been defined: the first category constitutes the hard core of poverty, it concerns women who combine both forms of poverty (monetary and multidimensional), the second category includes women who are poor in the sense of multidimensional poverty only and finally women who are subject to the monetary form of poverty only. The weight of these three groups determines the overall poverty rate of women in Morocco.

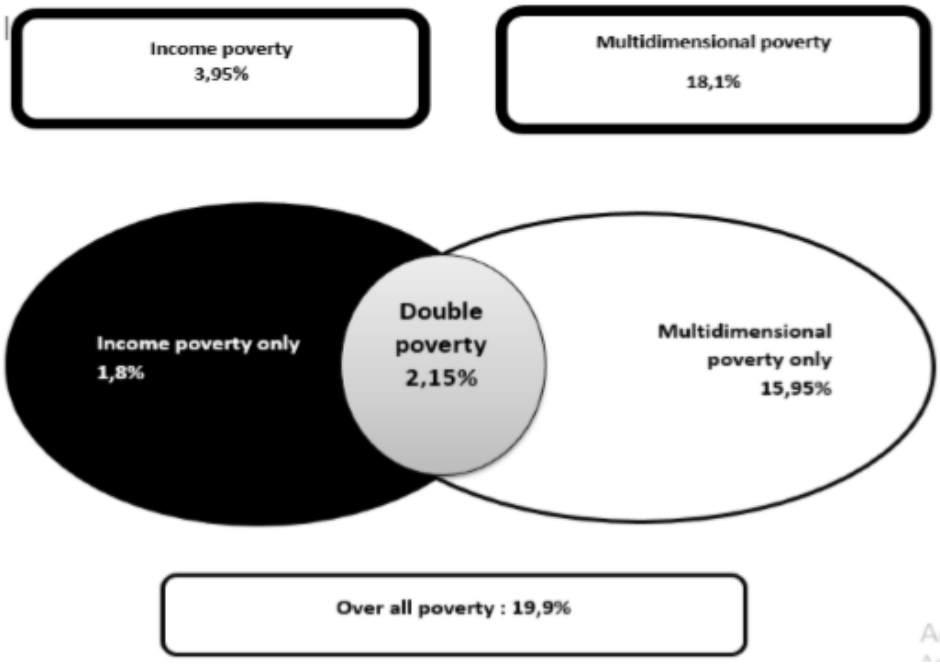
This typology highlighted that two million women suffer a single poverty (17.7% of the female population): 1.8 million women suffer exclusively from multidimensional poverty (15.9%) and 199 thousand women (1.8%) from monetary poverty only, 244 thousand women (2.1%), forming the hard core of poverty, are affected by both multidimensional poverty and monetary poverty. With these three indices, the volume of poverty in its monetary and multidimensional forms is 2.2 million women, representing an overall poverty rate of 19.9% nationally, 7.8% in urban areas and 40.4% in rural areas.

The hard core of poverty is 0.4% in urban areas and 5.2% in rural areas. At the regional level, the Draa Tafilalet region has the highest rate (6.8%) followed by Beni Mellal Khenifra (4.5%). In the provinces, the hard core of female poverty is most pronounced in Azilal (13.4%), followed by Zagora (12.4%), then Tinghir (9.1%) and Midelt (7.5%). Double poverty affects less than 5% of women in 68.4% of municipalities and urban centres, between 5% and 10% in 20.3%, between 10% and 20% in 8.9% and above 20% in 2%.

As for overall poverty, significant disparities emerge from the analysis on the territorial level: At the regional level, the overall poverty rate reaches 29.7% in Draa Tafilalet, 27.3% in Beni Mellal Khenifra, 24.8% in Fez Meknes, against

only 5% in Dakhla Oued Eddahab, 8.5% in Laayoune Sakia Al Hamra and 12.1% in Casablanca Settat; At the provincial level, Taounate province is the most affected by overall poverty with a rate of 62.8%, followed by Azilal (50.9%), Ouezzane (50.7%), Zagora (50.1%)%, Chefchaouen (47.1%), Moulay Yacoub (42.9%), Figuig (42.7%), Essaouira (42.4%), Chichaoua (39.4%).

At the communal level, nearly 15% of communes and urban centres have an overall female poverty rate of less than 10%, 17% between 10% and 20%, 29% between 20% and 40%, 28% between 40% and 70%, and 11% of communes and urban centres have an overall poverty rate above 70%.



4. Discussion and Conclusion

This work has enabled us to establish a database on the situation of Moroccan women's poverty in its non-monetary form. The breakdown of multidimensional female poverty by deprivation has provided us with information on the sources behind this phenomenon. In that respect, we chose to implement the MPI to the exhaustiveness of census data in order to analyse this scourge at the finest territorial and geographical level. This will make it possible to put in place local social policies specific to women in order to remedy each of the deprivations from which they suffer in their respective social context. Moreover, by combining the results of the mapping of income poverty already carried out by the HCP with the one we have just established, it has been possible to identify the "poorest of the poor" women, in particular those who combine both income poverty and multidimensional poverty, whom should be targeted by urgent social measures. In addition, the analysis of the dynamics of multidimensional poverty between 2004 and 2014 is a tool

for assessing the effectiveness of social programs already in place to combat poverty and inequality. Nevertheless, the choice to use census data has forced us to make compromises. It has not been possible to approach the health dimension by an indicator other than disability, which is debatable. It would therefore be relevant to implement the MPI for this category of the population in household survey data such as the national survey on consumption and household expenditure - representative at the regional level - which covers different areas of life including health and education, and in details. In addition to this, many questions have been implemented in the labour survey questionnaire, which will allow to monitor this form of poverty among women, but also among different categories of the population such as children and elderly. This will undoubtedly require a great deal of reflexion in order to develop specific multidimensional poverty indices whose structure respond to the deprivation situations of each of these categories

References

1. Sabina Alkire, James Foster, Suman Seth, Maria Emma Santos, José Manuel Roche and Paola Ballon Multidimensional poverty measurement and anlysis, 2015 Haut Commissariat au Plan, Poverty and shared prosperity in Morocco in the 3rd Millennium, 2016
   Link: https://www.hcp.ma/Pauvrete-et-prosperite-partagee-au-Maroc-du-troisieme-millenaire2001-2014_a2055.html

Factors influencing the economic growth using state Gross Domestic Product (GDP): A case study of Negeri Sembilan

Lim Kok-Hwa, Kon Mee-Hwa, Yaacob Hartini
Department of Statistics Malaysia

Abstract

The purpose of this article is to study the factors influencing the economic growth of Negeri Sembilan. The data used in this paper focuses on regional economic growth rate of GDP, particularly in the state of Negeri Sembilan. The growth of twelve year's time series data from Gross Domestics Product (GDP) by State from year 2005 to 2017 were employed in this study. This paper demonstrates the regression and correlation qualitative analysis of statistics data obtaining from the Department of Statistics Malaysia using an E-Views Program. In the first phase of the study, regression analysis has been used to construct appropriate model as well as to determine the correlation among Negeri Sembilan's GDP and the main economic indicators, namely agriculture, mining and quarrying, manufacturing, construction, services and import duties. Least Square Method analysis was applied to identify the significance of relationships among the economic indicators. Then, the second phase analysis focuses on the selected significant variables with other related economic indicators which consist of the consumer price index, unemployment rate and interest rate. The empirical result shows manufacturing and services are remaining as significant factors that influencing the growth of GDP in Negeri Sembilan. Besides, three linear econometric models were generated which indicate the correlation among economic indicators. This paper is initiated by the statisticians in the Department of Statistics Malaysia and to be served as a guideline to junior statisticians as well as the junior analyst on implement statistical analysis using time series of GDP data at regional level particularly at state level.

Keywords

Manufacturing; service; unemployment; CPI; correlation analysis.

1. Introduction

Negeri Sembilan is one of Malaysia's states located at 2°45'N 102°15'E coordinates on the western coast of Peninsular Malaysia. It covers 6,665 square kilometres land area of Malaysia with seven (7) districts namely Seremban, Port Dickson, Jempol, Jelebu, Kuala Pilah, Rembau and Tampin. The population in Negeri Sembilan is 1.128 million which is equal to 3.5% of total population in Malaysia in year 2017. The Gross Domestics Product (GDP) per capita for Negeri Sembilan is RM 39,763 in year 2017. Furthermore, the economic growth

in Negeri Sembilan was also influenced with the support from various project implementations, such as Economic Transformation Program (ETP), New Key Economic Area (NKEA), and Liberalisation impacted from the 10th Malaysia Plan towards Malaysia Vision 2020 at national level.

The objective of this paper is to study factors that influencing the economic growth and to identify the significant economic activities that performed as the major contributor towards the state economy in Negeri Sembilan. It also aims to present some endogenous analysis of statistical data with economic growth in Negeri Sembilan as well as to provide exposure to the regional economic data user particularly statisticians and analyst. The indicators of economic factors in Negeri Sembilan are illustrated via percentage growth shown by specific statistical indicators. Firstly, regression analysis using Least Square Method has been applied to construct econometric models and determine the correlation among the economic indicators in Negeri Sembilan. The relationship among GDP Negeri Sembilan with the five (5) economic main sectors, i.e. agriculture, mining & quarrying, manufacturing, constructions and services as well as import duties were being analyses. The analysis has been applied on the statistical indicators that measure regional economic growth to determine the causal relationship between state GDP and the selected available economic indicators. Then, further analysis focuses on the significant sectors and the growth of other economic indicators were implemented, namely the consumer price index, unemployment rate and interest rate.

Petrakos G., Kallioras D. and Anagnostou A. (2007) The opening of national borders, together with the rapid technological and scientific progress, has exposed regional economies to an extremely competitive, free-market, integrated economic environment, affecting their patterns of development. The paper develops a generalized econometric model for the investigation of the determinants of regional economic growth in 249 EU NUTS II regions, for the period 1990-2003. Fauzi H. and Soo Y.Y. (2012) examined the contribution of economic sectors to economic growth in China and India by using time series data from 1978 to 2007.The correlation analysis indicated that each agriculture, manufacturing and services sectors has strong, positive and significant linear relationship with economic growth in both countries. The results show that manufacturing sector contributes the highest to China's economic while services sector is the highest contributor to India's economic growth.

2. Methodology

This section described briefly about the concepts and statistical techniques applied in the GDP data analysis for economic growth in Negeri Sembilan. In the first phase, this paper studied the correlation and relationship among Negeri Sembilan's GDP with the main economic components, i.e. agriculture,

mining and quarrying, manufacturing, construction, services and import duties, were all applied in the Least Square Method analysis. Then, the second phase of analysis focuses on the significant variables with other related variables which consist of the consumer price index, unemployment rate and interest rate. In addition, this study has utilized Least Squares Analysis, Ordinary Correlations Analysis, descriptive analysis and linear relationship econometric model in identify the significant economic indicators to the economy.

Economic growth is a measure of the value of output of goods and services within a time period while economic development is a measure on welfare, investment, and development in a society.

In line with the objective of this study and the available time series data of State's GDP in Malaysia, a general econometrics model explains direct relationship of the state GDP as dependent variable Y, and X for economic activities as below:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \ldots\ldots\ldots + \beta nXn, \quad \text{where} \quad i = 1 \ldots n$$

The analysis begins with an examination of the coefficients significant test for each sector toward state GDP in Negeri Sembilan. The test of significance for βi or αi was carried out with Least Squares Method to obtain a specific econometrics model for state GDP Negeri Sembilan. Time series data used for this study is obtained from the growth of state GDP Negeri Sembilan by economic activities from year 2006 to 2017 obtained from the Department of Statistics, Malaysia. The modified linear econometrics model for the growth of the five economic main sectors and import duties as the influencing factors of GDP Negeri Sembilan have been developed as below:

$$GDPNegeriSembilan = \beta 0 + \beta 1\, X1 + \beta 2\, X2 + \beta 3\, X3 + \beta 4\, X4 + \beta 5\, X5 + \beta 6\, X6;$$ where,

The first phase with value of coefficients (βi),

Y = GDP Negeri Sembilan; X4 = Construction (CON);
X1 = Agriculture (AGR); X5 = Services (SER);
X2 = Mining and Quarrying (MNQ); X6 = Import Duties (IMD);
X3 = Manufacturing (MFG);

The second phase with value of coefficients (αi),

Y = GDP Negeri Sembilan; X4 = Consumer Price Index (CPI);
X1 = Agriculture (AGR); X5 = Unemployment Rate (UEM);
X2 = Manufacturing (MFG); X6 = Interest Rate (INT);
X3 = Services (SER);

and simple linear model relationship,

Y = GDP Malaysia; X1 = GDP Negeri Sembilan.

The main statistical tool applied for data analysis on Negeri Sembilan economic performance in this study is E-Views program. E-Views program is comprehensive and user friendly that managed to provide complicated

specialised functions for statistical operations particularly in econometric modelling and business performance analysis. Correlation analysis has been applied in order to identify the relationship among indicators in the time series data for both GDP at Malaysia level and Negeri Sembilan state level

3. Result

Base on the least square analysis result by sectors obtained from E-Views in table 1 as below, the result revealed that agricultural, manufacturing and services sectors had significantly impact positively the state GDP growth in Negeri Sembilan.

Table 1: Least Squares Analysis on GDP Negeri Sembilan by sectors

Dependent Variable: GDPNS
Method: Least Squares
Sample: 2006 – 2017 (percentage change)
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGRICULTURAL	0.108505	0.015969	6.794703	0.0011
MINING QUARRYING	0.023834	0.013718	1.737421	0.1428
MANUFACTURING	0.452906	0.027849	16.26312	0.0000
CONSTRUCTION	0.022950	0.010025	2.289174	0.0707
SERVICES	0.379021	0.045139	8.396732	0.0004
IMPORT DUTIES	0.004576	0.001002	4.567980	0.0060
C	0.051985	0.379862	0.136853	0.8965
R-squared	0.996050	Mean dependent var		4.676327
Adjusted R-squared	0.991310	S.D. dependent var		2.050857
S.E. of regression	0.191182	Akaike info criterion		-0.179981
Sum squared resid	0.182753	Schwarz criterion		0.102881
Log likelihood	0.079887	Hannan-Quinn criter.		-0.284707
Prob (F-statistics)	0.000008	Durbin-Watson stat		1.999832

Thus, the relationship among each sector towards GDP by State for Negeri Sembilan has formed an econometric model which represented as below:

$$\widehat{GDP}_{NegeriSembilan} = 0.051985 + 0.108505AGR + 0.023834MNQ + 0.452906MFG + 0.022950CON + 0.379021SER + 0.004576MD$$

Based on the model, the significant t-statistics where t > 2 and the significant probability value where p < 0.05 shown that manufacturing, services and agricultural sectors are influencing factors to the economic growth in GDP

Negeri Sembilan. Besides, further analysis has been done on the selected significant factors and other additional indicators, which consist of consumer price index, unemployment rate and interest rate. The second phase of Least Squares Analysis on GDP Negeri Sembilan by indicators is presented in Table 2 as below:

Table 2: Least Squares Analysis on GDP Negeri Sembilan by indicators

Dependent Variable: GDPNS
Method: Least Squares
Sample: 2006 –2017 (percentage change)
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGRICULTURAL	0.063009	0.025022	2.518119	0.0533
MANUFACTURING	0.381044	0.76844	4.958675	0.0043
SERVICES	0.292467	0.076303	3.832970	0.0122
CPI	0.021911	0.100659	0.217678	0.8363
UNEMPLOYMENTRATE	0.003087	0.014005	0.236181	0.8227
INTEREST RATE	1.959340	1.596790	1.227049	0.2744
C	1.288697	0.474867	2.924393	0.0328
R-squared	0.984605	Mean dependent var		4.676327
Adjusted R-squared	0.966130	S.D. dependent var		2.050857
S.E. of regression	0.377435	Akaike info criterion		1.180363
Sum squared resid	0.712287	Schwarz criterion		1.463226
Log likelihood	-0.082180	Hannan-Quinn criter.		1.075638
Prob (F-statistics)	0.000227	Durbin-Watson stat		2.534558

The data analysis in phase two with the involvement of consumer price index, the unemployment rate and the interest rate was run as the influencing factors. The empirical result indicated the most significant indicators that impact and influencing the economy of Negeri Sembilan is precisely driven by services sector and manufacturing sector. Therefore, the relationship among those indicators towards GDP for Negeri Sembilan has formed a comprehensive econometric model that demonstrated as below:

$$\widehat{GDP}_{NegeriSembilan} = 1.388697 + 0.063009AGR + 0.381044MFG + 0.292467SER - 0.021911CPI + 0.003308UEM + 1.959340INT$$

Table 3 illustrated the Ordinary Correlations Analysis for GDP Negeri Sembilan resulted from E-Views software. The tabulation indicated covariance, correlation, t-statistics and probability's magnitude of the covariance analysis among GDP in Negeri Sembilan and the selected economic indicators. In addition, Table 4 is the summary of descriptive analysis for the growth of GDP time series data in Negeri Sembilan from 2005 to 2017.

Table 3: Ordinary Correlations Analysis for GDP Negeri Sembilan

Covariance Analysis: Ordinary
Sample : 2006 2017
Included observations : 12

Covariance Correlation t-Statistic Probability	GDPNS	MANUFACT...	SERVICES	CPI	GDPMSIA
GDPNS	3.855514				
	1.000000				

MANUFACTURING	5.522222	8.803332			
	0.947871	1.000000			
	9.406517	-----			
	0.0000	-----			
SERVICES	3.116714	4.001746	4.022675		
	0.791404	0.672463	1.000000		
	4.094010	2.873162	-----		
	0.0022	0.0166	-----		
CPI	0.969436	1.080693	0.863929	1.659336	
	0.383275	0.282756	0.334390	1.000000	
	1.312233	0.932194	1.122024	-----	
	0.2188	0.3732	0.2881	-----	
GDPMSIA	2.432161	3.538510	1.984987	1.001560	4.202857
	0.604197	0.581734	0.482756	0.379260	1.000000
	2.397787	2.261683	1.743192	1.296163	-----
	0.0375	0.0472	0.1119	0.2240	-----

Table 4: Descriptive Analysis for the growth of GDP Negeri Sembilan, 2006-2017

Descriptive Analysis Result	GDP NS	Agricultural	Mining and Quarrying	Manufacturing	Construction	Services	CPI	Unemployment Rate	Interestate
Mean	4.676327	3.355942	8.191173	3.070887	7.785067	6.192173	2.669816	-1.372385	0.020960
Median	4.696182	3.671254	6.469865	3.076739	8.632011	5.516579	2.314738	-1.612903	0.032380
Maximum	9.262872	13.85576	19.32793	9.010444	17.00132	10.29931	5.393743	25.80645	0.248917
Minimum	1.042256	-6.995270	0.000000	-2.353305	-6.822516	3.196707	0.614125	-20.51282	-0.390072
Std. Dev.	2.050857	6.178692	6.424441	3.098973	7.240240	2.094844	1.345431	12.44534	0.157825
Skewness	0.457324	-0.041028	0.364103	0.233731	-0.606159	0.625771	0.500425	0.551830	-1.250050
Kurtosis	3.607106	2.202114	1.891585	2.643173	2.585558	2.617344	2.551803	3.115367	5.150334

On the other hand, GDP at national level for Malaysia has been targeted to grow 5-6% annually in the 11th Malaysia Plan 2016-2020 (EPU, 2015). Table 5 shows correlations analysis and the coefficients of Pearson Correlations analysis result between both GDP Malaysia and Negeri Sembilan. Analysis result shows the GDP Negeri Sembilan is moderately influencing the economy of GDP Malaysia

Table 5: Correlations analysis between GDP Malaysia and GDP Negeri Sembilan

Dependent Variable: MALAYSIA GDP
Method: Least Squares
Sample: 2006 2017
Observations period: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GDPNS	0.675359	0.2933948	2.297548	0.0507
C	1.681080	1.536680	1.093969	0.3058

R-squared	0.391533	Mean dependent var	4.910808
Adjust R-squared	0.322224	Durbin-Watson stat	2.304239
S.E. of regression	1.962940	Akaike info criterion	4.363620
Prob (F-statistics)	0.050664	Schwatz criterion	4.424137

Base on the correlations analysis result between GDP Malaysia and GDP Negeri Sembilan as shown in Table 5, the linear relationship model between the GDP at Malaysia level and the GDP Negeri Sembilan is illustrated as below:
$\widehat{GDP}$Malaysia = 1.931062+ 0.630827 $\widehat{GDP}$NegeriSembilan

Among the contribution of analysis presented the impact factor involvement of GDP Negeri Sembilan coefficient contributes towards GDP Malaysia is 0.630827. Despite of the impact from the implementation of Malaysia Plan (MP) towards the GDP rates in Negeri Sembilan, the fast growing of economic growth in Negeri Sembilan is also dominated by the physical economic projects development throughout the implementation of various projects under Economic Transformation Plan (ETP) and implementation of The Malaysian Urban-Rural National Indicators Network on Sustainable Development (MURNInets) program which has been initiated the by local authority under monitor and coordinate by the Federal Department of Town and Country Planning with the cross sectorial indicators, such as urbanisation, housing developments, heritage conservation and tourism encouragement, etc. Furthermore, in line with the objectives of ETP and the initiative on generating high income nation with the target to lift six per cent per annum (6%) of Malaysia's gross national income (GNI) per capita, the improvement of sustainable level in Negeri Sembilan also supported by both the domestic and foreign capital investment with the implementation of projects approved by MIDA towards Negeri Sembilan. The capital investment towards Negeri Sembilan had grown significantly over the years and the largest amount of approved capital investment for Negeri Sembilan was RM 5,905 million in year 2011. Among the notable investment projects approved by MIDA for Negeri Sembilan are the RM34.8 million Energy-efficient Investment, the port project, and the Oil Palm Trunk (OPT) project by a wholly Malaysian owned company based in Siliau, Negeri Sembilan, which provide job opportunities for about 100 Malaysians (MIDA, 2012).

4. Discussion and Conclusion

In conclusion, Negeri Sembilan is an open economy and is dependent on the trade investment for its economic development. Statistical data focus on the economy is proven to be input and catalyst to the advancement for the implementation of local socio-economic development programs and projects. Quality of the statistical data is also vital and important to the formation of policy for both state and the Malaysia country's development.

The empirical result from E-views analysis shows that manufacturing and services sectors remained as the key engine influencing the GDP growth in Negeri Sembilan. Hence, the causal link holds both the short-run and long-run to the state GDP growth and three linear econometric models that indicated economic indicators correlated in influencing the growth of GDP in Negeri Sembilan were generated as below:

$\widehat{GDP}_{\text{NegeriSembilan}}$ = 0.052 + 0.109AGR + 0.024MNQ + 0.453MFG + 0.023CON + 0.379SER + 0.005MD

$\widehat{\text{GDP}}$NegeriSembilan = 1.407 + 0.059AGR + 0.348 MFG + 0.327 Services - 0.043 CPI + 0.004 Unemployment Rate + 0.026 Interest rate

$\widehat{\text{GDP}}$Malaysia = 1.681+ 0.675 $\widehat{\text{GDP}}$NegeriSembilan

The indicators of regional economic development can be illustrated via the growth of GDP as well as the GDP per capita shown with other similar economic indicators or projects development. It is recommended that similar exercise of time series data analysis methods for other regional statistics data could be encouraging to apply in measuring the economic growth in Malaysia by using various indicators towards State GDP which served as an extra effort for junior statisticians to enhance the initial regional economic analysis by each State. This paper is to be enhanced and to be expanded in future for regional economic studies.

References

1. David L. K., James E. P. and Mildred E. W. (2007). *Role of Services in Regional Economy Growth*, Blackwell Publishing, Oxford, United Kingdom.
2. EPU. (2015). *Ringkasan Eksekutif Rancangan Malaysia Kesebelas 2016-2020*. Economic Planning Unit, Jabatan Perdana Menteri (JPM). Putrajaya.
3. EC, IMF, OECD, UN and World Bank (2009). *System of National Accounts (SNA) 2008*. European Commission (EC), International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD), United Nations (UN) and World Bank. New York.
4. Fauzi H. and Soo Y. Y. (2012). The *Contribution of Economic Sectors to Economic Growth: The cases of China and India.* Research in Applied Economics, ISSN 1948-5433 Vol. 4(4), Macrothink Institute, Universiti Utara Malaysia.
5. Jabatan Perancangan Bandar dan Desa (2013). *Laporan Kemampanan Bandar Negeri Sembilan 2012. Malaysian Urban-Rural National Indicators Network on Sustainable Development* (MURNInets). Jabatan Perancangan Bandar dan Desa (JPBD), Negeri Sembilan.
6. M. Idham M. R., Norazira M. A. and Noor Junaini A. Y. (2015). An *Overview of Primary Sector in Malaysia*. International Journal of Economics, Commerce and Management, ISSN 2348 0386, Vol. III Issue 2. Creative Common, United Kingdom.
7. MIDA (2013). *MIDA Approved Projects by State, 2000-2011*. Malaysian Investment Development Authority (MIDA). Kuala Lumpur.
8. Website: *http: //www.mida.gov.my/home/investment-data/posts/*
9. Online official website: http://www.bnm.gov.my/index.php?ch=statistic&pg=stats_convinterbkrates
10. Online official website: https://stats.oecd.org/glossary/detail.asp?ID=1163. *Glossary of Statistical Term*. Organisation for Economic Co-operation and Development (OECD)
11. Online official website e-services portal: http://www.statistics.gov.my. *GDP by State 2005-2017*, National Account. Department of Statistics Malaysia (DOSM).
12. Petrakos G., Kallioras D. and Anagnostou A. (2007). *A Generalized Model of Regional Economic Growth in the European Union*, University of Thessaly, Gree

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 978-967-2000-73-0

#ISIWSC2019